

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ALEXANDRE ALKMIM CHAMON

Quantificação de Carbono por Satélite

RIO DE JANEIRO
2025

ALEXANDRE ALKMIM CHAMON

Quantificação de Carbono por Satélite

Trabalho de conclusão de curso de graduação
apresentado ao Instituto de Computação da
Universidade Federal do Rio de Janeiro como
parte dos requisitos para obtenção do grau de
Bacharel em Ciência da Computação.

Orientador: Prof. Daniel Sadoc Menasche

RIO DE JANEIRO

2025

CIP - Catalogação na Publicação

C448q Chamon, Alexandre Alkmim
Quantificação de carbono por satélite / Alexandre
Alkmim Chamon. -- Rio de Janeiro, 2025.
33 f.

Orientador: Daniel Sadoc Menasche.
Trabalho de conclusão de curso (graduação) -
Universidade Federal do Rio de Janeiro, Instituto
de Computação, Bacharel em Ciência da Computação,
2025.

1. Biomassa florestal. 2. IA. 3. Sensoriamento
remoto. 4. Mercado de carbono. I. Menasche, Daniel
Sadoc, orient. II. Título.

ALEXANDRE ALKMIM CHAMON

Quantificação de Carbono por Satélite

Trabalho de conclusão de curso de graduação
apresentado ao Instituto de Computação da
Universidade Federal do Rio de Janeiro como
parte dos requisitos para obtenção do grau de
Bacharel em Ciência da Computação.

Aprovado em 18 de dezembro de 2025

BANCA EXAMINADORA:

Prof. Daniel Sadoc Menasché,
Orientador
Ph.D. (UFRJ)

Prof. Claudio Miceli de Farias
D.Sc. (UFRJ)

Prof. Rafael Maiani de Mello
D.Sc. (UFRJ)

AGRADECIMENTOS

Este trabalho é fruto de minha colaboração com colegas da IBM Research, onde trabalhei entre os anos de 2022 e 2023. Sou muito grato ao pesquisador Ademir da Silva e todos os co-autores dos trabalhos que deram origem a este TCC, por todas as suas contribuições e ensinamentos

Agradeço em especial à minha mãe Maria Alice que, sendo professora de geografia formada pela UFRJ, sempre me mostrou a importância da educação e sinto que seus ensinamentos sobre geografia de uma forma ou de outra me inspiraram na realização dessa monografia que integra nossas áreas de estudo. Agradeço também ao meu pai Carlos Eduardo que mesmo estando longe, conseguiu me ensinar muito sobre paciência e perseverança. E também ao meu irmão Guilherme Chamon que sempre esteve e estará comigo em todas as fases da minha vida.

Agradeço também aos meus amigos de infância Daniel Lucas, Raissa Leite, Derek de Carvalho e Renato Faria que me tornaram a pessoa que eu sou hoje. O apoio deles foi essencial para eu chegar a onde estou.

Agradeço também a minha esposa Julia Maria que me deu o amor, suporte e as forças necessárias durante o período mais difícil da minha vida.

RESUMO

Este trabalho de conclusão de curso investiga o potencial e as limitações do sensoriamento remoto por satélites e da inteligência artificial para a estimativa da biomassa florestal acima do solo (AGBD) em cenários de monitoramento de carbono em larga escala. A partir de um estudo de referência recente, analisamos uma abordagem de fusão de sensores que combina dados ópticos do Sentinel-2, dados de radar em banda C do Sentinel-1 e medições LiDAR da missão GEDI, utilizadas como referência para calibração. A metodologia é reproduzida e discutida sob a ótica do aprendizado de máquina, comparando-se Regressão Linear, métodos de ensemble baseados em árvores (Random Forest e XGBoost) e uma rede neural convolucional do tipo U-Net.

O foco central do trabalho é a comparação entre modelos “específicos por bioma”, treinados separadamente para regiões de baixa, média e alta biomassa, e um modelo “geral” único, treinado com todos os dados combinados. Os resultados confirmam que os modelos específicos tendem a apresentar erros ligeiramente menores; contudo, os ganhos são sistematicamente marginais, frequentemente inferiores a 1 Mg/ha em termos de RMSE, o que suscita dúvidas sobre o custo-benefício dessa estratégia em aplicações de escala nacional. O trabalho discute ainda o impacto da saturação dos sensores em florestas densas e explica por que o Random Forest superou a U-Net neste cenário, dada a natureza esparsa e pontual dos rótulos do GEDI. Por fim, são discutidas implicações para o mapeamento de carbono no Brasil, destacando o compromisso entre acurácia e complexidade operacional e apontando direções futuras envolvendo radar em banda L, arquiteturas de deep learning mais adequadas e validação direta nos biomas brasileiros para subsidiar políticas climáticas e mercados de carbono.

Palavras-chave: Biomassa Florestal; Inteligência Artificial; Sensoriamento Remoto; Comparação de Modelos; Sentinel; GEDI; Mercado de Carbono.

ABSTRACT

This undergraduate thesis investigates the potential and limitations of satellite remote sensing and artificial intelligence for estimating forest aboveground biomass density (AGBD) in the context of large-scale carbon monitoring. Building on a recent reference study, we examine a multi-sensor fusion approach that combines optical data from Sentinel-2, C-band SAR data from Sentinel-1, and LiDAR footprints from NASA’s GEDI mission as calibration reference. The methodology is reproduced and analysed from a machine learning perspective, comparing linear regression, tree-based ensemble methods (Random Forest and XGBoost) and a convolutional neural network (U-Net).

A central focus of the work is the comparison between “biome-specific” models, trained separately for regions with low, medium, and high biomass, and a single “general” model trained on pooled data. The results confirm that biome-specific models tend to achieve slightly lower errors; however, the gains are systematically marginal, often below 1 Mg/ha in RMSE, raising questions about their cost–benefit at national scale.

The thesis also discusses the impact of sensor saturation in dense forests and explains why Random Forest outperforms U-Net in this setting, given the sparse, point-based nature of GEDI labels. Finally, we draw implications for carbon mapping in Brazil, highlighting the trade-off between accuracy and operational complexity, and outlining research directions involving L-band radar, more suitable deep learning architectures, and validation in Brazilian biomes to support carbon accounting and emerging carbon markets.

Keywords: Forest Biomass; Artificial Intelligence; Remote Sensing; Model Comparison; Sentinel; GEDI; Carbon Market.

SUMÁRIO

1	INTRODUÇÃO	8
1.1	CONTEXTUALIZAÇÃO E PROBLEMA	8
1.2	OBJETIVOS	9
1.3	CONTRIBUIÇÕES	10
1.4	ORGANIZAÇÃO DO TRABALHO	10
2	FUNDAMENTAÇÃO TEÓRICA E ESTADO DA ARTE . . .	11
2.1	SENSORIAMENTO REMOTO PARA ESTIMATIVA DE BIOMASSA FLORESTAL	11
2.1.1	Sensores Ópticos (Sentinel-2): Princípios, Índices de Vegetação e Limitações	11
2.1.2	Sensores de Radar de Abertura Sintética (SAR - Sentinel-1): Interação com a Vegetação, Polarimetria e o Efeito de Saturação da Banda C	12
2.1.3	Tecnologia LiDAR (GEDI): Medição Direta da Estrutura Vertical e seu Uso como Referência	12
2.1.4	A Sinergia Multi-Sensor para Mapeamento de AGBD	13
2.2	MODELOS DE APRENDIZADO DE MÁQUINA PARA REGRESSÃO GEOESPACIAL	14
2.2.1	Métodos de Ensemble: Random Forest (Bagging) e XGBoost (Boosting)	14
2.2.2	Redes Neurais Convolucionais para Regressão Pixel a Pixel: A Arquitetura U-Net	15
2.3	DESAFIOS NO MAPEAMENTO DE BIOMASSA EM BIOMAS BRASILEIROS	16
2.3.1	Amazônia: Saturação de Sensores e Cobertura de Nuvens . .	16
2.3.2	Cerrado: Heterogeneidade Estrutural e Sazonalidade	17
3	DESENHO EXPERIMENTAL E PROCESSAMENTO DE DADOS	18
3.1	AQUISIÇÃO E CARACTERIZAÇÃO DOS DADOS	18
3.1.1	Dados Preditores (Variáveis Independentes)	18
3.1.2	Dados de Referência (Variável Dependente)	18
3.2	ÁREAS DE ESTUDO: CARACTERIZAÇÃO DE BIOMAS REPRESENTATIVOS	19
3.3	PIPELINE DE PRÉ-PROCESSAMENTO E FUSÃO DE DADOS	19
3.4	ESTRATÉGIA DE VALIDAÇÃO CRUZADA ESPACIAL	20
3.4.1	O Problema da Autocorrelação Espacial em Dados Geoespaciais	20

3.4.2	Implementação do Método de Discos Concêntricos	21
3.5	IMPLEMENTAÇÃO E TREINAMENTO DOS MODELOS DE REGRESSÃO .	21
4	ANÁLISE COMPARATIVA DOS RESULTADOS	23
4.1	AVALIAÇÃO QUANTITATIVA DO DESEMPENHO DOS MODELOS	23
4.2	ANÁLISE DA DISTRIBUIÇÃO DE ERROS E DA ACURÁCIA PREDITIVA	24
4.3	AVALIAÇÃO VISUAL E QUALITATIVA DOS MAPAS DE BIOMASSA GE- RADOS	26
5	DISCUSSÃO E IMPLICAÇÕES PARA O MAPEAMENTO DE CARBONO NO BRASIL	27
5.1	SÍNTESE DOS ACHADOS: A VANTAGEM MARGINAL DOS MODELOS ESPECÍFICOS	27
5.2	ANÁLISE CRÍTICA DOS MODELOS: POR QUE O RANDOM FOREST SUPEROU A U-NET NESTE CENÁRIO?	27
5.3	LIMITAÇÕES DA ABORDAGEM E O IMPACTO DA SATURAÇÃO EM FLO- RESTAS DENSAS	28
5.4	IMPLICAÇÕES PRÁTICAS PARA PROJETOS DE LARGA ESCALA . . .	29
6	CONCLUSÃO	30
6.1	RECAPITULAÇÃO DO PROBLEMA, METODOLOGIA E RESULTADOS .	30
6.2	CONSIDERAÇÕES FINAIS	30
6.3	RECOMENDAÇÕES PARA TRABALHOS FUTUROS	30
	REFERÊNCIAS	32

1 INTRODUÇÃO

A crise climática global, impulsionada pelo aumento das concentrações de gases de efeito estufa (GEE) na atmosfera, impõe a necessidade urgente de monitoramento preciso dos ciclos biogeoquímicos, especialmente o ciclo do carbono. As florestas desempenham um papel crucial neste equilíbrio, atuando como grandes sumidouros de carbono através da fotossíntese e armazenamento de biomassa. Estima-se que a vegetação terrestre armazene entre 450 e 650 Gigatoneladas de Carbono (GtC) (NATHANIEL et al., 2022). Consequentemente, a quantificação precisa da Densidade de Biomassa Acima do Solo (AGBD - *Aboveground Biomass Density*) tornou-se um requisito fundamental para a implementação de estratégias de mitigação, como os projetos de Redução de Emissões por Desmatamento e Degradação Florestal (REDD+) e o mercado de créditos de carbono (SILVA et al., 2023).

1.1 CONTEXTUALIZAÇÃO E PROBLEMA

Tradicionalmente, a estimativa de biomassa é realizada através de inventários florestais de campo. Embora precisos, esses métodos são onerosos, limitados espacialmente e inviáveis para o monitoramento contínuo de vastas áreas, como o território brasileiro. O sensoriamento remoto surge como a única alternativa viável para o monitoramento em larga escala. No entanto, o uso de satélites para esta finalidade enfrenta desafios técnicos significativos.

Sensores ópticos (como o Sentinel-2) dependem da iluminação solar e são frequentemente obstruídos por nuvens (Figura 1), uma característica persistente em florestas tropicais como a Amazônia (ASNER, 2000). Sensores de radar de abertura sintética (SAR), como o Sentinel-1, superam a barreira das nuvens, mas operam em comprimentos de onda (Banda C) que interagem predominantemente com o topo do dossel, saturando rapidamente em florestas densas e perdendo a sensibilidade para altos valores de biomassa (XIAO et al., 2019).

Recentemente, a disponibilização de dados LiDAR orbitais, como os da missão GEDI (*Global Ecosystem Dynamics Investigation*), abriu novas perspectivas. Embora o GEDI forneça medições diretas e precisas da estrutura vertical da floresta, sua cobertura é amostral e descontínua (Figura 2). A solução emergente, portanto, reside na fusão de dados multi-sensor, utilizando algoritmos de Inteligência Artificial (IA) para extrapolar as medições precisas do GEDI através da cobertura contínua fornecida por satélites como Sentinel-1 (radar) e Sentinel-2 (óptico) (DUBAYAH et al., 2020; SHENDRYK, 2022).

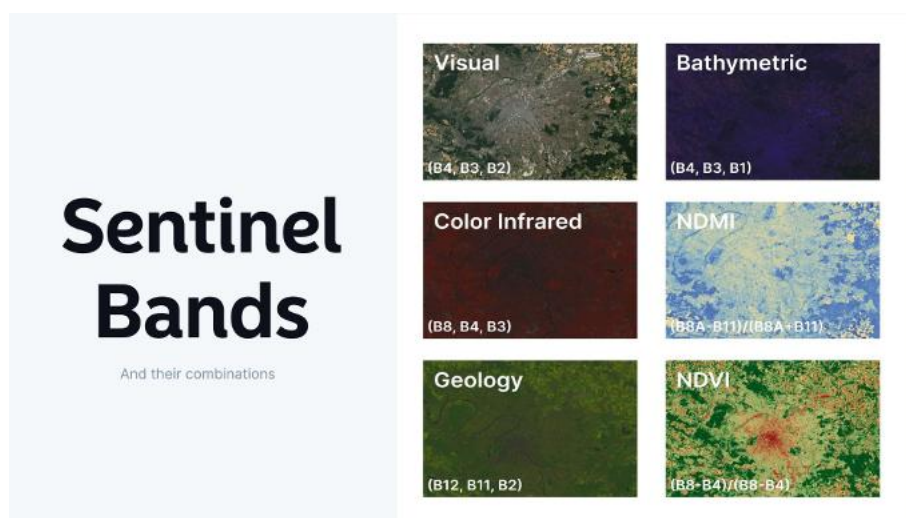


Figura 1 – Visualização das bandas espectrais do Sentinel-2

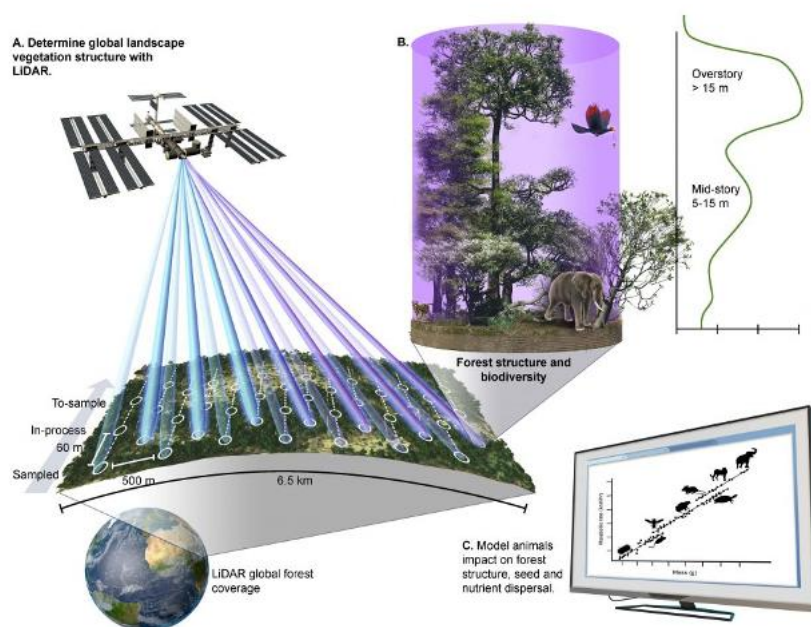


Figura 2 – Visualização da cobertura descontínua do GEDI

1.2 OBJETIVOS

O objetivo principal deste trabalho é realizar uma análise crítica e experimental de metodologias de IA para o mapeamento de biomassa. Baseando-se em um estudo de referência recente (SILVA et al., 2023), investigamos a eficácia de diferentes algoritmos de aprendizado de máquina — incluindo Regressão Linear, Random Forest, XGBoost e U-Net — e avaliamos o impacto de estratégias de modelagem gerais versus específicas por bioma. A motivação central é compreender se a customização de modelos para as características locais de cada bioma justifica a complexidade operacional adicional, e se essa abordagem é transferível para a realidade dos biomas brasileiros, notadamente a

Amazônia e o Cerrado.

1.3 CONTRIBUIÇÕES

Este trabalho apresenta as seguintes contribuições principais para a área de sensoria-
mento remoto e monitoramento ambiental:

- **Avaliação Comparativa de Estratégias de Modelagem:** Uma análise experimental rigorosa quantificando a diferença de desempenho entre modelos de Inteligência Artificial “Específicos por Bioma” versus modelos “Gerais”. Demonstra-se que, embora os modelos específicos ofereçam uma precisão superior, o ganho é marginal (frequentemente inferior a 1 Mg/ha em RMSE), levantando questões sobre o custo-benefício operacional dessa abordagem para monitoramento em escala nacional.
- **Análise Crítica da Aplicabilidade em Biomas Tropicais Densos:** Uma discussão aprofundada sobre as limitações físicas dos sensores utilizados (Sentinel-1 banda C e Sentinel-2 óptico) frente aos altos estoques de carbono da Amazônia. O trabalho identifica a saturação do sinal como um gargalo crítico que impede a transferência direta da metodologia do estudo de referência (realizado em biomas de menor densidade) para a realidade amazônica, sem a integração de dados de maior penetração (como banda L).
- **Diagnóstico de Desempenho de Arquiteturas de IA:** Uma investigação sobre as razões pelas quais algoritmos de *ensemble* baseados em árvores (Random Forest) superaram arquiteturas de *Deep Learning* (U-Net) neste cenário específico. Identifica-se o desalinhamento entre a natureza esparsa dos dados de referência (GEDI) e a necessidade de supervisão densa para o treinamento eficaz de redes convolucionais como a U-Net.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma: O **Capítulo 2** apresenta a fundamentação teórica sobre sensoriamento remoto, ciclo do carbono e as arquiteturas de redes neurais utilizadas. O **Capítulo 3** detalha a metodologia, incluindo o pré-processamento dos dados Sentinel e GEDI e a configuração dos experimentos. O **Capítulo 4** apresenta os resultados quantitativos e a análise comparativa entre os modelos. O **Capítulo 5** discute as implicações desses resultados para o monitoramento ambiental no Brasil. Por fim, o **Capítulo 6** apresenta as conclusões e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA E ESTADO DA ARTE

2.1 SENSORIAMENTO REMOTO PARA ESTIMATIVA DE BIOMASSA FLORESTAL

A estimativa de biomassa florestal por sensoriamento remoto baseia-se na premissa de que as propriedades da vegetação influenciam a maneira como a radiação eletromagnética interage com ela. Diferentes tipos de sensores capturam aspectos distintos dessa interação, cada um com suas potencialidades e limitações.

Neste contexto, destacam-se os instrumentos ópticos e de radar das missões Sentinel-2 e Sentinel-1, respectivamente, bem como a tecnologia LiDAR do sensor GEDI, cujas medições da estrutura vertical da floresta servem como referência fundamental para a calibração dos modelos.

2.1.1 Sensores Ópticos (Sentinel-2): Princípios, Índices de Vegetação e Limitações

Sensores ópticos passivos, como o MultiSpectral Instrument (MSI) a bordo dos satélites Sentinel-2, operam medindo a energia solar refletida pela superfície da Terra em diversas bandas espectrais, que vão do visível (VIS) ao infravermelho de onda curta (SWIR) (SILVA et al., 2023; NATHANIEL et al., 2022).

A vegetação saudável absorve fortemente a radiação na faixa do vermelho (para a fotossíntese) e reflete intensamente no infravermelho próximo (NIR), devido à estrutura celular das folhas. Essa assinatura espectral característica permite a criação de Índices de Vegetação (VIs), que são combinações matemáticas de diferentes bandas espectrais projetadas para realçar a presença e o vigor da vegetação. O Índice de Vegetação por Diferença Normalizada (NDVI), por exemplo, é um dos mais utilizados. O estudo base emprega 12 das bandas espectrais do Sentinel-2 como variáveis preditoras, explorando a riqueza de informações contida em todo o espectro medido (SILVA et al., 2023).

Apesar de sua ampla utilização, os sensores ópticos possuem duas limitações críticas para o monitoramento de florestas tropicais. A primeira é a sua dependência da iluminação solar e de condições atmosféricas claras. A presença de nuvens, frequente em regiões como a Amazônia, impede a aquisição de imagens, criando lacunas espaciais e temporais nos dados (ASNER, 2000; XIAO et al., 2019). A segunda limitação é a saturação do sinal em áreas de alta densidade de biomassa. Em florestas com dossel fechado e múltiplos estratos, os VIs atingem um valor máximo (platô) e perdem a sensibilidade a variações adicionais de biomassa. Isso ocorre porque o sensor capta apenas a reflectância da camada superior do dossel, não conseguindo “enxergar” a estrutura e a biomassa contidas abaixo dela (NATHANIEL et al., 2022; SHENDRYK, 2022).

2.1.2 Sensores de Radar de Abertura Sintética (SAR - Sentinel-1): Interação com a Vegetação, Polarimetria e o Efeito de Saturação da Banda C

Diferentemente dos sensores ópticos, os sensores de Radar de Abertura Sintética (SAR) são ativos, ou seja, emitem seus próprios pulsos de micro-ondas e medem a energia que é retroespalhada (backscatter) pela superfície. Essa capacidade lhes permite operar independentemente das condições de iluminação (dia e noite) e atmosféricas (atravessam nuvens e chuva), tornando-os particularmente valiosos para o monitoramento de regiões tropicais (BROWN, 2002; IMHOFF, 1995; SAATCHI; DUNCANSON; KELLNER, 2015).

O sinal de radar interage com os componentes estruturais da floresta, como troncos, galhos e folhas, e é sensível à sua geometria, orientação e conteúdo de umidade. A polarização do sinal (a orientação do campo elétrico da onda) é uma propriedade importante. O estudo base utiliza as polarizações VV (transmissão e recepção vertical) e VH (transmissão vertical e recepção horizontal) do satélite Sentinel-1 (SILVA et al., 2023). A polarização cruzada (VH) é geralmente mais sensível à dispersão volumétrica causada pela copa das árvores, sendo frequentemente mais correlacionada com a biomassa.

A principal limitação do Sentinel-1 para a estimativa de biomassa reside no seu comprimento de onda. Ele opera na banda C, com um comprimento de onda de aproximadamente 5.6 cm. Esse sinal curto interage predominantemente com os elementos menores da vegetação, como folhas e galhos finos, localizados na parte superior do dossel. A capacidade de penetração na estrutura densa da floresta é limitada. Como resultado, o sinal de retroespalhamento da banda C satura rapidamente em níveis de AGBD relativamente baixos, com estudos indicando um platô em valores entre 60 e 75 Mg/ha (IMHOFF, 1995; SAATCHI; DUNCANSON; KELLNER, 2015; NATHANIEL et al., 2022; SHENDRYK, 2022).

Este nível de saturação é significativamente inferior à biomassa encontrada em muitas florestas tropicais maduras, incluindo grandes porções da Amazônia e da Mata Atlântica, o que representa um desafio substancial para a sua utilização nessas áreas.

2.1.3 Tecnologia LiDAR (GEDI): Medição Direta da Estrutura Vertical e seu Uso como Referência

A tecnologia LiDAR (Light Detection and Ranging) também é um sistema de sensoriamento ativo, mas que utiliza pulsos de laser em vez de micro-ondas. Ao medir o tempo que um pulso de laser leva para atingir a vegetação e retornar ao sensor, o LiDAR é capaz de criar um perfil tridimensional extremamente preciso da estrutura vertical da floresta, desde o topo do dossel até o solo (BROWN, 2002; DUBAYAH et al., 2020).

A missão GEDI (Global Ecosystem Dynamics Investigation), da NASA, consiste em um sensor LiDAR a bordo da Estação Espacial Internacional. Ele fornece medições amostrais (footprints de aproximadamente 25 m de diâmetro) da estrutura florestal em escala

quase global. O GEDI é considerado o padrão-ouro para a calibração e validação de modelos de biomassa em larga escala, pois suas medições diretas da altura e do perfil vertical da vegetação estão fortemente correlacionadas com a AGBD e não sofrem do mesmo efeito de saturação que afeta os sensores ópticos e de radar (SILVA et al., 2023; BROWN, 2002; DUBAYAH et al., 2020; SHENDRYK, 2022).

A principal limitação do GEDI é sua natureza de amostragem. Ele não fornece uma cobertura contínua (“wall-to-wall”) do terreno, mas sim uma série de trilhas de medições pontuais. Isso torna necessário o uso de outras fontes de dados e de modelos estatísticos para extrapolar essas medições precisas e gerar mapas de biomassa espacialmente contínuos (DUBAYAH et al., 2020; XIAO et al., 2019).

2.1.4 A Sinergia Multi-Sensor para Mapeamento de AGBD

A fusão de dados de diferentes sensores não é apenas uma simples combinação de informações, mas uma estratégia de extrapolação inteligente para superar as limitações individuais de cada tecnologia. O processo pode ser entendido da seguinte forma:

1. **O Problema:** O objetivo é criar um mapa de biomassa preciso e de alta resolução para uma grande área, como o Brasil.
2. **A Referência Precisa, mas Esparsa:** O GEDI fornece medições de AGBD de alta qualidade, consideradas como a “verdade terrestre”, mas apenas em pontos discretos e dispersos no território. É impossível construir um mapa contínuo apenas com esses dados (DUBAYAH et al., 2020; XIAO et al., 2019).
3. **A Cobertura Contínua, mas Indireta:** Os satélites da família Sentinel fornecem uma cobertura completa e regular de todo o território. No entanto, os sinais que eles medem (reflectância óptica e retroespalhamento de radar de banda C) são proxies indiretos da biomassa e, crucialmente, saturam em florestas densas, o que pode levar a grandes subestimações (NATHANIEL et al., 2022; SHENDRYK, 2022; IMHOFF, 1995).
4. **A Solução Sinérgica:** A estratégia consiste em usar o melhor de cada sensor. Os pontos de medição do GEDI são utilizados como a variável alvo (a resposta, y) em um modelo de aprendizado de máquina. Os valores dos pixels correspondentes nos dados Sentinel (as 12 bandas ópticas, as 2 polarizações de radar, e outras variáveis derivadas como índices de vegetação ou texturas) são usados como as variáveis preditoras (as features, X).
5. **O Papel do Aprendizado de Máquina:** O algoritmo de ML é treinado para aprender a função matemática (f) que mapeia as características dos dados Sentinel

para os valores de biomassa do GEDI, ou seja, encontrar a relação $y \approx f(X)$ nos locais onde ambos os dados estão disponíveis.

6. **A Extrapolação:** Uma vez que o modelo f é treinado e validado, ele pode ser aplicado a todos os pixels do mapa Sentinel para os quais não existem medições GEDI. Desta forma, o modelo “preenche as lacunas”, extrapolando a acurácia das medições pontuais do GEDI para toda a área de cobertura dos dados Sentinel.

Esta abordagem sinérgica, que combina a precisão vertical do LiDAR com a cobertura espacial de sensores ópticos e de radar, é a base da metodologia do estudo de referência e representa o estado da arte no mapeamento de biomassa em larga escala (SILVA et al., 2023; KACIC; HIRNER; PONTE, 2021; SHENDRYK, 2022; NATHANIEL et al., 2022; XIAO et al., 2019; DUBAYAH et al., 2020).

2.2 MODELOS DE APRENDIZADO DE MÁQUINA PARA REGRESSÃO GEOESPACIAL

2.2.1 Métodos de Ensemble: Random Forest (Bagging) e XGBoost (Boosting)

Métodos de ensemble combinam as predições de múltiplos modelos mais simples (chamados de “preenche as lacunas”, tipicamente árvores de decisão) para produzir uma predição final mais robusta e precisa.

O Random Forest (RF) é um método de ensemble baseado na técnica de bagging (Bootstrap Aggregating). Ele constrói um grande número de árvores de decisão em paralelo. Cada árvore é treinada em uma amostra ligeiramente diferente dos dados de treinamento (amostragem bootstrap, com reposição) e, em cada nó da árvore, apenas um subconjunto aleatório das variáveis preditoras é considerado para encontrar o melhor ponto de corte. Para uma tarefa de regressão, a predição final do RF é a média das predições de todas as árvores individuais (BREIMAN, 2001; CHEN; GUESTIN, 2016; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O RF não otimiza uma função de custo global por meio de descida de gradiente. Em vez disso, cada árvore é construída de forma independente e “gulosa”, onde cada divisão de nó é escolhida para maximizar uma métrica local de pureza. Para regressão, essa métrica é tipicamente o Erro Quadrático Médio (MSE), e maximizar sua redução é equivalente a maximizar a redução da variância dos dados nos nós filhos (BREIMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Suas principais vantagens são a robustez ao sobreajuste (overfitting), a capacidade de lidar com relações não lineares e dados ruidosos, e a eficiência computacional, pois a construção das árvores pode ser facilmente paralelizada (SHENDRYK, 2022; BREIMAN, 2001; CHEN; GUESTIN, 2016).

O XGBoost (Extreme Gradient Boosting), por outro lado, é um método de boosting. Ele também constrói um ensemble de árvores de decisão, mas de forma sequencial e aditiva. A primeira árvore é treinada nos dados, e a segunda árvore é treinada para corrigir os erros (os resíduos) da primeira. A terceira corrige os erros combinados das duas primeiras, e assim por diante. Cada nova árvore é adicionada para minimizar uma função de custo global, utilizando um processo de otimização baseado em gradiente.

O XGBoost é uma implementação altamente otimizada e escalável deste conceito, que inclui técnicas de regularização (L1 e L2) para prevenir o overfitting (BREIMAN, 2001; CHEN; GUESTRIN, 2016). Frequentemente, o XGBoost atinge níveis de acurácia superiores ao Random Forest, especialmente em competições de ciência de dados, sendo considerado um dos algoritmos mais poderosos para dados tabulares (XIAO et al., 2019; BREIMAN, 2001; CHEN; GUESTRIN, 2016).

2.2.2 Redes Neurais Convolucionais para Regressão Pixel a Pixel: A Arquitetura U-Net

A U-Net é uma arquitetura de rede neural convolucional (CNN) originalmente desenvolvida para a segmentação de imagens biomédicas, uma tarefa que exige a classificação de cada pixel da imagem (RONNEBERGER; FISCHER; BROX, 2015; NATHANIEL et al., 2022). Sua arquitetura simétrica em forma de “U” a tornou extremamente popular e eficaz para diversas tarefas de mapeamento pixel a pixel. Ela é composta por duas partes principais:

- **Encoder (Caminho de Contração):** Esta parte da rede funciona como um extrator de características. Ela consiste em uma sequência de blocos, cada um contendo camadas convolucionais seguidas por uma camada de max-pooling. As convoluções extraem padrões e texturas da imagem em diferentes escalas, enquanto o pooling reduz a resolução espacial do mapa de características, forçando a rede a aprender representações mais abstratas e semânticas. Essencialmente, o encoder aprende a identificar “o que” está presente na imagem (RONNEBERGER; FISCHER; BROX, 2015; NATHANIEL et al., 2022).
- **Decoder (Caminho de Expansão):** Esta parte é responsável por reconstruir um mapa de saída com a mesma resolução da imagem de entrada. Ela utiliza operações de upsampling (como a convolução transposta) para aumentar gradualmente a resolução espacial dos mapas de características. O decoder utiliza as informações semânticas de baixa resolução do encoder para localizar com precisão as características, respondendo à pergunta “onde” (RONNEBERGER; FISCHER; BROX, 2015; NATHANIEL et al., 2022).

- **Skip Connections (Conexões de Salto):** Esta é a característica mais distintiva da U-Net. As skip connections ligam diretamente os mapas de características do encoder aos mapas de características de resolução correspondente no decoder. Essa fusão permite que o decoder recupere informações espaciais de alta frequência (detalhes finos, bordas) que foram perdidas durante as operações de pooling no encoder. O resultado é uma capacidade muito maior de produzir mapas de saída com limites precisos e detalhes espaciais preservados (RONNEBERGER; FISCHER; BROX, 2015).

Para adaptar a U-Net, originalmente uma arquitetura de classificação, para uma tarefa de regressão como a estimativa de biomassa, duas modificações fundamentais são necessárias. Primeiro, a camada de saída da rede, que tipicamente usa uma ativação softmax ou sigmoid para produzir probabilidades de classe, é modificada. Para a regressão, a última camada convolucional (geralmente com um kernel de 1x1) terá uma ativação linear ou nenhuma ativação, permitindo que a rede produza um valor contínuo e não limitado para cada pixel, que representará a AGBD em Mg/ha . O número de filtros nesta camada final é ajustado para 1, correspondendo ao único valor de biomassa a ser previsto por pixel (KACIC; HIRNER; PONTE, 2021; SHENDRYK, 2022).

Segundo, a função de perda (loss function) utilizada durante o treinamento é alterada. A perda de Entropia Cruzada (Cross-Entropy), padrão para problemas de classificação, é substituída por uma métrica de erro de regressão, como o Erro Quadrático Médio (MSE) ou o Erro Absoluto Médio (MAE), que quantificam a magnitude da diferença entre os valores de biomassa previstos e os valores de referência (SHENDRYK, 2022; XIAO et al., 2019). O estudo de base emprega o MSE para o treinamento da U-Net (SILVA et al., 2023).

2.3 DESAFIOS NO MAPEAMENTO DE BIOMASSA EM BIOMAS BRASILEIROS

A aplicação de metodologias de sensoriamento remoto e IA para mapeamento de biomassa no Brasil enfrenta desafios únicos, impostos pela vasta extensão e pela heterogeneidade de seus biomas.

2.3.1 Amazônia: Saturação de Sensores e Cobertura de Nuvens

A Floresta Amazônica é o maior bioma brasileiro e um dos mais importantes reguladores climáticos do planeta. Seu mapeamento de biomassa é dificultado por dois fatores principais. Primeiro, a alta densidade de biomassa, que frequentemente excede 250-300 Mg/ha , causa uma saturação severa tanto em sensores ópticos quanto em sensores de radar de banda C. Isso significa que, acima de um certo limiar, os sensores perdem a sensibilidade a aumentos na biomassa, resultando em subestimações sistemáticas e na incapacidade de diferenciar florestas maduras de florestas em estágios avançados

de regeneração (NATHANIEL et al., 2022; IMHOFF, 1995; SAATCHI; DUNCANSON; KELLNER, 2015).

Segundo, a cobertura de nuvens é quase persistente durante grande parte do ano, o que limita drasticamente a disponibilidade de imagens ópticas livres de contaminação. Embora os dados de radar superem essa limitação, a saturação da banda C do Sentinel-1 continua sendo um obstáculo significativo (ASNER, 2000; XIAO et al., 2019).

A imensa área do bioma também torna os inventários de campo, necessários para a calibração de modelos, uma tarefa logisticamente complexa e cara (ALENCAR et al., 2020).

2.3.2 Cerrado: Heterogeneidade Estrutural e Sazonalidade

O Cerrado, a savana mais biodiversa do mundo, apresenta um desafio diferente: a sua extrema heterogeneidade estrutural. Este bioma não é uma paisagem uniforme, mas um mosaico complexo de diferentes fisionomias vegetais, que variam desde campos abertos dominados por gramíneas (campo limpo), passando por savanas com árvores esparsas (cerrado sensu stricto), até formações florestais densas (cerradão) (KACIC; HIRNER; PONTE, 2021; ALENCAR et al., 2020; RIBEIRO; MACHADO; FILHO, 2009; JR. et al., 2020).

Essa variação estrutural ocorre em escalas espaciais curtas, tornando difícil para um único modelo de regressão capturar com precisão a relação entre o sinal do sensor e a biomassa em todas essas fisionomias. Além disso, o Cerrado possui uma forte sazonalidade, com uma estação seca e uma chuvosa bem definidas. Essa variação climática altera significativamente a resposta espectral da vegetação (por exemplo, a queda de folhas na estação seca), o que exige uma seleção cuidadosa do período das imagens a serem utilizadas ou a incorporação de métricas que capturem essa dinâmica temporal (fenologia) (KACIC; HIRNER; PONTE, 2021; RIBEIRO; MACHADO; FILHO, 2009). A alta frequência de incêndios, tanto naturais quanto antrópicos, é outro fator de perturbação que adiciona complexidade ao monitoramento da biomassa neste bioma (ALENCAR et al., 2020).

3 DESENHO EXPERIMENTAL E PROCESSAMENTO DE DADOS

Este capítulo descreve em detalhe a metodologia empregada no estudo de referência, que serve de base para este trabalho. O objetivo é fornecer uma descrição clara e replicável do pipeline, desde a aquisição dos dados até o treinamento e a avaliação dos modelos de inteligência artificial.

3.1 AQUISIÇÃO E CARACTERIZAÇÃO DOS DADOS

A metodologia se baseia na fusão de dados de três fontes satelitais distintas, cada uma desempenhando um papel específico no processo de modelagem.

3.1.1 Dados Preditores (Variáveis Independentes)

As variáveis utilizadas para treinar os modelos e prever a biomassa foram extraídas de dois satélites do programa Copernicus da Agência Espacial Europeia (ESA):

- **Sentinel-1:** Foram utilizadas imagens de radar de abertura sintética (SAR) operando em banda C. Especificamente, o produto Ground Range Detected (GRD) no modo Interferometric Wide swath (IW), que fornece dados de intensidade do retroespalhamento. As duas polarizações disponíveis, co-polarizada (VV) e cruzada (VH), foram empregadas como duas variáveis preditoras distintas (SILVA et al., 2023).
- **Sentinel-2:** Foram utilizadas imagens multiespectrais do produto Nível-2A, que fornece dados de reflectância de superfície já corrigidos para efeitos atmosféricos. Do total de 13 bandas espectrais disponíveis, 12 foram utilizadas, descartando-se apenas a banda 10 (1375 nm), que é primariamente usada para detecção de nuvens do tipo cirrus (SILVA et al., 2023).

3.1.2 Dados de Referência (Variável Dependente)

A “verdade terrestre” para o treinamento e a validação dos modelos foi obtida a partir da missão GEDI da NASA:

- **GEDI:** Foi utilizado o produto de Nível 4A (L4A), Versão 2.1, denominado Aboveground Biomass Density (AGBD). Este produto fornece estimativas diretas da densidade de biomassa acima do solo, em Mg/ha , para cada footprint (pegada) individual do sensor LiDAR, que tem aproximadamente 25 m de diâmetro (SILVA et al., 2023; DUBAYAH et al., 2020).

3.2 ÁREAS DE ESTUDO: CARACTERIZAÇÃO DE BIOMAS REPRESENTATIVOS

Para avaliar o desempenho dos modelos em diferentes condições de vegetação, o estudo selecionou três áreas quadradas de 50×50 km nos Estados Unidos. Essas áreas foram escolhidas por apresentarem densidades de biomassa muito distintas, servindo como análogos (proxies) para diferentes tipos de biomas encontrados globalmente (SILVA et al., 2023):

- **Biome 1 (Jayton, Texas):** Caracterizado por uma vegetação de baixa biomassa, com uma AGBD média de apenas 4.1 Mg/ha . Este ambiente pode ser considerado análogo a ecossistemas semiáridos ou savanas muito abertas, como partes da Caatinga ou do Cerrado mais ralo.
- **Biome 2 (Goliad, Texas):** Apresenta uma biomassa intermediária, com AGBD média de 17.3 Mg/ha . Corresponde a um ambiente de savana ou floresta aberta, análogo a formações típicas do Cerrado sensu stricto.
- **Biome 3 (Oberlin, Louisiana):** Possui uma vegetação de alta biomassa, com AGBD média de 94.6 Mg/ha . Este local representa um ambiente de floresta temperada densa, servindo como análogo para florestas tropicais de menor porte, como trechos da Mata Atlântica ou da Amazônia em regeneração.

A distribuição dos valores de AGBD em cada um desses biomas, visualizada através de histogramas, confirma suas características distintas, com o Bioma 1 concentrando valores muito baixos, o Bioma 2 com uma distribuição intermediária e o Bioma 3 exibindo uma ampla gama de valores, chegando a centenas de Mg/ha (SILVA et al., 2023).

3.3 PIPELINE DE PRÉ-PROCESSAMENTO E FUSÃO DE DADOS

Um rigoroso processo de pré-processamento foi aplicado para garantir a consistência e a qualidade dos dados antes da modelagem (Figura 4).

- **Seleção Temporal e Co-registro:** Para cada área de estudo, foi selecionada uma única cena do Sentinel-2 com a menor cobertura de nuvens possível, adquirida durante o verão de 2019 (entre junho e setembro). As imagens do Sentinel-1 foram adquiridas com uma diferença temporal de, no máximo, uma semana em relação à imagem do Sentinel-2, para garantir que ambas representassem o mesmo estado da vegetação. Todos os dados GEDI disponíveis dentro desse período foram coletados (SILVA et al., 2023).
- **Construção do Cubo de Dados:** A fim de criar um conjunto de dados coeso, todas as bandas dos sensores Sentinel-1 e Sentinel-2 foram reamostradas para uma

resolução espacial comum de 20 metros. Esta resolução foi escolhida para ser compatível com o diâmetro aproximado dos footprints do GEDI. Em seguida, as 12 bandas ópticas e as 2 bandas de radar foram empilhadas, criando um “cubo de dados” onde cada pixel é representado por um vetor de 14 características (SILVA et al., 2023).

- **Tratamento dos Dados GEDI:** Para assegurar a alta qualidade dos dados de referência, foram aplicados dois filtros. Primeiro, apenas os footprints do GEDI com um valor de sensitivity superior a 0.95 e com as flags de qualidade L2A e L4A ativadas foram retidos. Segundo, pontos com AGBD estimada acima de 400 *Mg/ha* foram removidos, pois foram considerados outliers potenciais, representando menos de 0.1% do total de dados (SILVA et al., 2023).

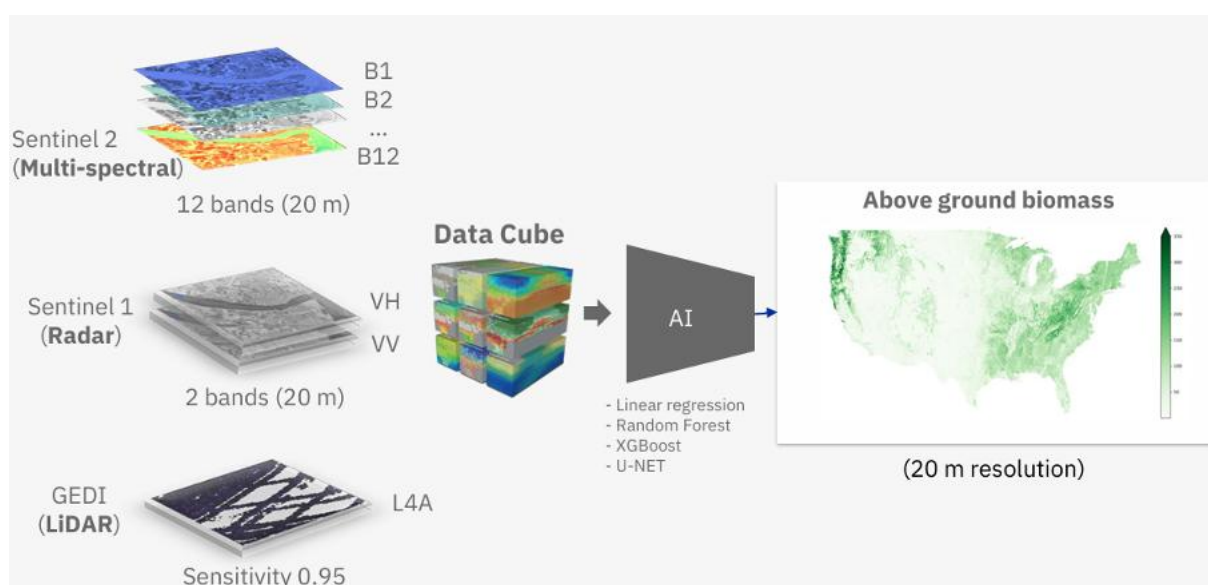


Figura 3 – Pipeline de transformação dos dados dos satélites em um “data cube” para o uso de entrada em algoritmos de aprendizagem de máquina

3.4 ESTRATÉGIA DE VALIDAÇÃO CRUZADA ESPACIAL

3.4.1 O Problema da Autocorrelação Espacial em Dados Geoespaciais

Dados geoespaciais violam uma premissa fundamental de muitos algoritmos de aprendizado de máquina: a de que as amostras são independentes e identicamente distribuídas (i.i.d.). De acordo com a Primeira Lei da Geografia de Tobler, “coisas próximas no espaço são mais semelhantes entre si do que coisas distantes” (TOBLER, 1970; ROBERTS et al., 2017). Esse fenômeno, conhecido como autocorrelação espacial, significa que os valores de biomassa em dois pontos próximos tendem a ser mais parecidos do que em dois pontos distantes. Se uma estratégia de validação cruzada aleatória padrão (como k-fold) for aplicada a esses dados, é muito provável que amostras geograficamente próximas acabem

sendo divididas entre os conjuntos de treinamento e teste. Um modelo treinado com uma amostra pode, então, prever com alta acurácia o valor de uma amostra de teste vizinha, não porque aprendeu uma relação geral robusta, mas simplesmente porque a amostra de teste é muito semelhante à de treino. Isso leva a uma avaliação excessivamente otimista do desempenho do modelo e a uma falsa sensação de sua capacidade de generalização para áreas verdadeiramente novas e independentes (ROBERTS et al., 2017).

3.4.2 Implementação do Método de Discos Concêntricos

Para mitigar o efeito da autocorrelação espacial e obter uma avaliação mais realista do desempenho do modelo, o estudo de base empregou uma técnica de validação cruzada espacial em blocos, implementada através de um método de discos concêntricos (SILVA et al., 2023):

1. Em cada uma das três áreas de estudo, 100 pontos foram selecionados aleatoriamente para servirem como centros.
2. Ao redor de cada um desses 100 centros, dois discos concêntricos foram definidos: um interno com 1 km de diâmetro (500 m de raio) e um externo com 2 km de diâmetro (1 km de raio).
3. **Conjunto de Teste:** Todos os footprints do GEDI localizados dentro dos discos internos (raio < 500 m) foram alocados para o conjunto de teste.
4. **Conjunto de Treinamento:** Todos os footprints do GEDI localizados fora dos discos externos (raio > 1 km) foram alocados para o conjunto de treinamento.
5. **Zona de Exclusão (Buffer):** Os footprints do GEDI que caíram na área anelar entre os dois discos ($500 \text{ m} < \text{raio} < 1 \text{ km}$) foram completamente descartados do experimento.

Essa abordagem garante que cada ponto no conjunto de teste esteja a uma distância mínima de 500 metros de qualquer ponto no conjunto de treinamento. Essa separação espacial força o modelo a ser avaliado em sua capacidade de extrapolação local, que é precisamente o que se deseja ao criar um mapa contínuo a partir de dados de treinamento esparsos. É uma metodologia muito mais robusta para avaliar modelos geoespaciais do que a validação aleatória tradicional (ROBERTS et al., 2017).

3.5 IMPLEMENTAÇÃO E TREINAMENTO DOS MODELOS DE REGRESSÃO

Com os dados devidamente preparados e divididos, os modelos de regressão foram treinados sob duas abordagens distintas:

- **Modelos Específicos:** Para cada um dos três biomas, os quatro algoritmos selecionados (Regressão Linear, Random Forest, XGBoost e U-Net) foram treinados utilizando exclusivamente o conjunto de treinamento correspondente àquele bioma. Isso resultou em um conjunto de três modelos para cada algoritmo, cada um especializado em uma faixa de densidade de biomassa.
- **Modelo Geral:** Os conjuntos de treinamento dos três biomas foram agregados para formar um único conjunto de dados maior e mais diversificado. Os quatro algoritmos foram, então, treinados neste conjunto de dados unificado, resultando em um único modelo “geral” para cada algoritmo, projetado para abranger todas as faixas de biomassa.

Os detalhes de implementação seguiram as melhores práticas. Para Random Forest e XGBoost, uma busca em grade (grid search) com validação cruzada de 4 folds foi usada para otimizar os principais hiperparâmetros, como `max_features` para RF e `max_depth` e `learning_rate` para XGBoost (SILVA et al., 2023). Para a U-Net, o treinamento foi realizado com patches (recortes) de imagem de 100×100 pixels, utilizando o otimizador Adam e uma taxa de aprendizado de 0.01, com um agendador (scheduler) que reduzia a taxa de aprendizado quando a métrica de erro (MSE) parava de melhorar em um platô (SILVA et al., 2023).

4 ANÁLISE COMPARATIVA DOS RESULTADOS

4.1 AVALIAÇÃO QUANTITATIVA DO DESEMPENHO DOS MODELOS

A avaliação do desempenho dos modelos foi realizada utilizando duas métricas de erro padrão para tarefas de regressão: a Raiz do Erro Quadrático Médio (RMSE) e o Erro Médio Absoluto (MAE), ambas expressas em Mg/ha . A Tabela 4.1 consolida os resultados obtidos nos conjuntos de teste para cada combinação de bioma, abordagem (Específica vs. Geral) e algoritmo de IA, com base nos dados apresentados no estudo de referência (SILVA et al., 2023).

Tabela 1 – Comparação de Métricas de Erro (RMSE e MAE em Mg/ha) para Modelos de Estimativa de AGBD.

Bioma	Abordagem	Modelo	AGBD Média (Mg/ha)	RMSE (Mg/ha)	MAE (Mg/ha)
1 (Baixa Biomassa)	Específica	Regressão Linear	4.1	8.95	3.63
		Random Forest		8.85	3.34
		XGBoost		8.87	3.34
		U-Net		9.14	2.67
	Geral	Regressão Linear		21.58	16.85
		Random Forest		9.19	3.72
		XGBoost		9.41	3.87
		U-Net		9.22	3.89
2 (Biomassa Intermediária)	Específica	Regressão Linear	17.3	20.76	10.39
		Random Forest		16.83	9.50
		U-Net		16.42	9.83
		XGBoost		16.57	9.29
	Geral	Regressão Linear		24.56	17.41
		Random Forest		17.23	9.61
		U-Net		22.61	10.59
		XGBoost		18.28	10.01
3 (Alta Biomassa)	Específica	Regressão Linear	94.6	68.85	52.00
		Random Forest		54.79	36.18
		U-Net		60.85	37.55
		XGBoost		55.39	36.79
	Geral	Regressão Linear		70.73	52.98
		Random Forest		54.79	36.09
		U-Net		58.36	36.56
		XGBoost		53.80	35.85

A análise da Tabela 4.1 revela várias tendências importantes. Primeiramente, os modelos baseados em árvores de decisão (ensemble), Random Forest e XGBoost, demonstram um desempenho consistentemente superior em comparação com a Regressão Linear, que serve como linha de base. A U-Net, embora seja uma arquitetura mais complexa, não superou os modelos de ensemble na maioria dos cenários, com exceção do MAE no Bioma 1.

Em segundo lugar, ao comparar as abordagens “Específica” e “Geral”, observa-se que os modelos específicos apresentam, na maioria dos casos, valores de RMSE e MAE ligeiramente inferiores. Por exemplo, para o Random Forest no Bioma 1, o RMSE do modelo específico foi de $8.85 Mg/ha$, enquanto o do modelo geral foi de $9.19 Mg/ha$. No Bioma 2, a diferença foi de $16.83 Mg/ha$ para $17.23 Mg/ha$. Essa tendência sugere que a

calibração local confere uma pequena vantagem em termos de acurácia. No entanto, é notável que a magnitude dessa vantagem é bastante marginal, frequentemente inferior a 1 Mg/ha (SILVA et al., 2023).

4.2 ANÁLISE DA DISTRIBUIÇÃO DE ERROS E DA ACURÁCIA PREDITIVA

Para além dos valores médios de erro, a análise da distribuição dos erros de predição oferece uma compreensão mais profunda do comportamento dos modelos. Os gráficos de violino (Figura 4), que ilustram a densidade de probabilidade dos erros para cada modelo e abordagem, revelam padrões claros (SILVA et al., 2023).

- **Aumento do Erro com a Biomassa:** A dispersão dos erros (a “largura” dos violinos) aumenta progressivamente do Bioma 1 para o Bioma 3. Isso indica que a incerteza e a magnitude dos erros de predição são maiores em ecossistemas de alta biomassa, um resultado esperado, dadas as limitações dos sensores.
- **Semelhança entre Abordagens para RF e XGBoost:** Para os modelos Random Forest e XGBoost, as distribuições de erro dos modelos “Específico” e “Geral” são visualmente muito semelhantes. Isso reforça a conclusão quantitativa de que o ganho de acurácia ao especializar o modelo é modesto.

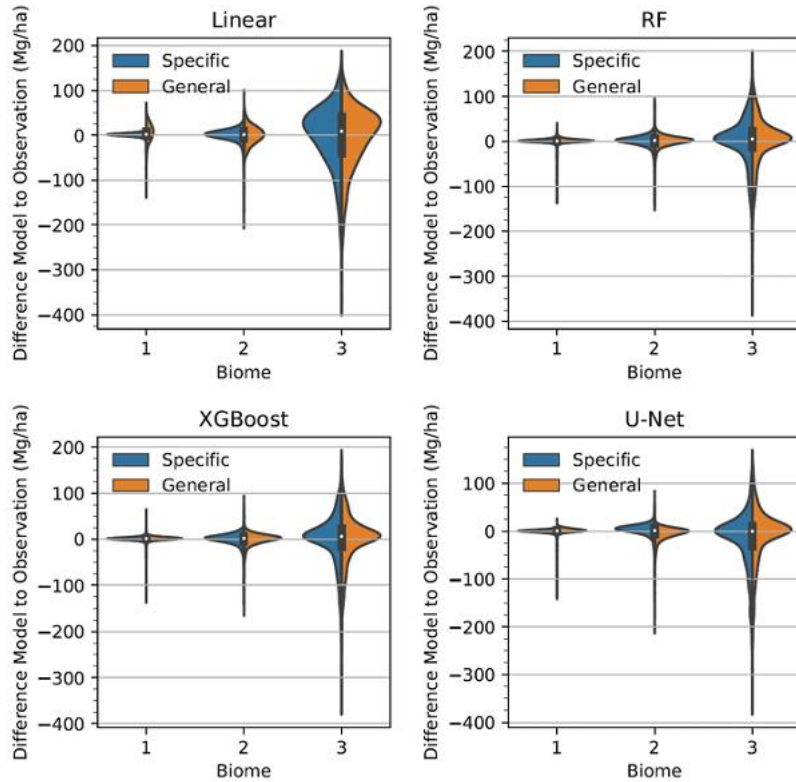


Figura 4 – Comparação do erro entre modelos para biomas específicos e gerais

A análise dos gráficos de dispersão (Figura 5), que comparam os valores de AGBD preditos pelos modelos com os valores de referência do GEDI, fornece mais detalhes sobre a natureza dos erros (SILVA et al., 2023).

- **Concentração em Baixa Biomassa:** Nos Biomas 1 e 2, os pontos se agrupam de forma relativamente compacta em torno da linha diagonal (que representa a predição perfeita), indicando um bom ajuste do modelo para valores de biomassa mais baixos.
- **Heterocedasticidade em Alta Biomassa:** No Bioma 3, o padrão é diferente. A dispersão dos pontos aumenta significativamente para valores mais altos de AGBD. Isso é um sinal clássico de heterocedasticidade, onde a variância do erro não é constante, mas aumenta com o valor da variável predita. Este comportamento é uma consequência direta do efeito de saturação dos sensores ópticos e de radar de banda C. À medida que a biomassa real aumenta, o sinal do sensor para de responder, e o modelo perde sua capacidade de discriminar entre valores altos de AGBD, resultando em maior incerteza e erros mais dispersos (SILVA et al., 2023).

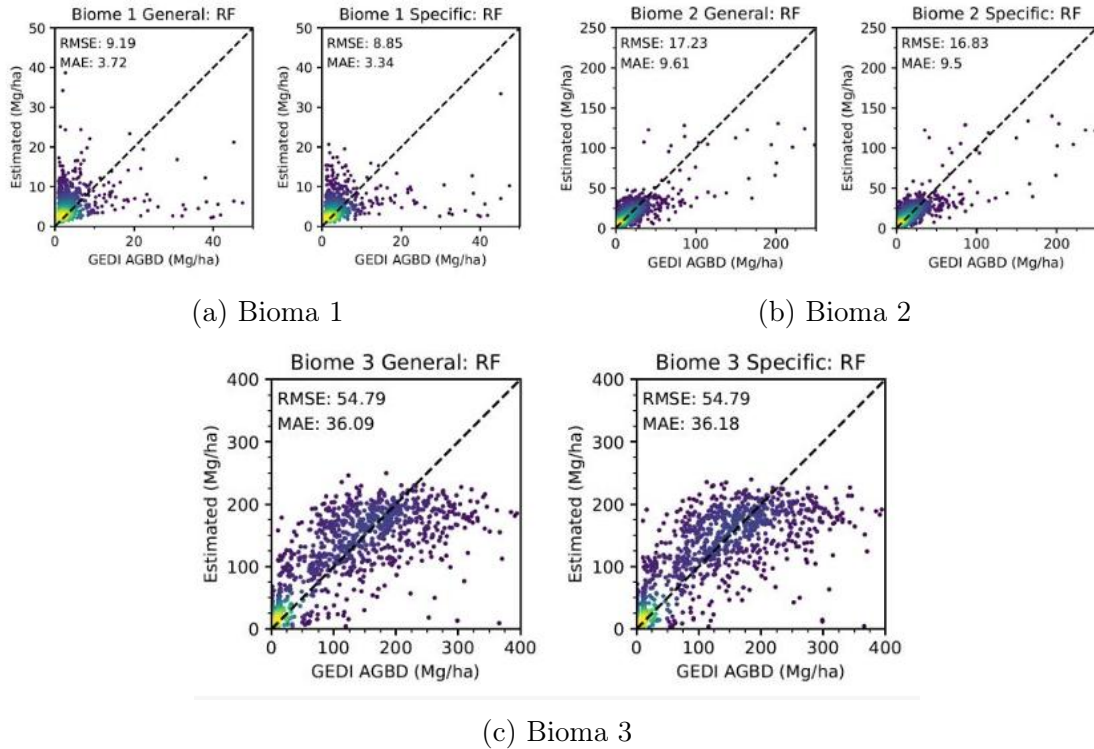


Figura 5 – Gráficos de dispersão de biomassa estimada e biomassa medida pelo GEDI para os três biomas

4.3 AVALIAÇÃO VISUAL E QUALITATIVA DOS MAPAS DE BIOMASSA GERADOS

A avaliação final dos resultados envolve a inspeção visual dos mapas de AGBD espacialmente contínuos, gerados pela aplicação dos modelos Random Forest (tanto específico quanto geral) a toda a área de estudo (SILVA et al., 2023). A comparação visual entre o mapa gerado pelo modelo específico de cada bioma e o mapa gerado pelo modelo geral revela diferenças muito sutis como pode ser visto na Figura 6. Mesmo em áreas distantes dos pontos de treinamento do GEDI, os padrões espaciais e a distribuição dos valores de biomassa preditos são notavelmente consistentes entre as duas abordagens. Isso sugere que, do ponto de vista prático e de interpretação visual, a escolha entre um modelo geral e um específico pode não resultar em diferenças significativas na aparência ou na distribuição espacial da biomassa mapeada. Esta observação é particularmente relevante para aplicações que dependem da análise de padrões espaciais, onde a robustez e a consistência do mapa são tão importantes quanto a acurácia pontual.

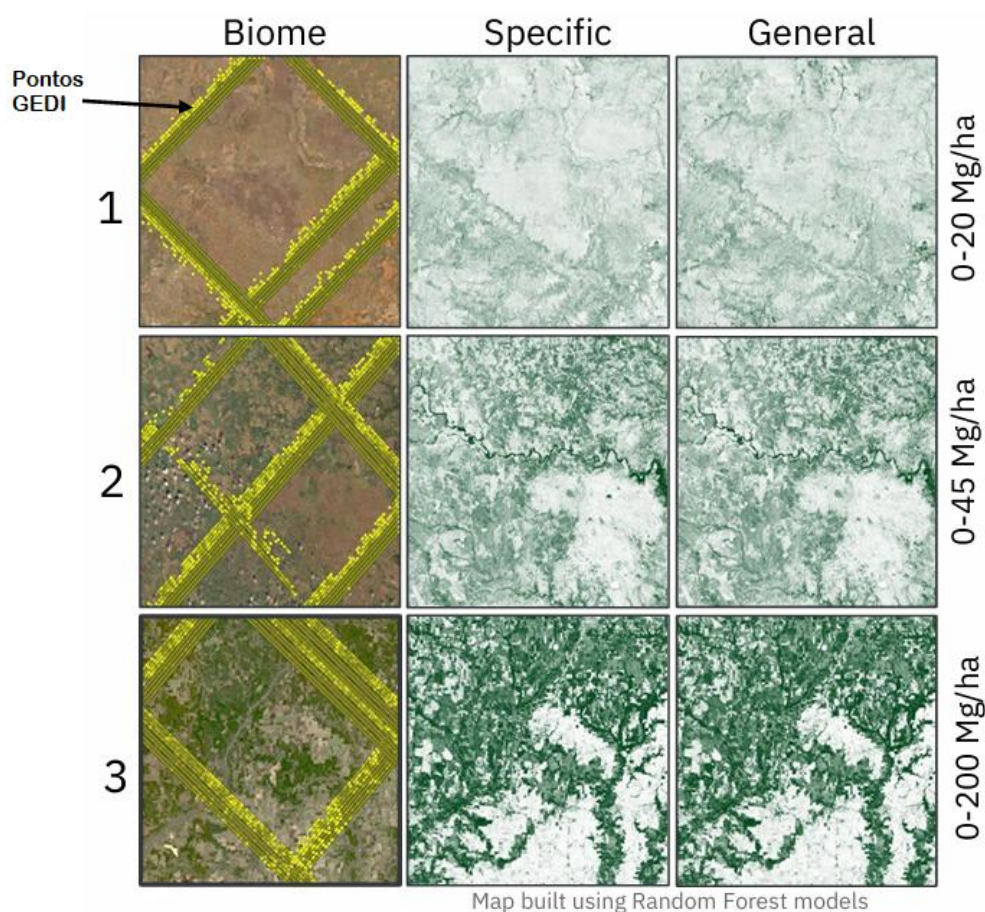


Figura 6 – Comparação visual de cada bioma com a imagem gerada por cada tipo de dado de treinamento

5 DISCUSSÃO E IMPLICAÇÕES PARA O MAPEAMENTO DE CARBONO NO BRASIL

5.1 SÍNTESE DOS ACHADOS: A VANTAGEM MARGINAL DOS MODELOS ESPECÍFICOS

A conclusão central do estudo experimental é que os modelos calibrados para biomas específicos tendem a ser, na maioria dos cenários, ligeiramente mais precisos do que um modelo geral treinado com um conjunto de dados unificado (SILVA et al., 2023). No entanto, a magnitude dessa vantagem é consistentemente pequena, com reduções no RMSE frequentemente inferiores a 1 Mg/ha . Este resultado, embora estatisticamente observável, levanta uma questão prática fundamental sobre o custo-benefício de cada abordagem quando se considera a implementação em larga escala. A decisão de adotar uma abordagem de modelos específicos implica um aumento significativo na complexidade operacional.

Para um país como o Brasil, isso significaria desenvolver, treinar, validar e manter pipelines de dados e modelos separados para a Amazônia, o Cerrado, a Mata Atlântica, a Caatinga, e outros biomas. Cada modelo exigiria seus próprios dados de referência, ajuste de hiperparâmetros e um sistema de inferência capaz de aplicar o modelo correto a cada pixel do território nacional.

Em contrapartida, um único modelo geral, embora marginalmente menos preciso em alguns casos, oferece uma solução muito mais simples, escalável e fácil de manter. Um modelo geral robusto, treinado em um conjunto de dados vasto e diversificado, pode até mesmo apresentar maior resiliência ao ser aplicado em zonas de transição entre biomas (ecótonos), onde a definição de um modelo “específico” é ambígua.

Portanto, para iniciativas de mapeamento em escala continental, como as conduzidas pelo MapBiomas ou pelo Instituto Nacional de Pesquisas Espaciais (INPE), o compromisso entre a máxima acurácia pontual e a escalabilidade operacional torna-se uma decisão estratégica de engenharia de sistemas. A simplicidade e a robustez de um modelo geral podem, na prática, superar os pequenos ganhos de precisão de uma abordagem fragmentada.

5.2 ANÁLISE CRÍTICA DOS MODELOS: POR QUE O RANDOM FOREST SUPEROU A U-NET NESTE CENÁRIO?

Os resultados mostraram uma clara superioridade dos modelos de ensemble baseados em árvores (Random Forest e XGBoost) sobre a arquitetura de deep learning U-Net. Uma análise superficial poderia concluir que as árvores de decisão são inerentemente melhores para esta tarefa. No entanto, uma análise mais profunda sugere que o resultado está

mais relacionado a um desalinhamento entre a arquitetura da U-Net e a forma como o problema foi formulado.

O problema foi estruturado como uma tarefa de regressão tabular: para cada ponto de referência do GEDI, um vetor de 14 características foi extraído do pixel correspondente nos dados Sentinel. O objetivo era mapear este vetor de entrada para um único valor de saída (AGBD). O Random Forest é um algoritmo excepcionalmente eficaz para este tipo de problema, lidando de forma robusta com características heterogêneas e relações não lineares (SHENDRYK, 2022; BREIMAN, 2001). A principal força da U-Net, e das CNNs em geral, reside na sua capacidade de aprender padrões espaciais complexos a partir do contexto local de um pixel, utilizando informações de vizinhança (texturas, formas, gradientes) capturadas pelos filtros convolucionais (RONNEBERGER; FISCHER; BROX, 2015; NATHANIEL et al., 2022). No entanto, no experimento realizado, a U-Net foi treinada com recortes de imagem (patches) de 100×100 pixels, mas a informação de referência (o valor de AGBD do GEDI) era um único ponto dentro desse patch. A função de perda (MSE) foi calculada comparando a predição de um único pixel de saída com o valor do GEDI. Isso significa que a rede não recebeu um feedback denso (um mapa de referência completo) que a orientasse a aprender as relações espaciais dentro do patch. Essencialmente, a arquitetura foi subutilizada, incapaz de alavancar sua principal vantagem — o aprendizado de contexto espacial — devido à natureza esparsa e pontual dos dados de treinamento.

O Random Forest, portanto, não venceu por ser uma arquitetura inerentemente superior, mas porque o problema foi enquadrado de uma maneira que se alinha perfeitamente com seus pontos fortes (regressão tabular ponto a ponto).

5.3 LIMITAÇÕES DA ABORDAGEM E O IMPACTO DA SATURAÇÃO EM FLORESTAS DENSAS

A limitação mais significativa da metodologia, explicitamente reconhecida no estudo de referência, é o aumento substancial do erro em áreas de alta biomassa (Biome 3) (SILVA et al., 2023). Este fenômeno é atribuído diretamente à saturação dos sinais dos sensores Sentinel-1 (banda C) e Sentinel-2 (óptico). Esta limitação tem implicações críticas para a aplicação da metodologia no Brasil, especialmente na Amazônia. O Bioma 3, usado como análogo de floresta densa no estudo, possui uma AGBD média de 94.6 Mg/ha . Em contraste, grandes extensões da Floresta Amazônica apresentam valores de AGBD que podem facilmente exceder $250\text{-}300 \text{ Mg/ha}$. Os sensores utilizados no estudo saturam em níveis de biomassa muito inferiores a esses, o que significa que eles se tornam “cegos” para as variações de carbono nas florestas mais maduras e com maiores estoques (NATHANIEL et al., 2022; IMHOFF, 1995; SAATCHI; DUNCANSON; KELLNER, 2015).

A aplicação direta desta metodologia na Amazônia resultaria, muito provavelmente,

em uma subestimação severa e sistemática da biomassa, comprometendo a acurácia do balanço de carbono nacional e a validade dos mapas para fins de políticas climáticas e mercados de carbono.

5.4 IMPLICAÇÕES PRÁTICAS PARA PROJETOS DE LARGA ESCALA

Apesar das limitações, o estudo oferece lições valiosas para projetos de monitoramento ambiental em larga escala no Brasil (e.g., MapBiomass, INPE).

- A validação do Random Forest como um algoritmo robusto e de alto desempenho para esta tarefa reforça sua utilização em projetos como o MapBiomass, que já emprega esta técnica extensivamente para a classificação de uso e cobertura da terra (JR. et al., 2020; ALENCAR et al., 2020).
- A pesquisa destaca a importância crítica de dados de referência de alta qualidade e com cobertura espacial para a calibração de modelos em escala nacional. A integração sistemática dos dados GEDI disponíveis para o território brasileiro nos fluxos de trabalho de projetos como o PRODES e o DETER, do INPE, poderia aprimorar significativamente a acurácia das estimativas de desmatamento e degradação, e permitir a transição para um monitoramento direto dos estoques de carbono (SHENDRYK, 2022).
- A discussão sobre modelos gerais versus específicos fornece um quadro analítico para que essas instituições avaliem suas estratégias de modelagem, ponderando o custo computacional e logístico de desenvolver modelos regionais em face dos ganhos de acurácia esperados.

6 CONCLUSÃO

6.1 RECAPITULAÇÃO DO PROBLEMA, METODOLOGIA E RESULTADOS

Este trabalho de conclusão de curso se propôs a realizar uma análise aprofundada sobre a quantificação de carbono florestal por meio de imagens de satélite e inteligência artificial, com foco na comparação entre abordagens de modelagem gerais e específicas por bioma. Baseando-se em um estudo de referência, foi explorada uma metodologia que funde dados de sensores ópticos (Sentinel-2), de radar (Sentinel-1) e LiDAR (GEDI) para treinar e avaliar diferentes algoritmos de regressão, incluindo Regressão Linear, Random Forest, XGBoost e a rede neural convolucional U-Net.

A análise dos resultados demonstrou que os modelos de ensemble baseados em árvores, como o Random Forest, apresentaram o desempenho mais robusto e preciso na maioria dos cenários avaliados. A principal conclusão do estudo indica que modelos treinados especificamente para as características de cada bioma oferecem uma vantagem de acurácia em relação a um modelo geral unificado, embora essa vantagem seja consistentemente marginal. A discussão aprofundada revelou as limitações críticas da metodologia, especialmente o efeito de saturação dos sensores em áreas de alta biomassa, e questionou a adequação da arquitetura U-Net para o problema, dada a natureza esparsa dos dados de referência.

6.2 CONSIDERAÇÕES FINAIS

A capacidade de monitorar os estoques de carbono da vegetação em escala global, com precisão e em tempo hábil, é uma das ferramentas mais poderosas que a ciência da computação e o sensoriamento remoto podem oferecer no combate à crise climática. Os resultados analisados neste trabalho demonstram tanto o imenso potencial da fusão de dados multi-sensor com algoritmos de inteligência artificial quanto os desafios que ainda persistem. A escolha entre modelos gerais e específicos, a seleção de arquiteturas de aprendizado de máquina adequadas e, acima de tudo, a superação das limitações físicas dos sensores em ecossistemas complexos como os do Brasil, são fronteiras ativas de pesquisa. A continuidade da investigação nestas áreas é fundamental para aprimorar a acurácia e a confiabilidade dos sistemas de monitoramento que, em última análise, subsidiarão as decisões globais de gestão ambiental e climática nas próximas décadas.

6.3 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Com base na análise crítica da metodologia e suas limitações, diversas avenidas de pesquisa podem ser exploradas para avançar no mapeamento de biomassa no Brasil.

- **Exploração de Outros Sensores:** Para superar a limitação da saturação em florestas densas como a Amazônia, é fundamental investigar o uso de sensores de radar com comprimentos de onda mais longos, como a banda L (e.g., do satélite ALOS-2 PALSAR-2 ou da futura missão NISAR). A banda L possui maior capacidade de penetração no dossel florestal, interagindo mais com troncos e galhos grossos, e, conseqüentemente, satura em níveis de biomassa muito mais elevados, tornando-a mais adequada para ecossistemas de alta AGBD (SAATCHI; DUNCANSON; KELLNER, 2015).
- **Adoção de Arquiteturas de Deep Learning Mais Avançadas:** Embora a U-Net não tenha se destacado neste estudo específico, o potencial do deep learning para aprender padrões espaciais complexos permanece vasto. Arquiteturas como a Feature Pyramid Network (FPN) são explicitamente projetadas para extrair e fundir características em múltiplas escalas de resolução, o que pode ser particularmente eficaz para capturar as diferentes texturas e padrões espaciais associados a diferentes níveis de biomassa, desde vegetação rasteira até florestas densas (LIN et al., 2017).
- **Aplicação e Validação no Contexto Brasileiro:** O passo mais importante é a aplicação e validação rigorosa desta e de outras metodologias avançadas diretamente nos diversos e complexos biomas brasileiros. Isso envolveria a utilização de dados de inventários florestais locais como validação adicional e a integração de dados de plataformas nacionais, como os mapas de uso e cobertura do MapBiomas ou os alertas de desmatamento do INPE (PRODES, DETER), que podem servir como variáveis preditoras adicionais ou como máscaras para estratificar a análise (ALENCAR et al., 2020).

REFERÊNCIAS

- ALENCAR, A. et al. Mapping three decades of changes in the brazilian savanna native vegetation using landsat data processed in the google earth engine platform. **Remote Sensing**, v. 12, n. 6, p. 924, 2020.
- ASNER, G. P. Cloud cover in landsat observations of the brazilian amazon. **International Journal of Remote Sensing**, v. 21, n. 18, p. 3855–3862, 2000.
- BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.
- BROWN, S. Measuring carbon in forests: current status and future challenges. **Environmental Pollution**, v. 116, n. 3, p. 363–372, 2002.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2016. p. 785–794.
- DUBAYAH, R. et al. The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. **Science of Remote Sensing**, v. 1, p. 100002, 2020.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York: Springer, 2009.
- IMHOFF, M. L. Radar backscatter and biomass saturation: Ramifications for global biomass inventory. **IEEE Transactions on Geoscience and Remote Sensing**, v. 33, n. 2, p. 511–518, 1995.
- JR., C. M. S. et al. Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. **Remote Sensing**, v. 12, n. 17, p. 2735, 2020.
- KACIC, P.; HIRNER, A.; PONTE, E. D. Fusing sentinel-1 and-2 to model gedi-derived vegetation structure characteristics in gee for the paraguayan chaco. **Remote Sensing**, v. 13, n. 24, p. 5105, 2021.
- LIN, T.-Y. et al. Feature pyramid networks for object detection. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. p. 2117–2125.
- NATHANIEL, J. et al. Aboveground carbon biomass estimate with physics-informed deep network. **arXiv preprint arXiv:2210.13752**, 2022.
- RIBEIRO, K. G.; MACHADO, V. L. L.; FILHO, M. D. L. A. B. d. S. Seasonality in insect abundance in the "cerrado" of goiás state, brazil. **Revista Brasileira de Entomologia**, v. 53, n. 1, p. 94–99, 2009.
- ROBERTS, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. **Ecography**, v. 40, n. 8, p. 913–929, 2017.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015**. [S.l.]: Springer International Publishing, 2015. p. 234–241.

SAATCHI, S. S.; DUNCANSON, L. E.; KELLNER, J. R. The saturation of l-band radar backscatter to forest aboveground biomass. **Remote Sensing of Environment**, v. 166, p. 152–165, 2015.

SHENDRYK, Y. Fusing gedi with earth observation data for large area aboveground biomass mapping. **International Journal of Applied Earth Observation and Geoinformation**, v. 115, p. 103108, 2022.

SILVA, A. F. da et al. Comparison of biome-specific ai models to estimate biomass. In: **IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium**. Pasadena, CA, USA: [s.n.], 2023. p. 1–4.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. **Economic Geography**, v. 46, p. 234–240, 1970.

XIAO, J. et al. Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. **Remote Sensing of Environment**, v. 233, p. 111383, 2019.