Predição dos parâmetros de qualidade e quantidade de forragem

Samuel Rodrigues[1], Ricardo Santos[1], Guilherme Defalque[1], João Serrano[2]

[1]Faculdade de Computação – Universidade Federal de Mato Grosso do Sul (UFMS)
Campo Grande/MS – Brazil

[2]MED - Mediterranean Institute for Agriculture, Environment and Development
and CHANGE - Global Change and Sustainability Institute
Universidade de Évora, Évora — Portugal

{samuel_rodrigues, ricardo.santos, guilherme.defalque}@ufms.br

jmrs@uevora.pt

Abstract. *This work presents the design and development of predictive models for the parameters pasture moisture content (PMC), crude protein (CP), neutral detergent fiber (NDF), green matter (GM), and dry matter (DM). The predictors were built using an extensive dataset composed of spectral data from satellite platforms such as Landsat-8, Sentinel-2, and MODIS, combined with climatic variables from eight regions. The study also presents the methodology employed, which includes techniques such as feature selection, PCA, and GridSearchCV, achieving adjusted* R^2 *values of up to* 63% *for moisture content,* 60% *for crude protein,* 24% *for neutral detergent fiber,* 22% *for green matter, and* 34% *for dry matter.*

Resumo. *Este trabalho apresenta o projeto para o desenvolvimento de modelos preditivos para os parâmetros teor de umidade (PMC), proteína bruta (CP), fibra em detergente neutro (NDF), matéria verde (GM) e matéria seca (DM). Os preditores foram construídos com base em um extenso dataset de dados espectrais de satélites, como Landsat-8, Sentinel-2 e MODIS, juntamente com variáveis climáticas de oito regiões. O trabalho apresenta ainda a metodologia empregada, por meio de técnicas como feature selection, PCA, GridSearchCV e obtendo resultados de* R^2 *ajustado de até* 63% *para teor de umidade,* 60% *para proteína bruta,* 24% *para fibra em detergente neutro,* 22% *para matéria verde e* 34% *para matéria seca.*

1. Introdução

A agropecuária mundial, diante do rápido aumento populacional, enfrenta desafios crescentes com a demanda por proteína animal. No Brasil, a engorda baseada em pastagens é predominante, e o desempenho do gado depende diretamente da qualidade e quantidade

da forragem disponível. Isso porque, pastagens em quantidade suficiente e com maior qualidade são requisitos essenciais para obter maior aproveitamento dos nutrientes, maiores taxas de crescimento e ganho de peso [Senar, 2018].

A qualidade da pastagem pode ser inferida por meio de parâmetros químicos que caracterizam a composição nutricional da forragem. A proteína bruta (CP) e a fibra em detergente neutro (NDF) se destacam como métricas para estimar a qualidade da forragem, ao passo que níveis elevados de proteína bruta e baixos teores de fibras representam pastagens de alta qualidade [SERRANO et al., 2024]. Essas características são típicas de plantas mais jovens, nas quais a coloração verde e a maior presença de folhas são predominantes. Além disso, o teor de umidade de pastagem (PMC) representa a porcentagem de água presente no material forrageiro e está diretamente relacionado à sua qualidade. Uma alta porcentagem de umidade é característica de pastagens com mais proteína e digestíveis, o inverso sinaliza maior teor de matéria seca e de fibras, que, em excesso, reduzem a digestibilidade e o valor energético para os animais. Além disso, parâmetros como a matéria seca (DM) e a matéria verde (GM) sinalizam a quantidade de forragem disponivel, refletindo diretamente a produtividade da pastagem.

Tendo em vista a necessidade de monitorar a qualidade das pastagens de forma contínua, o sensoriamento remoto tem se mostrado uma ferramenta eficaz para extrair informações sobre as condições de uma determinada área, de forma não invasiva, reduzindo a necessidade de medições a campo de alto custo. Estudos [DEFALQUE et al., 2024; BRETAS et al., 2021] mostram que dados espectrais provenientes de satélites, como o *Landsat-8*, *Sentinel-2* e *MODIS*, em conjunto com técnicas de aprendizado de máquina, são eficazes para predizer, com alta precisão, parâmetros de qualidade e de quantidade de forragem. Nesse contexto, o objetivo desse trabalho é combinar dados espectrais e variáveis climáticas de diversas regiões ao longo de um amplo intervalo temporal, visando ao projeto de modelos preditivos para os parâmetros de CP, NDF, PMC, DM e GM.

2. Materiais e Métodos

2.1. Áreas Sob Estudo

Serrano [SERRANO et al., 2024] conduziu um estudo experimental durante os ciclos de crescimento vegetativo de pastagens nos períodos de 2018/2019, 2019/2020 e 2020/2021 em oito regiões distintas de Portugal e Espanha. Seis regiões (AZI, GRO, MIT, MUR, PAD e TAP) situam-se no sul do Alentejo, uma (QF) na região de Beira Interior, em Portugal, e outra na província espanhola de Extremadura (CUB), como ilustrado na Figura 1.

As amostras de pastagem foram coletadas entre janeiro e maio para o ciclo de 2019; de janeiro a junho (exceto fevereiro) e dezembro no ciclo de 2020; e de fevereiro

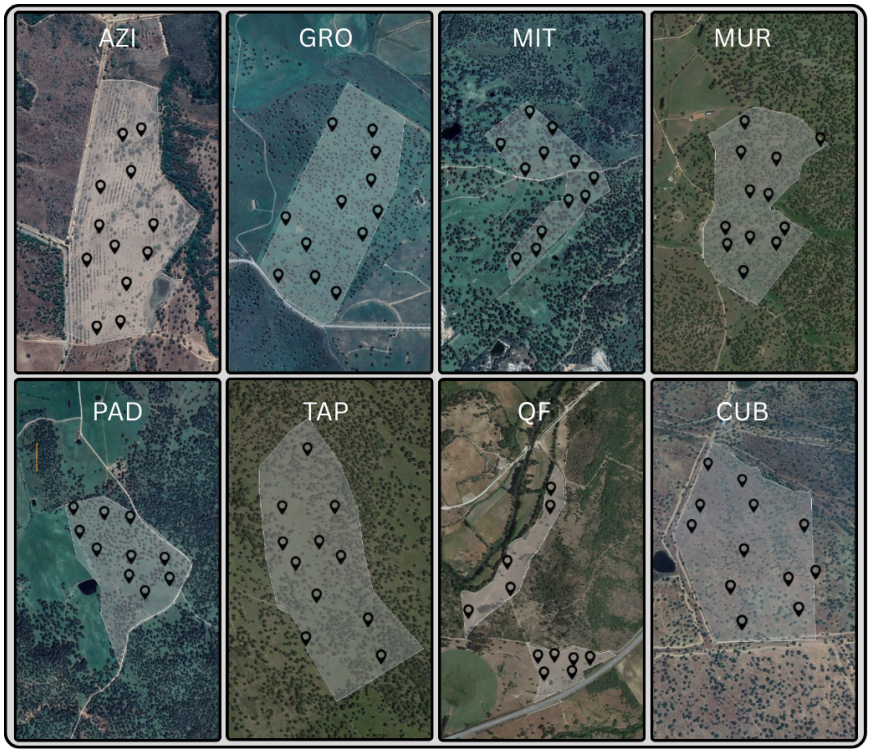


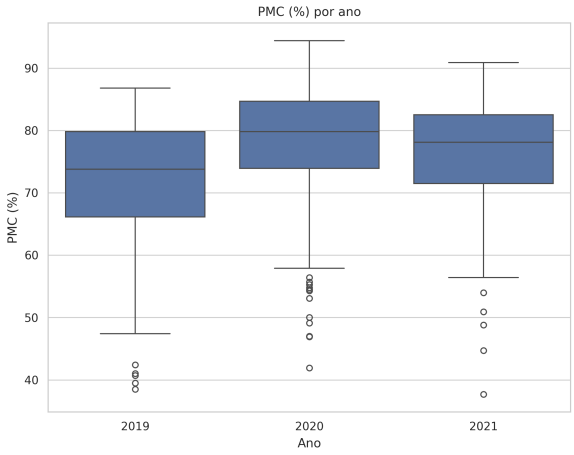
Figura 1. Imagens das áreas estudadas e localização das amostras coletadas.

a maio no ciclo de 2021. Utilizou-se um quadrante metálico (0,5 m × 0,5 m) e tesouras elétricas de precisão para a coleta de 10 a 12 amostras compostas por cinco subamostras cada. Além disso, foram utilizados dispositivos especializados para registrar as coordenadas geográficas de cada amostra. Posteriormente, as amostras foram submetidas à análise laboratorial, empregando métodos analíticos padrão (AOAC) [HORWITZ; LATIMER, 2005]. No laboratório, as amostras foram pesadas para determinar GM, em seguida, foram secadas em estufa a 65 °C durante 72 horas e pesadas novamente para determinar o PMC e DM. Para determinar CP, mediu-se o nitrogênio da amostra pelo método Kjeldahl, através de leitura colorimétrica em um autoanalisador Bran and Luebbe, em seguida, o valor de nitrogênio foi então convertido em proteína bruta usando o fator 6,25 (CP = nitrogênio medido x 6,25). O teor de NDF foi determinado segundo o método de Goering e Van Soest, por meio de um digestor de fibras (Foss Tecator).

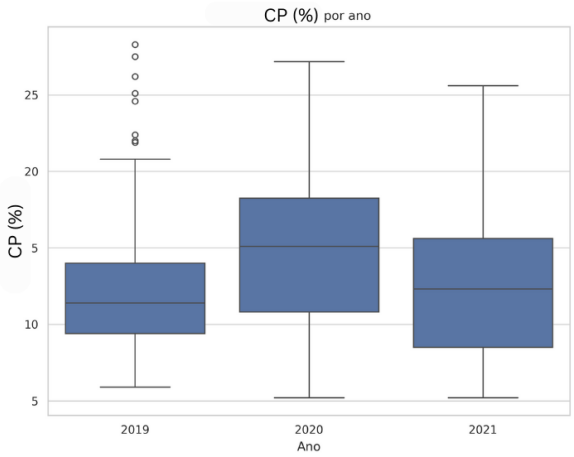
2.2. Análise estatística dos parâmetros

O objetivo deste trabalho é desenvolver modelos de aprendizado de máquina para as variáveis de qualidade e quantidade de forragem. Para fundamentar essa modelagem, realizou-se um estudo considerando tanto a variabilidade dos dados ao longo dos três anos de coleta, como as diferenças espaciais entre as regiões estudadas. Essa etapa de caracterização dos dados é fundamental para compreender a estrutura dos dados e determinar as abordagens mais adequadas para projeto dos modelos. As Figuras 2 e 3 apresentam os *boxplots* de cada variável-alvo ao longo dos anos, permitindo avaliar tendências,

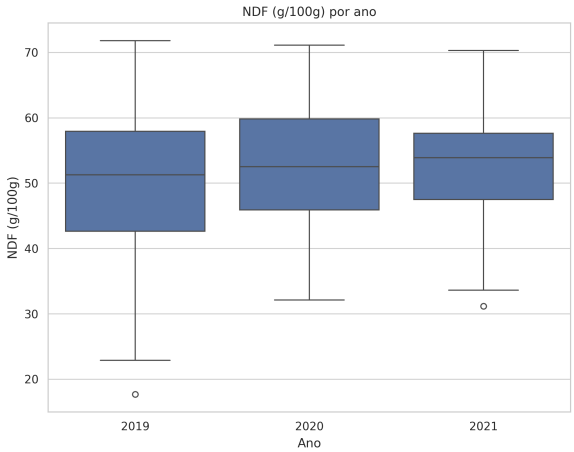
estabilidade ou oscilações entre os anos. Complementarmente, as Figuras 4, 5, 6, 7 e 8 apresentam a distribuição dessas mesmas variáveis entre as regiões estudadas, permitindo identificar possiveis padrões espaciais.



(a) *Boxplot de PMC por ano*

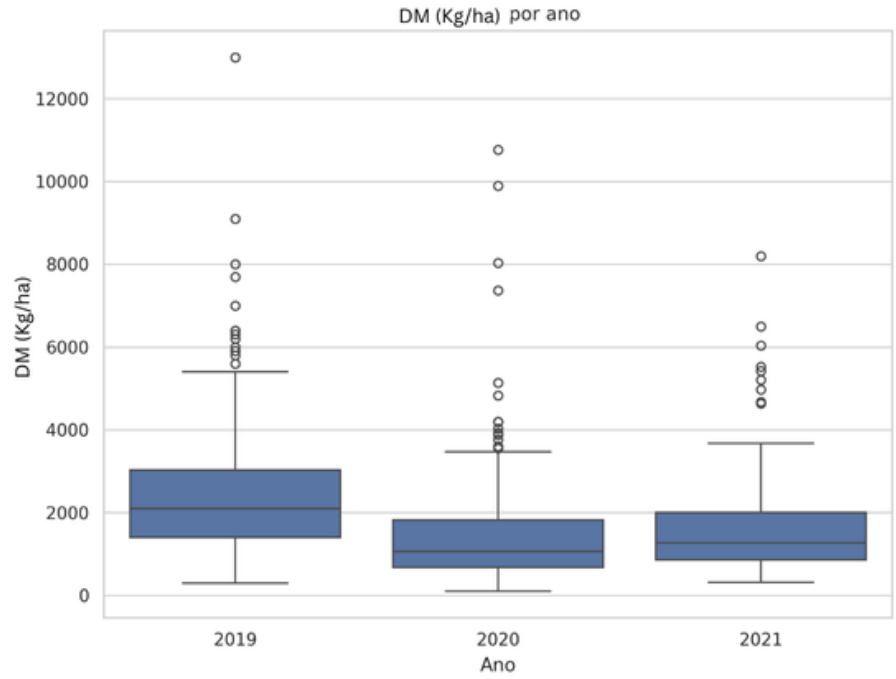


(b) Boxplot de CP por ano

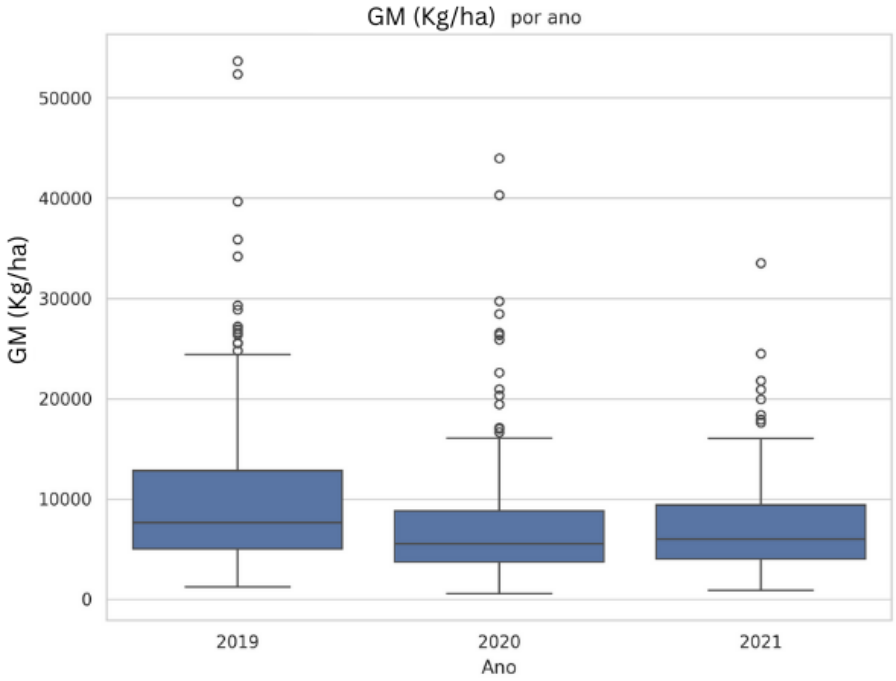


(c) Boxplot de NDF por ano

Figura 2. *Boxplots* para os parâmetros de qualidade de forragem por ano

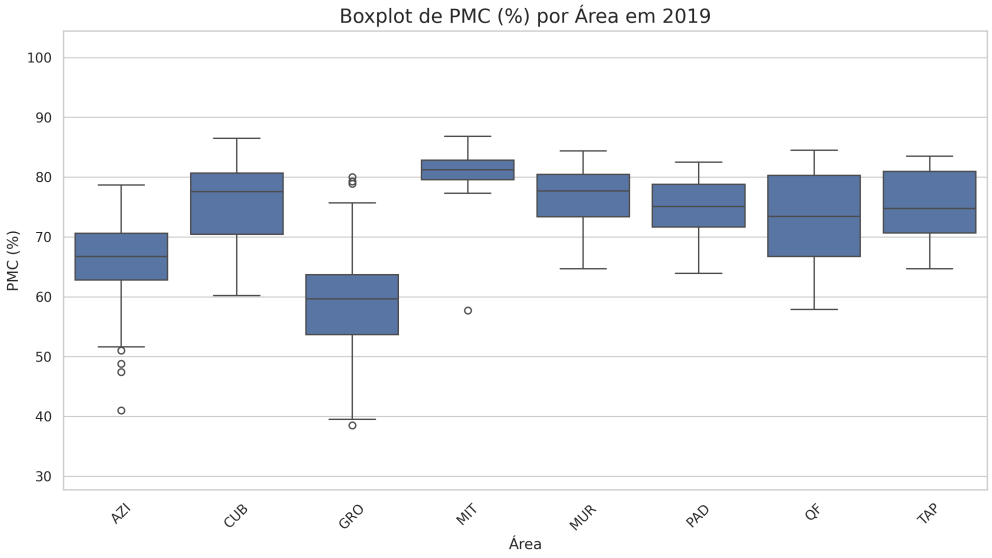


(a) Boxplot de DM por ano

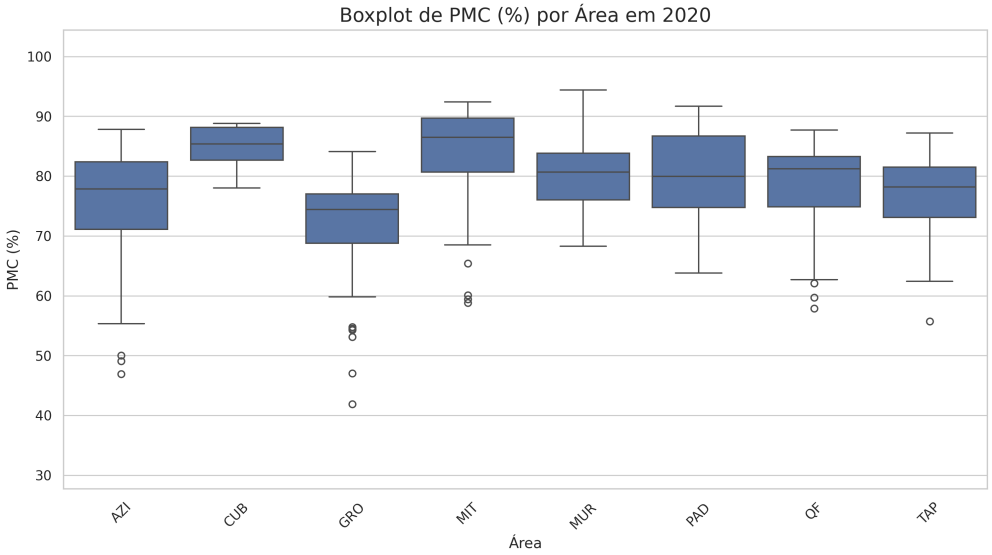


(b) Boxplot de GM por ano

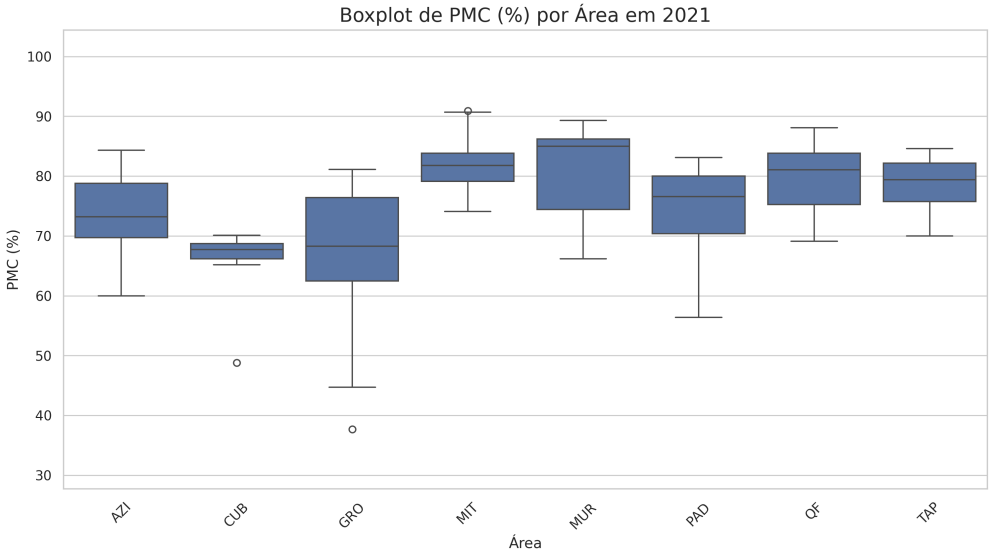
Figura 3. *Boxplots* para os parâmetros de quantidade de forragem por ano



(a) *Boxplot de PMC para cada região em 2019*

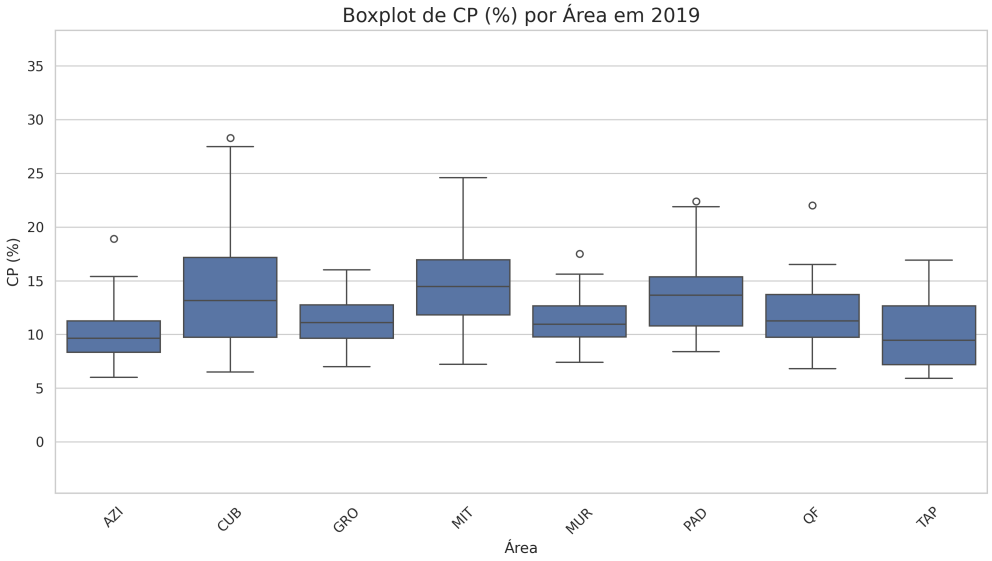


(b) *Boxplot de PMC para cada região em 2020*

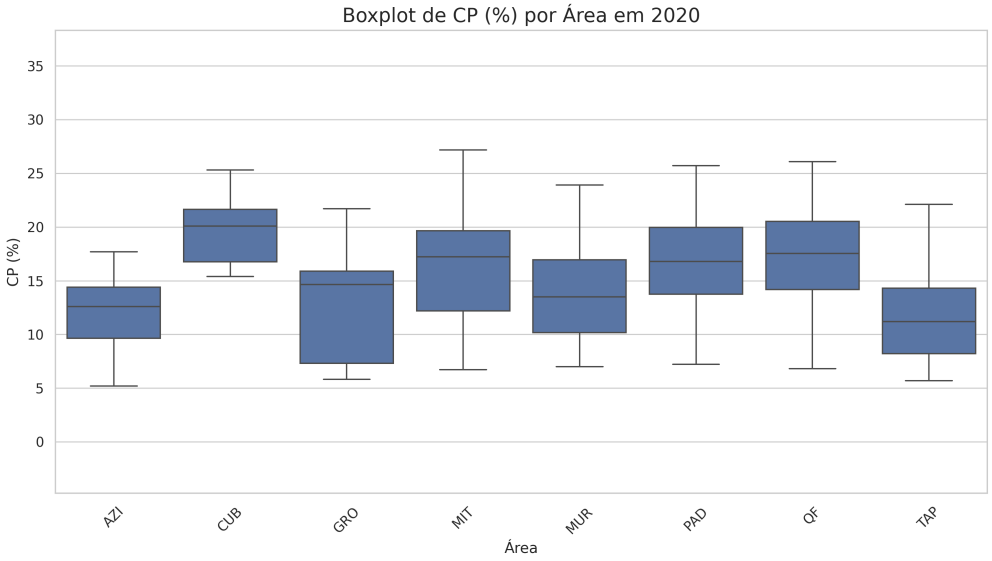


(c) *Boxplot de PMC para cada região em 2021*

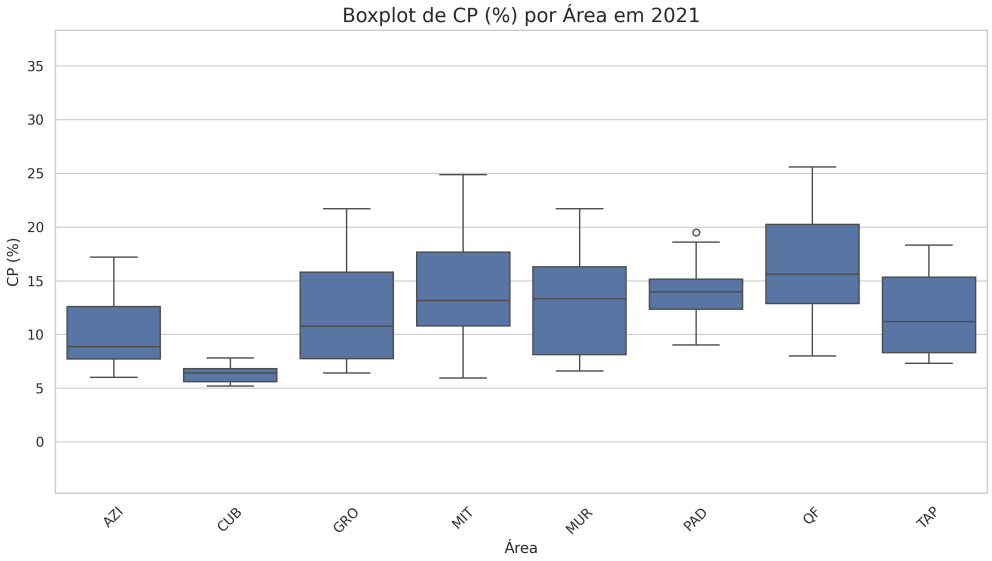
Figura 4. *Boxplots* de cada região para o parâmetro PMC



(a) Boxplot de CP para cada região em 2019

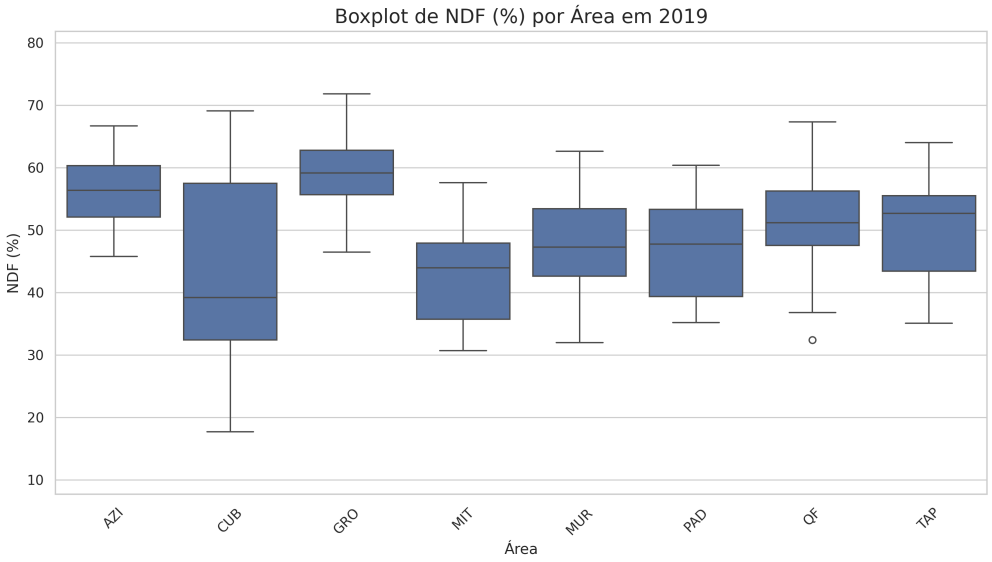


(b) Boxplot de CP para cada região em 2020

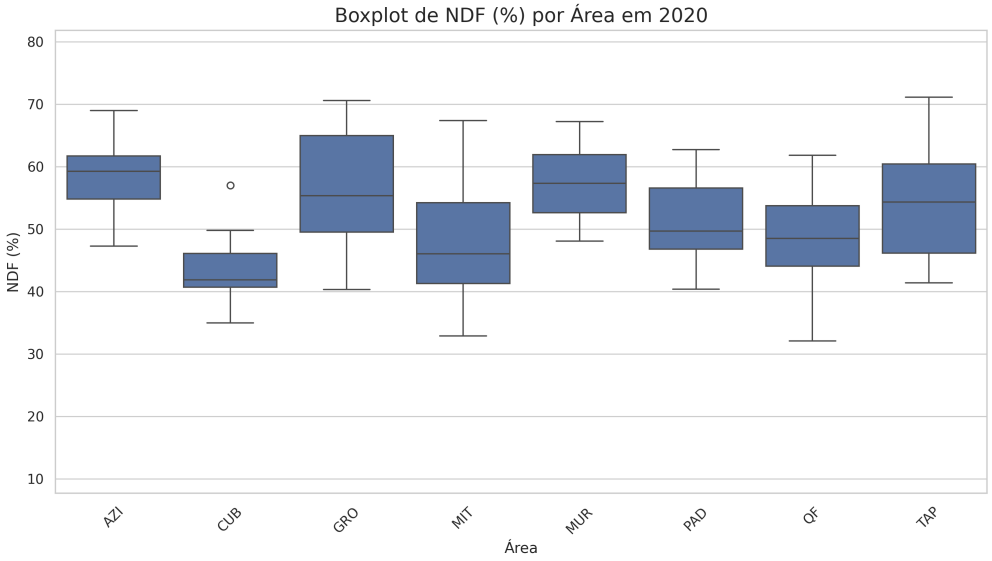


(c) Boxplot de CP para cada região em 2021

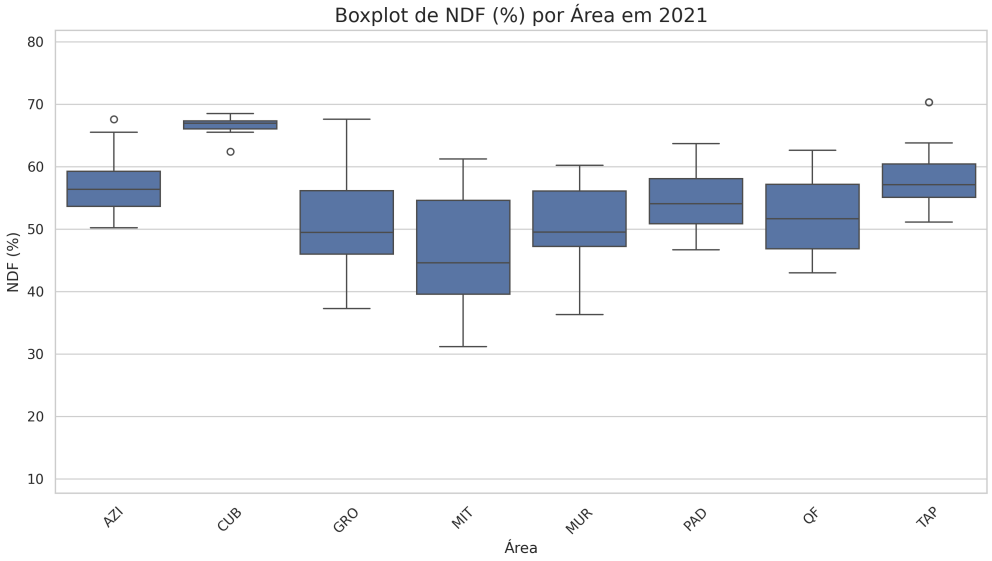
Figura 5. *Boxplots* de cada região para o parâmetro CP



(a) Boxplot de NDF para cada região em 2019

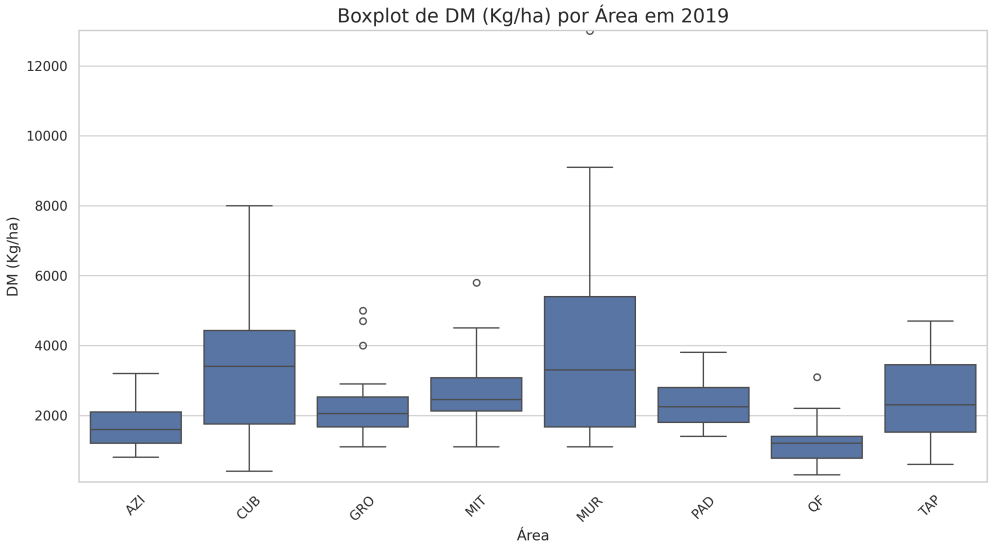


(b) Boxplot de NDF para cada região em 2020

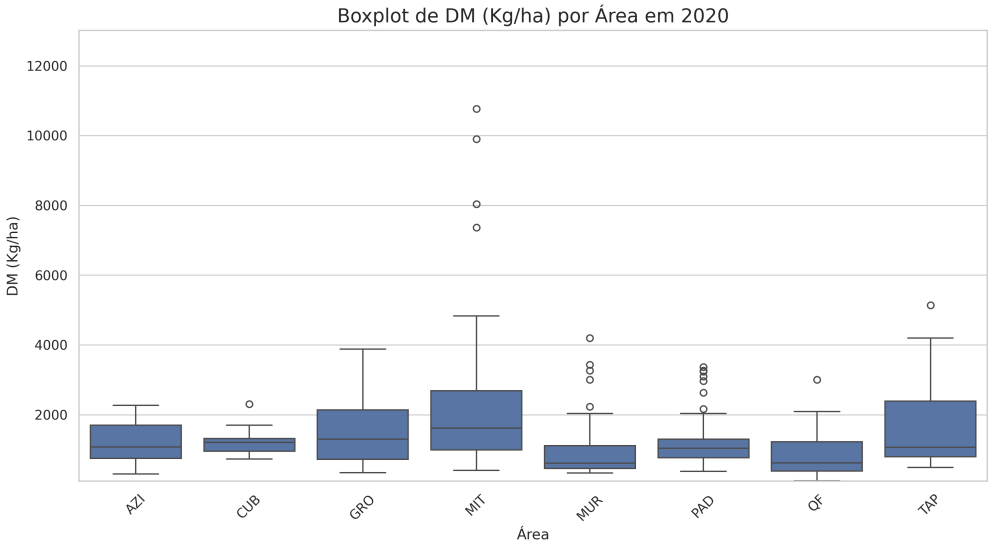


(c) Boxplot de NDF para cada região em 2021

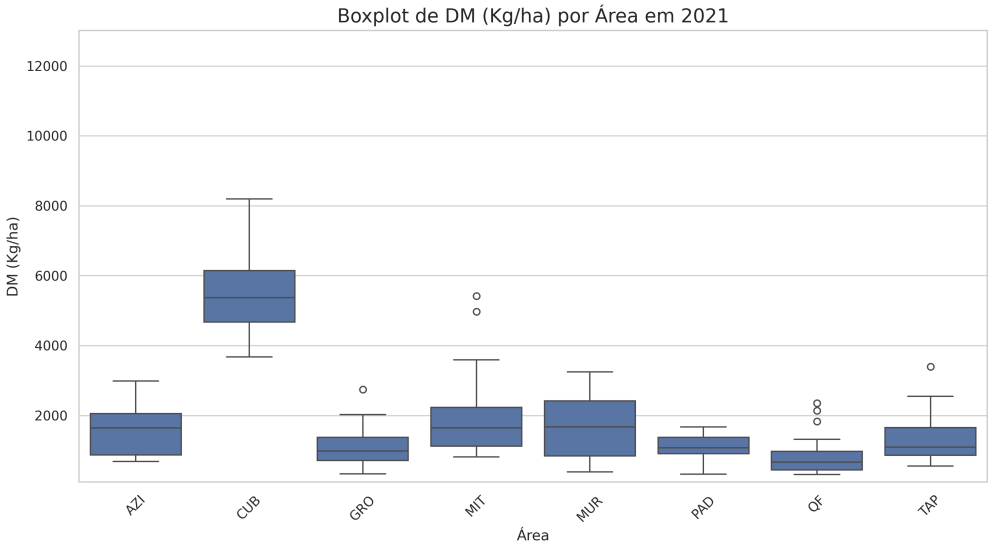
Figura 6. *Boxplots* de cada região para o parâmetro NDF



(a) Boxplot de DM para cada região em 2019

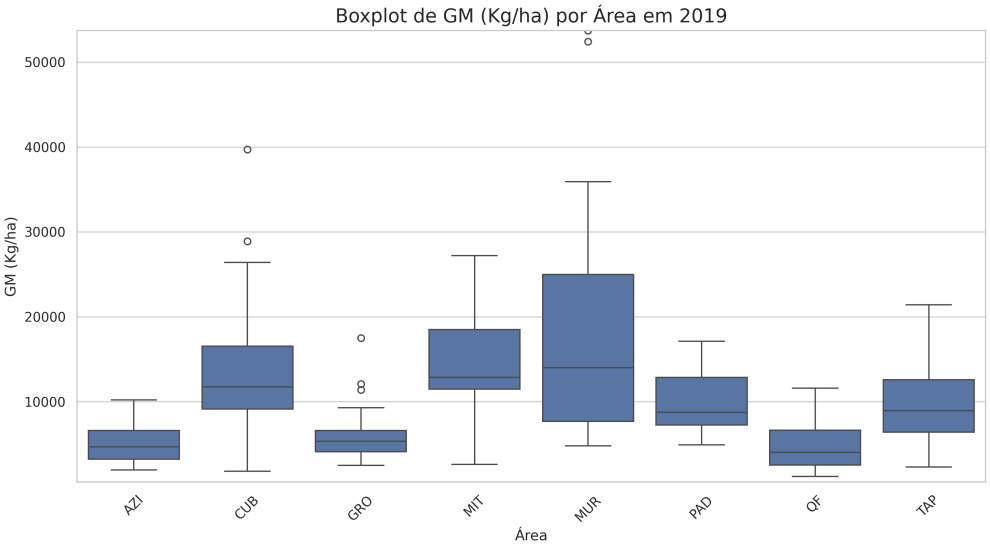


(b) Boxplot de DM para cada região em 2020

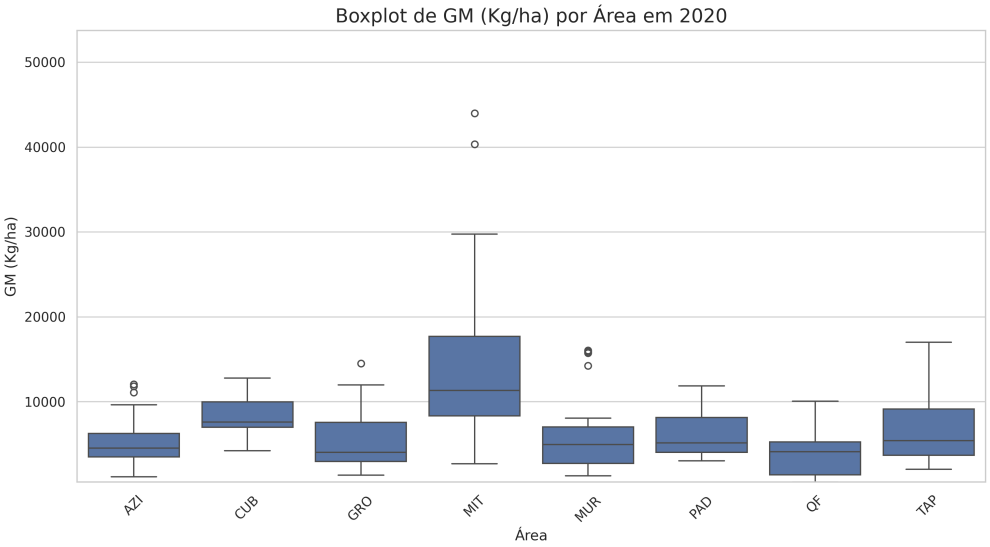


(c) Boxplot de DM para cada região em 2021

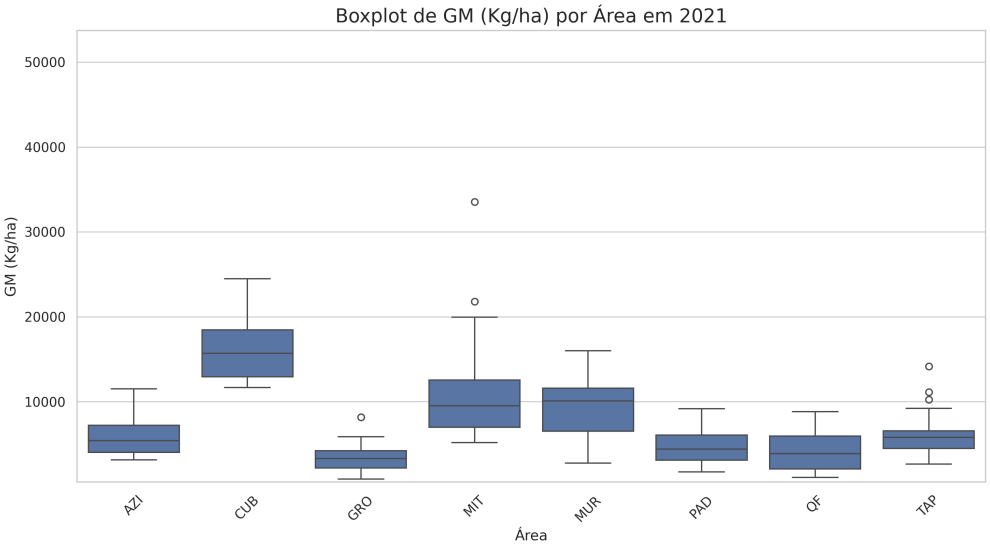
Figura 7. *Boxplots* de cada região para o parâmetro DM



(a) Boxplot de GM para cada região em 2019



(b) Boxplot de GM para cada região em 2020



(c) Boxplot de GM para cada região em 2021

Figura 8. *Boxplots* de cada região para o parâmetro GM

2.3. Construção do *Dataset*

2.3.1. Extração das Bandas e Variáveis Climáticas

Para a construção do *dataset* denominado MontadoDB [RODRIGUES et al., 2025], procedeu-se à coleta das bandas espectrais de diversos satélites orbitais com múltiplas coleções de imagens. Os satélites selecionados foram *Landsat-8*, *Sentinel-2* e *MODIS*, com coleções de reflectância de superfície, de topo da atmosfera e imagens sem processamento, com objetivo de obter o melhor de cada satélite. As bandas foram extraídas de forma automatizada por meio de algoritmos computacionais, em *Python*. Além disso, variáveis climáticas referentes ao dia da coleta também foram extraídas a partir das coordenadas geográficas de cada amostra.

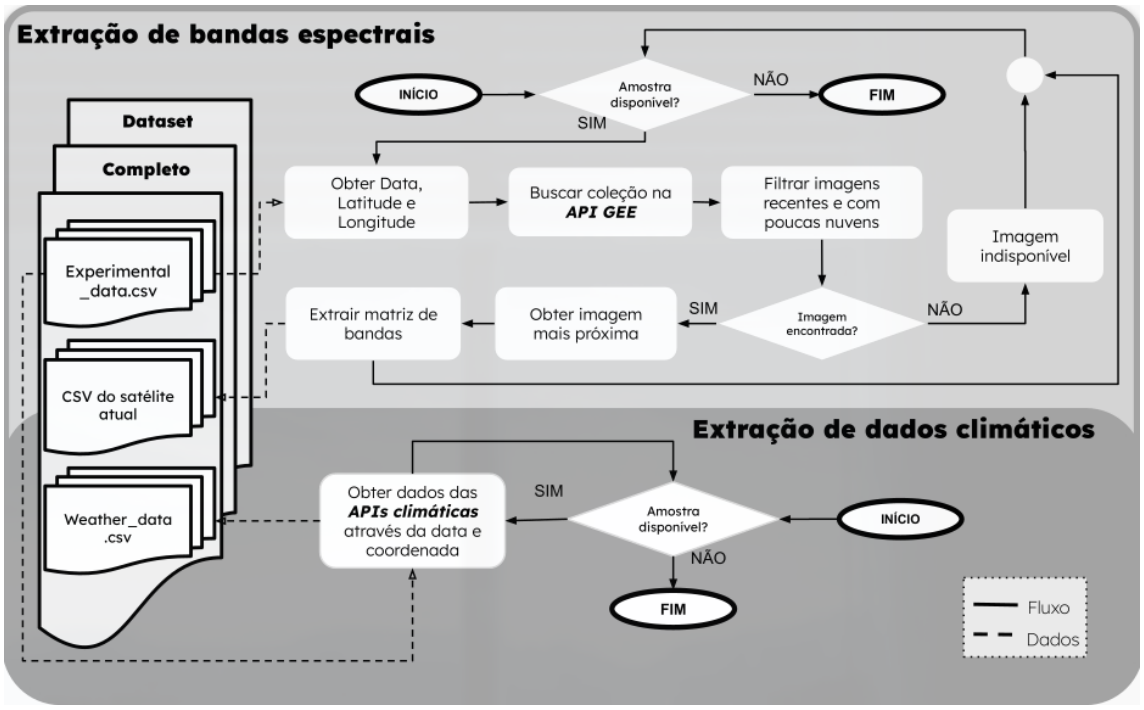


Figura 9. Diagrama detalhando o método de extração de bandas e de dados climáticos.

A Figura 9 ilustra as etapas do algoritmo para a obtenção das bandas e das variáveis climáticas. Para a extração de bandas espectrais, a partir da data e do *shapefile* das regiões, a API do *Google Earth Engine* (GEE) [GORELICK et al., 2017] foi utilizada para obter a coleção de imagens com intervalos temporais próximos à data de coleta, juntamente com filtros para remover imagens cuja visibilidade esteja comprometida por alta quantidade de nuvens. Se houver disponibilidade de imagens, é escolhida a imagem mais próxima no tempo e as bandas de toda a região são extraídas. Com o objetivo de incluir variáveis altamente correlacionadas com os parâmetros de interesse, a partir das bandas já obtidas, também foram calculados e adicionados ao *dataset* índices de vegetação como:

Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Enhanced Vegetation Index (EVI) e Leaf Area Index (LAI), entre outros. Os índices de vegetação são equações que combinam duas ou mais bandas com o objetivo de avaliar a vegetação e suas propriedades. Para a extração de dados climáticos, as variáveis foram obtidas por meio da API *Open-Meteo* [ZIPPENFENIG, 2023], com base na data e nas coordenadas geográficas das amostras.

2.3.2. *Pooling* Adaptativo 2D

Os dados utilizados para treinamento e teste dos modelos de predição são derivados de diversos satélites, cada qual com dimensões espaciais distintas. Para garantir maior homogeneidade dos dados, aplicou-se a técnica de *pooling* adaptativo 2D. Inicialmente, extraem-se as bandas espectrais de toda a região. Em seguida, define-se a resolução final da região para aplicação do *pooling*. Foram testadas cinco resoluções (5x5, 10x10, 15x15, 20x20 e 25x25), onde cada uma resultou em um *dataset* diferente. Por fim, foi calculada a média de todos os valores, resultando em um valor representativo para as bandas e para cada variável-alvo. Esse processo garante consistência espacial entre os dados espectrais utilizados no treinamento dos modelos.

2.3.3. *Data Augmentation*

Data augmentation é um conjunto de técnicas utilizadas para aumentar ou transformar o *dataset* a partir dos dados existentes, com o objetivo de melhorar o desempenho e a capacidade de generalização dos modelos de aprendizado de máquina [WANG et al., 2025]. Em situações onde a quantidade de dados é escassa, como no presente estudo que possui um número reduzido de amostras por região, o modelo tem dificuldade de generalização, resultando em sobreajuste (*Overfitting*). Nesses casos, técnicas de *Data augmentation* tornam-se essenciais para obter um conjunto de dados com uma representação mais contínua e explicativa.

Dentre as abordagens disponíveis, a interpolação SPLINE se destaca por enriquecer os dados de maneira coerente, sem introduzir valores aleatórios ou ruídos excessivos. Diferente da interpolação polinomial padrão, que interpola todo o conjunto de dados aproximando de uma única função de ordem elevada, a técnica SPLINE interpola o conjunto de dados por meio de polinômios de graus menores entre as subseções do conjunto, normalmente um polinômio cúbico, que resulta em curvas mais suaves e melhor explicação da tendência dos dados.

A Figura 10 ilustra a aplicação da técnica para a região AZI e parâmetro CP, evidenciando coerência na representação dos dados, onde a tendência se segue durante

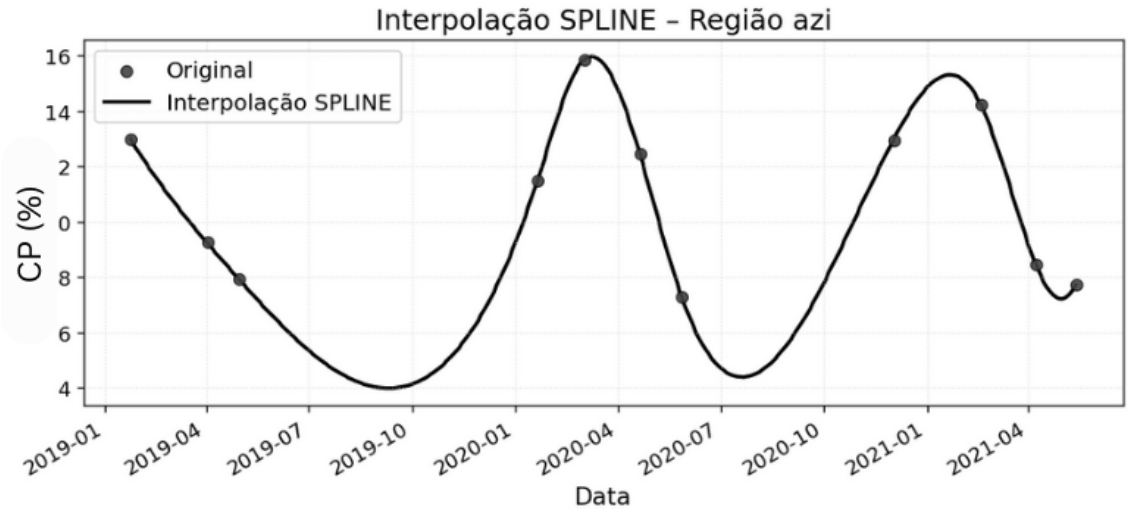


Figura 10. Interpolação SPLINE para a região AZI e parâmetro CP

todo o intervalo de coleta.

2.4. Projeto dos Modelos de Predição

A Figura 11 ilustra o *pipeline* desenvolvido para o projeto de modelos, destacando as etapas de pré-processamento e de treinamento. Na construção do *dataset*, obteve-se uma grande quantidade de variáveis independentes, muitas delas não apresentam características significativas para explicar as variáveis-alvo e apresentam grande correlação entre si (colinearidade).

Com o objetivo de reduzir essas variáveis, aplicou-se *feature selection*, etapa responsável por selecionar, por meio de técnicas como *Kbest* e *Recursive Feature Elimination* (RFE), ou por correlação de *spearman*, as variáveis mais correlacionadas com cada variável-alvo, além de remover a colinearidade entre as variáveis. Na etapa de transformação, aplicou-se a técnica *Principal Component Analysis* (PCA) que transforma o conjunto de variáveis originais em novas variáveis independentes que preservam a variância dos dados. Essa abordagem torna o conjunto de dados mais estável ao reduzir a dimensão.

Para o treinamento dos modelos foram selecionados diversos algoritmos de aprendizado de máquina amplamente utilizados na literatura e consolidados na área. Entre eles, destacam-se o *Linear Regression*, o *Support Vector Regression* (SVR) [DRUCKER et al., 1996], o *Random Forest* [BREIMAN, 2001], o *XGBoost* [CHEN et al., 2025], o *Gradient Boosting* [FRIEDMAN, 2001] e modelos baseados em redes neurais, como *Multilayer Perceptron* e rede pré-treinada para tarefas tabulares, o *TabPFN* [HOLLMANN et al., 2023]. Com exceção dos modelos neurais, uma busca extensiva pelos hiperparâmetros de cada algoritmo foi conduzida, utilizando técnicas de otimização com validação cruzada, como *GridSearchCV*.

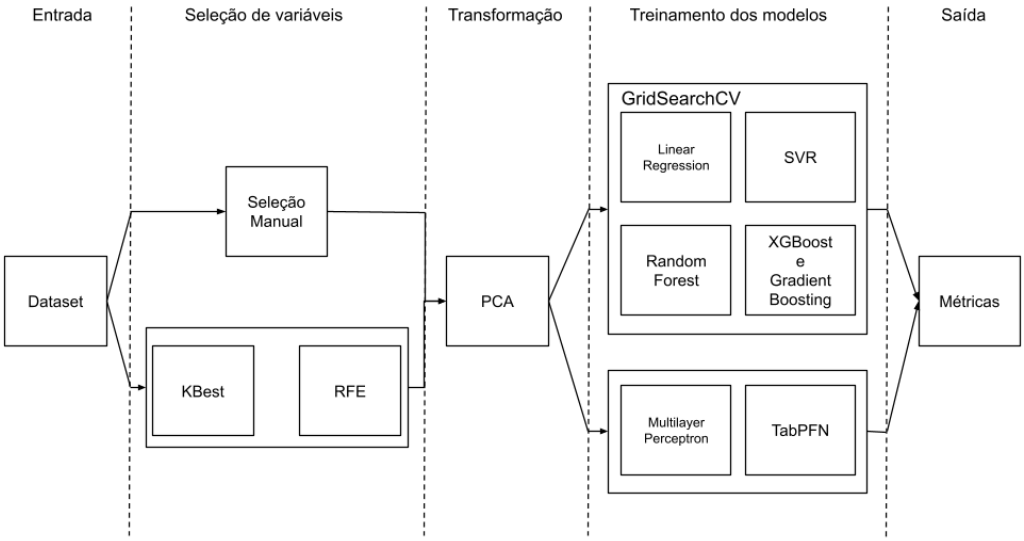


Figura 11. Pipeline utilizado para projeto dos modelos.

Devido à natureza temporal dos dados, a metodologia de estratificação temporal foi utilizada. Primeiramente, o conjunto de dados é dividido de modo que os anos de 2019 e 2020 sejam utilizados para o treinamento dos modelos, enquanto as amostras de 2021 ficam reservadas para teste. Foi então avaliado sistematicamente diversas configurações das técnicas mencionados, como a presença ou ausência do *pooling* adaptativo 2D e das técnicas de interpolação. As melhores configurações foram então selecionadas com base em comparações quantitativas do desempenho dos modelos no conjunto de teste.

Para avaliar o desempenho dos modelos, utilizou-se a métrica de validação por coeficiente de determinação ajustado (R^2 ajustado), uma variação da métrica R^2 que penaliza a inclusão de variáveis irrelevantes.

3. Resultados e Discussão

As Tabelas 1, 2, 3, 4 e 5 apresentam os valores de R^2 ajustado e RMSE/MAPE obtidos pelos modelos que melhor explicaram a variância dos dados para cada parâmetro avaliado. Além dos resultados quantitativos, a tabela também apresenta a configuração utilizada para o treinamento de cada modelo, detalhando a dimensão do *pooling* (ou sua ausência) e a abordagem empregada para a seleção de variáveis.

Tabela 1. Resultados para o parâmetros PMC

Pooling	**Seleção de Variáveis**	$R^2_{\textbf{aj.}}$ **(%)**	**RMSE**	**Modelo**
5×5	KBest + RFE	63	4.38	Linear Regression
5×5	Manual (Spearman)	60	4.64	TabPFN
15×15	Manual (Spearman)	57	4.86	Random Forest

Tabela 2. Resultados para o parâmetros CP

Pooling	Seleção de Variáveis	$R^2_{\text{aj.}}$ (%)	RMSE	Modelo
10×10	Manual (Spearman)	60	2.49	TabPFN
20×20	Manual (Spearman)	56	2.85	Gradient Boosting
5×5	KBest + RFE	54	2.89	Linear Regression

Tabela 3. Resultados para o parâmetros NDF

Seleção de Variáveis	$R^2_{\text{aj.}}$ (%)	RMSE	Modelo
KBest + RFE	24	51.55	Linear Regression
KBest + RFE	21	52.78	Gradient Boosting

Tabela 4. Resultados para o parâmetros GM

Pooling	Seleção de Variáveis	$R^2_{\text{aj.}}$ (%)	RMSE	MAPE (%)	Modelo
N/A	Manual	22	N/A	73.50	MultiLayer Perceptron
25×25	KBest + RFE	20	3284.42	N/A	XGBoost

Tabela 5. Resultados para o parâmetros DM

Seleção de Variáveis	$R^2_{\text{aj.}}$ (%)	MAPE (%)	Modelo
Manual	34	65.91	MultiLayer Perceptron

Para o PMC (Tabela 1), o modelo *Linear Regression* superou outros modelos mais complexos, com R^2 ajustado de 63%. Isso sugere que a relação entre as bandas espectrais e variáveis climáticas com PMC seja fortemente linear, tornando o uso de modelos mais complexos desnecessário. Em contrapartida, para o parâmetro CP (Tabela 2), modelos mais complexos (como TabPFN) obtiveram melhores resultados, indicando uma relação não-linear predominante, com R^2 ajustado de 60%. Ademais, o parâmetro NDF (Tabela 3) apresentou resultados insatisfatórios, com valor de R^2 ajustado máximo de 24%. Mesmo com a seleção de variáveis e transformação com PCA, as bandas e variáveis climáticas podem ser insuficientes para explicar o parâmetro NDF, ou fatores externos podem ter introduzido ruído excessivo nos dados relacionados ao NDF. Por fim, GM e DM também obtiveram desempenhos insatisfatórios, com R^2 ajustado de 22% e 34%, respectivamente. Diferentemente dos parâmetros de qualidade de forragem, pouco se foi trabalhado nos parâmetros de quantidade de forragem, com as Tabelas 4 e 5 refletindo resultados iniciais do estudo.

As tabelas 6, 7, 8, 9 e 10 apresentam as variáveis finais selecionadas nos melhores modelos para cada variável-alvo, bem como os resultados obtidos com o *GridSearchCV*, utilizado para identificar as combinações ótimas de hiperparâmetros. Observa-se predominância de índices de vegetação e variáveis climáticas entre os modelos selecionados, evidenciando a relevância desses parâmetros para o desempenho dos modelos.

Tabela 6. Variáveis e PCA selecionados após etapa de seleção de variáveis e *GridSearchCV* para PMC

Seleção de Variáveis	PCA	Modelo
KBest + RFE: TEMP_MAX (ºC), RAD_SOL (MJ/m²), EVAPOT (mm), MS_EVI.	2	Linear Regression
Manual (Spearman): MS_EVI, S2_SR_NDWI, EVAPOT (mm), TEMP_MAX (ºC).	1	TabPFN Random Forest

Tabela 7. Variáveis e PCA selecionados após etapa de seleção de variáveis e *GridSearchCV* para CP

Seleção de Variáveis	PCA	Modelo
Manual (Spearman): MS_EVI, S2_VARI, S2_SR_NDWI, EVAPOT (mm), TEMP_MAX (ºC), RAD_SOL (MJ/m²), MS_NDVI.	4, 3	TabPFN Gradient Boosting
KBest + RFE: TEMP_MAX (ºC), RAD_SOL (MJ/m²), EVAPOT (mm), MS_NDVI, MS_EVI.	2	Linear Regression

Tabela 8. Variáveis e PCA selecionados após etapa de seleção de variáveis e *GridSearchCV* para NDF

Seleção de Variáveis	PCA	Modelo
KBest + RFE: MS_NDVI, S2_SR_VARI.	1	Linear Regression Gradient Boosting

Tabela 9. Variáveis e PCA selecionados após etapa de seleção de variáveis e *GridSearchCV* para GM

Seleção de Variáveis	PCA	Modelo
Manual: Bandas do Sentinel-2, S2_NDVI, S2_NDRE.	N/A	Multilayer Perceptron
KBest + RFE: L8_RAW_1_MNLI, L8_RAW_1_NLI.	2	XGBoost

Tabela 10. Variáveis e PCA selecionados após etapa de seleção de variáveis e *GridSearchCV* para DM

Seleção de Variáveis	Modelo
Manual: Bandas do Sentinel-2, S2_NDVI, S2_NDRE, TEMP_MAX (ºC), HUM_REL (%).	Multilayer Perceptron

4. Conclusão

Este trabalho apresentou a construção de modelos preditivos para os parâmetros teor de umidade (PMC), proteína bruta (CP), fibra em detergente neutro (NDF), matéria verde (GM) e matéria seca (DM). Esses parâmetros são fundamentais para a avaliação da qualidade nutricional e estrutural da forragem. Destaca-se o parâmetro PMC, que apesar da sua relevância na avaliação qualitativa das pastagens, poucos trabalhos são focados na estimativa automatizada dos valores de PMC. O presente trabalho adotou um procedimento rigoroso que consiste em técnicas de redução de variáveis, como *KBest* e *RFE*, redução de dimensões com *PCA* e busca de hiperparâmetros com *GridSearchCV*, para a geração de preditores acurados. Obteve-se resultados de R^2 ajustado de 63% para PMC, 60% para CP, 24% para NDF, 22% para GM, e 34% para DM.

Observa-se baixo desempenho dos modelos preditivos para NDF, DM e GM, associado a um número reduzido de variáveis fortemente correlacionadas com essas variáveis-alvo, o que indica uma limitação explicativa do conjuntos de dados. Ademais, as coletas de campo foram realizadas nos ciclos vegetativos de cada região, o que dificulta a generalização dos modelos para outras épocas do ano. Por fim, restrições e interrupções

no processo de coleta nas regiões QF e CUB devido ao COVID-19 podem ter introduzido ruído e lacuna nos dados, agravando a instabilidade e baixa robustez das previsões.

Um dos enfoques para novos trabalhos reside em projetar modelos baseados em redes neurais para melhorar o desempenho obtido. Vislumbra-se também a pesquisa de novas técnicas de *Data Augmentation* para gerar dados artificiais, como opção para adequar o uso do *dataset* a novas técnicas de predição.

Referências

Senar. *Bovinocultura: manejo e alimentação de bovinos de corte em confinamento*. Senar, 2018. Acessado em: 12/08/2025. Disponível em: https://www.cnabrasil.org.br/assets/arquivos/232-BOVINOCULTURA.pdf.

SERRANO, J. et al. Pasture quality assessment through ndvi obtained by remote sensing: A validation study in the mediterranean silvo-pastoral ecosystem. *Agriculture*, v. 14, n. 8, 2024. ISSN 2077-0472.

DEFALQUE, G. et al. Machine learning models for dry matter and biomass estimates on cattle grazing systems. *Computers and Electronics in Agriculture*, v. 216, p. 108520, 2024. ISSN 0168-1699.

BRETAS, I. L. et al. Prediction of aboveground biomass and dry-matter content in brachiaria pastures by combining meteorological data and satellite imagery. *Grass and Forage Science*, v. 76, n. 3, p. 340–352, 2021.

HORWITZ, W.; LATIMER, G. W. *Official Methods of Analysis of AOAC International*. 18th. ed. [S.l.]: AOAC International, 2005.

RODRIGUES, S. et al. Montadodb: A comprehensive dataset of pasture parameters in the southern region of portugal. *Data in Brief*, v. 62, p. 112029, 2025. ISSN 2352-3409.

GORELICK, N. et al. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, v. 202, p. 18–27, 2017. ISSN 0034-4257.

ZIPPENFENIG, P. *Open-Meteo.com Weather API*. 2023.

WANG, Z. et al. A comprehensive survey on data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, p. 1–20, 2025.

DRUCKER, H. et al. Support vector regression machines. In: *Proceedings of the 10th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1996. (NIPS'96), p. 155–161.

BREIMAN, L. Random forests. Springer, v. 45, n. 1, p. 5–32, 2001.

CHEN, T. et al. *xgboost: Extreme Gradient Boosting*. [S.l.], 2025. R package version 3.2.0.0. Disponível em: https://github.com/dmlc/xgboost.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

HOLLMANN, N. et al. *TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second*. 2023. Disponível em: https://arxiv.org/abs/2207.01848.