

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DANIELLI BATISTELLA

**ESTIMATIVA DE PRODUTIVIDADE DE SOJA POR MEIO DE IMAGENS
ORBITAIS E *MACHINE LEARNING***

TESE DE DOUTORADO

PATO BRANCO

2023

DANIELLI BATISTELLA

**ESTIMATIVA DE PRODUTIVIDADE DE SOJA POR MEIO DE IMAGENS
ORBITAIS E *MACHINE LEARNING***

Soybean yield estimation using orbital images and *machine learning*

Tese de doutorado apresentada como requisito parcial à obtenção do título de Doutora em Agronomia - Área de Concentração: Solos e Sistemas Integrados de Produção Agropecuária da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Dr. José Ricardo da Rocha Campos.

Coorientador: Dr. Alcir José Modolo.

Dr. Vanderlei Aparecido de Lima.

PATO BRANCO

2023



[4.0 Internacional](https://creativecommons.org/licenses/by-nc/4.0/)

Esta licença permite remixe, adaptação e criação a partir do trabalho, para fins não comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.



**Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Pato Branco**



DANIELLI BATISTELLA

ESTIMATIVA DE PRODUTIVIDADE DE SOJA POR MEIO DE IMAGENS ORBITAIS E MACHINE LEARNING

Trabalho de pesquisa de doutorado apresentado como requisito para obtenção do título de Doutora Em Agronomia da Universidade Tecnológica Federal do Paraná (UTFPR).
Área de concentração: Solos E Sistemas Integrados De Produção Agropecuária.

Data de aprovação: 06 de Julho de 2023

Dr. Jose Ricardo Da Rocha Campos, Doutorado - Universidade Tecnológica Federal do Paraná
Dr. Gilvan Moises Bertollo, Doutorado - Universidade Federal de Santa Maria (Ufsm)
Dr. Giovani Benin, Doutorado - Universidade Tecnológica Federal do Paraná
Dr. Mario Sergio Muniz Tagliari, Doutorado - Faculdade Municipal de Meio Ambiente
Dr. Moeses Andriago Danner, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 06/07/2023.

Ao Dênio e a Dhália, com carinho.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por tudo!

Agradeço imensamente os orientadores Alcir José Modolo e José Ricardo da Rocha Campos, bem como o coorientador Vanderlei Aparecido de Lima, pela oportunidade, confiança e aprendizado.

À professora Ana Claudia dos Santos Luciano, pela contribuição dada a essa pesquisa.

Aos servidores da UTFPR, na condição de DIRPPG, Sandro César Bortoluzzi e NPPD, Marta Helena Dias da Silveira (*in memoriam*), pelo apoio à qualificação docente.

Aos produtores rurais dos municípios de Bom Sucesso do Sul, Clevelândia, Mangueirinha, Palmas e Pato Branco, pelo fornecimento dos dados de produção agrícola, e em especial, ao produtor Vitor Bandeira pela cordialidade.

À concessionária Verdesul da John Deere, ao Adoilson dos Passos e a Chaiane Borsatti, pela mediação no processo de obtenção dos dados de campo.

À Secretaria da Agricultura e do Abastecimento, regional de Pato Branco, pela disponibilização dos dados municipais, em especial, ao servidor Ivano Carniel, pelos diversos esclarecimentos acerca do contexto regional da agricultura.

Ao SIMEPAR, pela disponibilização dos dados meteorológicos.

Aos meus pais Vicente e Ivete, pela vida e pelos diversos ensinamentos.

Ao meu esposo Juliano, fiel companheiro e pai dedicado que Deus me presenteou, pelo grande apoio aos meus projetos de vida.

E ao meu filho Luiz Vicente, por ter trazido imensa alegria à minha vida.

Renova-se a esperança
Nova aurora a cada dia
E há que se cuidar do broto
Pra que a vida nos dê flor e fruto

(Milton Nascimento e Wagner Tiso)

RESUMO

Para uma região produtora de soja, estimar, de forma precisa e segura, a produtividade futura de suas lavouras, pode ser muito estratégico sob o ponto de vista comercial, porque gera segurança às cooperativas na comercialização da produção. Nesse sentido, o Sensoriamento Remoto (SR) orbital permite operacionalizar os programas de previsão e monitoramento de safras, além de caracterizar, em tempo hábil, a fenologia de culturas de interesse agrônômico. Para tal propósito, tem sido utilizado os métodos de Aprendizado de Máquina (AM) para uma melhor estimativa do estado da cultura e do ambiente para tomada de decisões. Dessa forma, o objetivo geral desse estudo foi aprimorar os métodos de estimativa da produção agrícola da cultura da soja já existentes por meio de imagens orbitais de maior resolução espacial, com dados de referência obtidos por monitores de colheita e processados por meio de algoritmos de AM, de forma a construir modelos de estimativa de safras com maior precisão. O estudo foi realizado no Núcleo Regional de Pato Branco/PR, composto por 15 municípios, localizados no Sudoeste do estado do Paraná. Para determinação das áreas de cultivo da soja foram criados perfis espectro-temporais RGB a partir da geração de imagens de máximo e mínimo NDVI (*Normalized Difference Vegetation Index*) e EVI (*Enhanced Vegetation Index*) obtidas do sensor MSI do satélite SENTINEL-2 e classificadas por método supervisionado do algoritmo *Random Forest* (RF) no aplicativo R Studio. Para gerar os modelos de estimativa de produtividade de soja, foram testados três algoritmos supervisionados de regressão linear por AM, denominados *Random Forest* (RF), *Support Vectors Machine* (SVM) e *Artificial Neural Networks* (ANN) no aplicativo R-Studio. Os dados de entrada para geração dos modelos preditivos foram: os índices de vegetação NDVI, EVI, NDMI (*Normalized Difference Moisture Index*), tanto do sensor MSI quanto do MODIS (AQUA e TERRA) das áreas de produção agrícola do Núcleo Regional Pato Branco-PR, dados de precipitação, dados de terreno referentes à altitude, tipo de solos, geologia e declividade e, dados reais de produção dos talhões, correspondentes às safras 2019/2020, 2020/2021 e 2021/2022. Os modelos de regressão preditivos foram comparados com a produtividade real dos próprios talhões utilizados no estudo. A estimativa de produção gerada para área de estudo, foi comparada com os dados oficiais a nível municipal e regional. O índice NDVI proporciona maior acurácia na determinação das áreas de cultivo da soja quando comparado ao EVI. O algoritmo de AM *Random Forest* gera um modelo de estimativa de produção com maior precisão quando comparado com os de *Support Vectors Machine* e *Artificial Neural Networks*. Os índices de vegetação NDMI, EVI e NDVI se configuraram como importantes variáveis preditoras na construção dos modelos. A utilização de índices de vegetação obtidos por imagens de alta resolução espacial, como o sensor MSI do SENTINEL-2, permite estimar a produção da soja de modo igual e superior as do sensor MODIS do TERRA/AQUA que possuem alta temporalidade quando associadas com algoritmos de Aprendizado Máquina e quando consideram a variabilidade espacial do talhão na variável resposta dos modelos preditivos de regressão linear.

Palavras-chave: Agricultura de Precisão. Sensoriamento Remoto. Inteligência Artificial.

ABSTRACT

For a soy-producing region, accurately and safely estimating the future productivity of its crops can be very strategic from a commercial point of view, because it creates security for cooperatives in producing production. In this sense, orbital Remote Sensing (RS) makes it possible to operationalize crop forecasting and monitoring programs, in addition to characterizing, in a timely manner, the phenology of crops of agronomic interest. For this purpose, Machine Learning (ML) methods have been used to better estimate the state of the crop and the environment for decision making. In this way, the general objective of this study was to improve the existing methods of belief of the agricultural production of the soybean crop through orbital images of higher spatial resolution, with reference data obtained by harvest monitors and processed through ML algorithms, in order to build more accurate yield estimation models. The study was carried out in the Regional Nucleus of Pato Branco/PR, composed of 15 municipalities. To cultivate the soybean cultivation areas, RGB spectro-temporal profiles were created from the generation of maximum and minimum NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index) images transmitted from the MSI sensor of the SENTINEL-2 satellite and classified by supervised method of the Random Forest (RF) algorithm in the R Studio application. To generate soybean yield estimation models, three supervised AM linear regression algorithms called RF; Support Vectors Machine (SVM) and Artificial Neural Networks (ANN) in the R-Studio app. The input data were used to generate the predictive models: the NDVI, EVI, NDMI (Normalized Difference Moisture Index) vegetation indices, both from the MSI and MODIS sensors (AQUA and TERRA) of the agricultural production areas in the region of Pato Branco -PR, precipitation data, terrain data referring to altitude, soil type, geology and slope, and actual plot production data, corresponding to the 2019/2020, 2020/2021 and 2021/2022 harvests. The predictive regression models were compared with the actual productivity of the plots used in the study. The production estimate generated for the study area was detected with official data at municipal and regional level. The NDVI index provides greater accuracy in the experience of soybean growing areas when compared to the EVI. The ML Random Forest algorithm generates a more accurate production estimation model when compared to Artificial Neural Networks and Support Vector Machines. The NDMI, EVI and NDVI vegetation indices are important predictors of variables in the construction of the models. The use of vegetation indices obtained from high spatial resolution images, such as the MSI of SENTINEL-2, allows estimating soybean production in an equal and superior way to that of the MODIS sensor of TERRA/AQUA, which have high temporality when associated with algorithms for Machine Learning and when to consider the spatial variability of the field in the response variable of predictive linear regression models.

Keywords: Precision Agriculture. Remote Sensing. Artificial Intelligence.

LISTA DE ILUSTRAÇÕES

Figura 1 - Histórico de produção e área colhida de soja no estado do Paraná e no Núcleo Regional Pato Branco.....	17
Figura 2 - Curva média de reflectância da vegetação fotossinteticamente ativa.....	22
Figura 3 - Hierarquia de aprendizado.....	26
Figura 4 – Esquema representativo do algoritmo da Regressão por Random Forest.....	28
Figura 5 - Exemplo de rede neural com uma camada oculta.....	29
Figura 6 - Localização do Núcleo Regional Pato Branco - PR.....	32
Figura 7 - Composição RGB espectro-temporal para a cultura da soja....	35
Figura 8 - Classificação por Random Forest.....	36
Figura 9 - Fluxograma da metodologia do mapeamento das áreas com a cultura da soja.....	38
Figura 10 - Distribuição das chuvas nas estações meteorológicas inseridas próximas ao Núcleo Regional Pato Branco.....	40
Figura 11 - Localização dos talhões utilizados para obtenção dos dados reais de produtividade de soja.....	42
Figura 12 - Esquema geral de obtenção dos dados de colheita da soja nos talhões.....	43
Figura 13 - Fluxograma da metodologia da estimativa de produtividade de soja para o Núcleo Regional Pato Branco.....	46
Figura 14 - Distribuição espacial da cultura da soja no Núcleo Pato Branco. UTFPR, Pato Branco - PR, 2023.....	49
Figura 15 - Dispersão da produtividade real e predita de soja por Random Forest. UTFPR, Pato Branco - PR, 2023.....	53
Figura 16 – Ranqueamento das variáveis preditoras para cada modelo gerado pelos sensores MODIS e MSI, para predição da estimativa de produtividade da cultura da soja por meio do algoritmo RF. UTFPR, Pato Branco - PR, 2023.....	55
Figura 17 - Mapa da distribuição da estimativa de produtividade da soja (t ha ⁻¹) no Núcleo Regional de Pato Branco – Safra 2019/2020. UTFPR, Pato Branco - PR, 2023.....	58
Figura 18 - Mapa da distribuição da estimativa de produtividade da soja (t ha ⁻¹) no Núcleo Regional de Pato Branco – Safra 2020/2021. UTFPR, Pato Branco - PR, 2023.....	59
Figura 19 - Mapa da distribuição da estimativa de produtividade da soja (t ha ⁻¹) no Núcleo Regional de Pato Branco – Safra 2021/2022. UTFPR, Pato Branco - PR, 2023.....	60

LISTA DE TABELAS

Tabela 1 - Produtividade da soja (Kg ha^{-1}) para o estado do Paraná e Núcleo Regional de Pato Branco	18
Tabela 2 - Bandas espectrais do visível e infravermelho do sensor MSI do SENTINEL-2	20
Tabela 3 - Descrição dos talhões utilizados no estudo.....	41
Tabela 4 - Matriz de confusão e porcentagem de erros para conjunto treinamento e teste das máscaras de NDVI e EVI da cultura da soja, Núcleo Pato Branco - Safra 2019/2020. UTFPR, Pato Branco - PR, 2023	47
Tabela 5 - Matriz de confusão e porcentagem de erros para conjunto treinamento e teste das máscaras de NDVI e EVI da cultura da soja, Núcleo Pato Branco - Safra 2020/2021. UTFPR, Pato Branco - PR, 2023	48
Tabela 6 - Matriz de confusão e porcentagem de erros para conjunto treinamento e teste das máscaras de NDVI e EVI da cultura da soja, Núcleo Pato Branco - Safra 2020/2021. UTFPR, Pato Branco - PR, 2023	48
Tabela 7 - Informações de área oficial (DERAL/SEAB) e obtidas pela classificação para o Núcleo Pato Branco - PR. UTFPR, Pato Branco - PR, 2023	50
Tabela 8 - Regressão com os algoritmos de Aprendizado de Máquina. UTFPR, Pato Branco - PR, 2023	51
Tabela 9 - Estatísticas descritivas da produtividade de soja (t ha^{-1}) real dos talhões e da estimativa do Núcleo Regional Pato Branco, por Aprendizado de Máquina com dados do sensor MSI. UTFPR, Pato Branco - PR, 2023 .	56
Tabela 10 - Estatísticas descritivas da produtividade de soja (t ha^{-1}) real dos talhões e da estimativa do Núcleo Regional Pato Branco, por Aprendizado de Máquina com dados do sensor MODIS. UTFPR, Pato Branco - PR, 2023.	57
Tabela 11 - Comparação entre produção oficial e estimada de soja (t) por município do Núcleo Regional de Pato Branco. UTFPR, Pato Branco - PR, 2023	61

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

AM	Aprendizado Máquina
ANN	Artificial Neural Networks
CONAB	Companhia Nacional de Abastecimento
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
EOS	Earth Observing System
ESA	European Space Agency
EVI	Enhanced Vegetation Index
IAF	Índice de Área Foliar
IDW	Inverso da Distância ao Quadrado
INMET	Instituto Nacional de Meteorologia
INPE	Instituto Nacional de Pesquisas Espaciais
MODIS	Moderate-Resolution Imaging Spectroradiometer
MSI	Multi Spectral Instrument
NDMI	Normalized Difference Moisture Index
NDVI	Normalized Difference Vegetation Index
OOB	Out-of-bag
PR	Paraná
RF	Random Forest
RMSE	Root Mean Squared Error
SC	Santa Catarina
SEAB	Secretaria de Estado da Agricultura e do Abastecimento
SIMEPAR	Sistema de Tecnologia e Monitoramento Ambiental do Paraná
SR	Sensoriamento Remoto
SVM	Support Vectors Machine
VANT's	Veículos Aéreos Não-Tripulados
VIF	Variance Inflation Fator

SUMÁRIO

1. INTRODUÇÃO	13
2. REVISÃO DE LITERATURA	16
2.1 Aspectos gerais e importância da cultura da soja	16
2.2 Sensores orbitais e uso na agricultura	19
2.3 Comportamento espectral dos alvos agrícolas	21
2.4 Índices de vegetação	24
2.5 Aprendizado de Máquina (Machine Learning) e suas aplicações na agricultura	26
2.5.1 Árvores de decisão	27
2.5.2 Redes Neurais Artificiais	28
2.5.3 Máquina de Vetores de Suporte	30
3. MATERIAL E MÉTODOS	32
3.1 Caracterização da área de estudo	32
3.2 Mapeamento das áreas com a cultura da soja	32
3.3 Modelo de classificação Random Forest	36
3.4 Análise de comparação de áreas	37
3.5 Modelo de estimativa de produtividade	38
3.6 Mineração de dados por Aprendizado Máquina	43
3.7 Regressão por Aprendizado de Máquina	44
3.8 Aplicação do modelo para estimativa da produção no Núcleo Regional Pato Branco	45
4. RESULTADOS E DISCUSSÃO	47
4.1 Mapeamento das áreas	47
4.2 Modelos de estimativa de produtividade de soja	51
4.3 Estimativa da produtividade no Núcleo Regional Pato Branco	57
5. CONCLUSÕES	63
6. CONSIDERAÇÕES FINAIS	64
REFERENCIAS BIBLIOGRÁFICAS	65
APÊNDICE A – Conjuntos de dados SEM aplicação de filtros tais como outliers, extremos valores e VIFs do sensor MSI	73
APÊNDICE B – Conjunto de dados SEM aplicação de filtros tais como outliers, extremos valores e VIFs do sensor MODIS	74
APÊNDICE C – Variáveis utilizadas COM aplicação de filtros tais como outliers, extremos valores e VIFs para os conjuntos do sensor MSI	76
APÊNDICE D – Variáveis utilizadas COM aplicação de filtros tais como outliers, extremos valores e VIFs para os conjuntos do sensor MODIS ...	77

1. INTRODUÇÃO

A produção de soja tem se tornado uma das atividades econômicas com maior crescimento nas últimas décadas, pois, além do comércio de seus subprodutos, houve a consolidação no mercado internacional como importante fonte de proteína vegetal que atende principalmente à produção animal. Essa valorização do cultivo viabilizou a expansão das áreas cultivadas em diversos lugares no mundo, inclusive no Brasil, que possui significativa participação na oferta e demanda de produtos agrícolas (BALBINOT JÚNIOR et al., 2022).

A safra brasileira de soja 2021/2022 foi de aproximadamente 125 milhões de toneladas (CONAB, 2022), sendo que o estado do Paraná produziu o equivalente a 12,4 milhões, com produtividade de 2.182 kg ha⁻¹. O núcleo Pato Branco obteve sexta posição no ranking paranaense da produção de soja (DERAL/SEAB, 2022).

O monitoramento das safras está diretamente relacionado ao planejamento e a estimativa precisa e, a nível local e regional, é importante para melhor utilizar os recursos da colheita dentro da micro e macroeconomia (JOHNSON, 2014). Nesse sentido, no que diz respeito às estatísticas agrícolas oficiais, estas estimativas são geradas a partir de questionários e entrevistas com técnicos, entidades do setor agrícola e produtores rurais (DERAL/SEAB, 2019), mas esse método não quantifica os erros envolvidos.

O Sensoriamento Remoto (SR) orbital permite operacionalizar os programas de previsão e monitoramento de safras (GALLEGO et al., 2010), além de adquirir informações sobre grandes extensões geográficas em tempo hábil e caracterizar a fenologia de culturas ou da vegetação, de forma geral. Entretanto, para quantificação de produção agrícola das principais *commodities* brasileiras, tem-se utilizado índices de vegetação das séries temporais com imagens de baixa resolução espacial, como MODIS para a soja (RICHETTI et al., 2018) e LANDSAT para a cultura da cana-de-açúcar (LUCIANO et al., 2021). Além disso, os trabalhos dos autores supracitados não consideram a variabilidade espacial dentro dos talhões usados como referências para a criação das estimativas, sendo apenas utilizados a média de produtividade de cada talhão.

A estimativa de produção é essencial na agricultura e o SR tem sido amplamente utilizado na construção de ferramentas de apoio para tomada de

decisões e no gerenciamento de sistemas agrícolas. Entretanto, isso requer o processamento de enorme quantidade de dados de diferentes plataformas, tanto orbitais quanto suborbitais. Para tais propósitos, métodos de Aprendizado de Máquina (AM) têm sido amplamente utilizados para geração de modelos preditivos, em função de sua capacidade de processar um grande número de entradas de dados. Nos últimos anos, os rápidos avanços em tecnologias de detecção de alvos por meio de sensores remotos e algoritmos de AM, proporcionaram maior custo-benefício e soluções para uma melhor estimativa do estado das culturas agrícolas. Essas novas abordagens farão parte da prática da agricultura de precisão num futuro próximo (CHLINGARYAN et al., 2018).

Exemplos de estudos envolvendo estimativas de produtividade com AM são encontrados nos trabalhos de Johnson (2014), que apresentou modelos cujos coeficientes de determinação (R^2) foram iguais a 0,71 para a cultura da soja e 0,77 para o milho; Liu et al. (2020), aplicaram o NDVI do MODIS para estimar rendimentos de cevada, canola e trigo, com valores de coeficientes de determinação entre 0,54 a 0,81 em regiões úmidas; Khanal et al. (2018), utilizaram imagem multiespectral de alta resolução referente de terra nua e dados do terreno, para prever propriedades do solo e rendimento do milho com coeficiente de determinação (R^2) de 0,53 com o algoritmo *Random Forest*; Pantazi et al. (2016) estimaram a produção de trigo com redes neurais artificiais de contra-propagação com precisão geral média de 78,3%; Gandhi et al. (2016), obtiveram R^2 0,69, ao testarem o algoritmo Máquina de Vetores de Suporte com parâmetros climáticos para estimar a produção de arroz em Maharashtra na Índia; Li et al. (2022), usaram os índices de vegetação de imagens pelo MODIS para estimar a produtividade do trigo no Noroeste da China, com R^2 0,74 e RMSE de 0,76.

Nesse contexto, a principal hipótese desse trabalho é que a utilização de índices de vegetação gerados a partir de imagens de maior resolução espacial, como as de 10 m do sensor MSI do SENTINEL-2, juntamente com dados do terreno e de precipitação, associados aos dados reais de produção que consideram a variabilidade espacial existente no talhão, possam gerar modelos mais precisos para a estimativa de produtividade que os trabalhos atualmente encontrados na literatura, como por exemplo, das pesquisas com o sensor MODIS, que possui alta resolução temporal e é amplamente utilizado no

monitoramento das culturas agrícolas.

E diante de tais explicações, o objetivo geral desse estudo foi aprimorar os métodos de estimativa da produção agrícola da cultura da soja já existentes por meio de modelos preditivos utilizando-se imagens orbitais de maior resolução espacial, dados de precipitação, geologia, solos e do terreno, associados com dados de referência obtidos por monitores de colheita e processados por algoritmos de AM.

Objetivos específicos:

- Mapear as áreas de produção da cultura da soja a partir da criação de um perfil-espectro temporal de índices de vegetação, utilizando-se imagens de maior resolução espacial, do sensor MSI do SENTINEL-2.
- Comparar três algoritmos de regressão por AM amplamente utilizados na literatura, como Floresta Aleatória, Redes Neurais Artificiais e Máquina de Vetores de Suporte, para realizar a estimativa de produção de soja a partir de dados reais dos talhões;
- Avaliar o desempenho dos índices de vegetação obtidos dos sensores MSI e MODIS, dados de terreno e precipitação na estimativa da produção agrícola da soja.
- Comparar o desempenho dos sensores MSI e MODIS nos modelos de estimativa de produtividade.
- Estimar regionalmente a produção agrícola mediante utilização de modelo de regressão por AM com melhor desempenho.

2. REVISÃO DE LITERATURA

2.1 Aspectos gerais e importância da cultura da soja

A soja passou a ser um produto, economicamente, importante para o Brasil no final da década de 60. Atualmente o país se tornou um grande produtor do grão no cenário internacional. Em 1966, a produção comercial de soja já era uma necessidade estratégica, sendo produzidas cerca de 500 mil toneladas no País (EMBRAPA, 2023). A maior parte da soja cultivada comercialmente tem sido destinada para a alimentação animal e humana, pois é um alimento rico em proteínas, o que permite além da fabricação de ração, também a produção de óleo para alimentação e uso industrial (FAO, 2017).

O preço da soja no mercado mundial elevou-se em meados da década de 1970, despertando interesse dos agricultores e do próprio governo brasileiro. Nesse sentido, o país passou a investir em pesquisa e tecnologia para adaptação da cultura às condições brasileiras, através da Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA, o que permitiu que a soja fosse adaptada ao clima tropical brasileiro. Os líderes mundiais na produção de soja são os Estados Unidos, Brasil, Argentina, China, Índia e Paraguai (EMBRAPA, 2023).

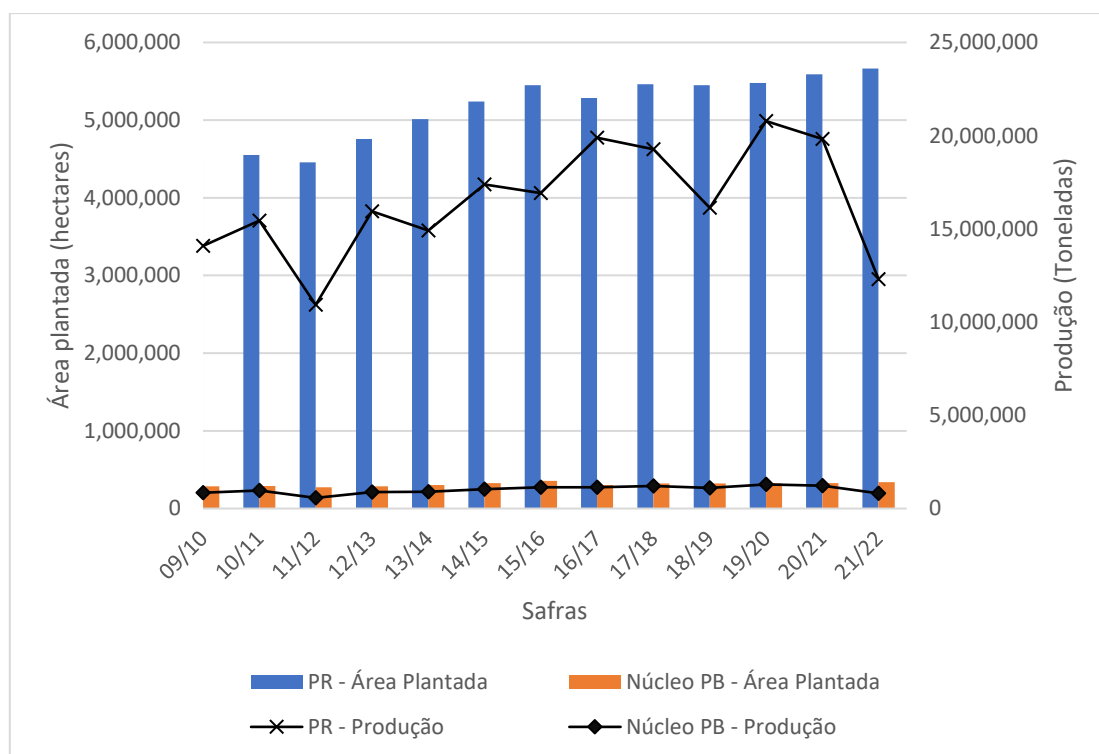
A produção brasileira de soja na safra 2019/2020 foi de aproximadamente 125 milhões de toneladas, representando aumento de 4,3% em relação ao período anterior e produtividade de 3.379 kg ha⁻¹. Já, na safra 2020/2021, ocorreu aumento da área plantada de soja de 4,3% em comparação à safra anterior, alcançando 38,5 milhões de hectares e uma produção recorde de 136 milhões de toneladas e produtividade de 3.527 kg ha⁻¹ (CONAB, 2020; CONAB, 2021).

Entretanto, a safra 2021/2022 sofreu influência do fenômeno La Niña na Região Sul e no Mato Grosso do Sul, o que reduziu drasticamente o volume de chuvas nessas regiões e conseqüentemente afetou a produção total de soja no país. Apesar do aumento da área plantada ser 4,9% superior ao semeado na safra 2020/2021, a produção obtida foi de apenas 125 milhões de toneladas e a produtividade de 3.029 kg ha⁻¹ (CONAB, 2022).

O estado do Paraná obteve na safra 2019/2020 a segunda posição nacional no volume produzido de soja, representando aproximadamente 17% da produção brasileira, com cerca de 21 milhões de toneladas (DERAL/SEAB,

2020). Enquanto isso, no núcleo Pato Branco, a área plantada para a cultura da soja foi de 324.950 hectares e a produção de 1.285.767 toneladas (Figura 1). Este núcleo, representou na safra citada, a quinta posição no ranking de produção do estado.

Figura 1 - Histórico de produção e área colhida de soja no estado do Paraná e no Núcleo Regional Pato Branco



Fonte: SEAB/DERAL (2023)

Para a safra 2020/2021, a produção estadual de soja foi de 19,8 milhões de toneladas e a do Núcleo Pato Branco de 1.223.946 toneladas. Já, para safra 2021/2022, apesar do aumento da área plantada ter sido de aproximadamente 4%, a produção estadual foi abaixo do esperado, com apenas 12,3 milhões de toneladas e a do Núcleo de 820.882 toneladas. Esse decréscimo da produção ocasionado pela falta de chuva foi de 38,6% no estado e 32,9% na região de estudo. Nesse contexto, a variação da produtividade ao longo das safras pode ser visualizada na Tabela 1.

Tabela 1 - Produtividade da soja (Kg ha⁻¹) para o estado do Paraná e Núcleo Regional de Pato Branco

Região	Safras		
	2019/2020	2020/2021	2021/2022
Paraná	3.794	3.548	2.182
N. R. Pato Branco	3.957	3.731	2.411

Fonte: DERAL/SEAB (2023)

Apesar da soja ser relativamente tolerante ao déficit hídrico na sua fase vegetativa, possui grande sensibilidade ao estresse durante a formação dos componentes do rendimento, que correspondem ao florescimento e enchimento de grãos (SENTELHAS et al., 2015). Além do estresse hídrico, existem vários fatores que impedem que a soja alcance todo o seu potencial produtivo. Seixas et al. (2020), destacam que as doenças causadas por fungos, bactérias, vírus e nematoides, podem afetar a cultura desde a germinação até o final do enchimento de grãos/sementes e, que a escolha de um cultivar resistente ajuda a minimizar o manejo de pragas. Oliveira Junior et al. (2020), enfatizam a necessidade de conhecimento da fertilidade do solo integrado com o manejo adequado da adubação e o uso racional de fertilizantes para produção de grãos.

A cultura da soja é descrita por estádios fenológicos que relacionam os períodos em que a lavoura atravessa durante o seu o ciclo de desenvolvimento (NEUMAIER et al., 2020). Fehr e Cavines (1977), propuseram uma divisão do ciclo da soja denominado estádios vegetativos (V) que compreende ao período de emergência da plântula até o início da floração e estádios reprodutivos (R), que corresponde a época de florescimento até sua a maturação.

As subdivisões da fase vegetativa são designadas numericamente como V1, V2, V3, até Vn, com exceção dos dois estádios iniciais, que são identificados como VE (emergência) e VC (estádio de cotilédone). O último estágio vegetativo é denominado de Vn, onde “n” consiste no número do último nó vegetativo, que varia conforme a variedade da soja e das condições ambientais. A fase reprodutiva contém oito estádios que compreendem a fase de florescimento (R1 e R2), desenvolvimento da vagem (R3 e R4), desenvolvimento da semente (R5 e R6) e maturação da planta (R7 e R8) (FEHR; CAVINESS, 1977).

2.2 Sensores orbitais e uso na agricultura

O Sensoriamento Remoto (SR) por meio de imagens obtidas por sensores embarcados em satélites é capaz de obter informações sobre a cultura agrícola, o solo e o relevo de extensas áreas, uma vez que apresentam grande potencial de identificação da variabilidade espacial do desenvolvimento das plantas. Por meio dessa informação, é possível investigar e analisar os causadores dessa variabilidade, assim como, propor intervenções quando a variabilidade for atribuída a certo fator, como ataque de pragas ou problemas de fertilidade (MOLIN et al., 2015). Além de monitorar o desenvolvimento vegetal espacial e temporal, contribui para a identificação de novas variedades que melhor se adaptem às diferentes regiões do país. Ainda, o SR pode ser utilizado para o monitoramento do uso de terras agrícolas e para a previsão da produção agrícola (WEISS et al., 2020).

Dados de satélite disponíveis gratuitamente, mesmo fornecendo menor resolução espacial e temporal, têm sido utilizados para monitorar a vegetação e estimar os rendimentos de diversas culturas agrícolas. Apesar de alguns sensores orbitais terem limitações em sua resolução e apresentarem nuvens no momento do imageamento, estes estão bem posicionados em plataformas estáveis, o que gera automaticamente imagens com menores distorções (SINGH et al., 2020).

Para utilização de imagens orbitais, é necessário identificar se a qualidade desses produtos satisfaz as demandas requeridas para determinado mapeamento. Diversos produtos gratuitos disponíveis com diferentes resoluções espaciais e temporais, podem ser aplicados na agricultura, dentre eles, o sensor MODIS, com resolução temporal de dois dias e, o sensor MSI do satélite SENTINEL-2, com resolução espacial de 10 metros e temporal de 5 dias (NASA, 2021; ESA, 2020).

O sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*) está a bordo dos dois satélites americanos TERRA e AQUA do Programa EOS (*Earth Observing System*) da Agência Espacial Americana, que realiza o mapeamento e o monitoramento da cobertura vegetal e uso da terra. Este possui alta sensibilidade radiométrica (12 bits) em 36 bandas espectrais que variam em comprimento de onda de 0,4 a 14,4 μm e resoluções espaciais de 250 a 1000

metros (NASA, 2021). A partir das bandas do sensor MODIS são desenvolvidos produtos como os índices de vegetação denominados MOD13 e MYD13, os quais têm sido amplamente utilizados para análise de séries temporais da dinâmica da vegetação (BECK et al., 2006).

O sensor MSI (*Multispectral Instrument*) está acoplado ao satélite SENTINEL-2 e se caracteriza como uma missão de imageamento multiespectral, de alta resolução, conjuntamente administrada pela Comunidade Europeia e a ESA (*European Space Agency*), com a finalidade de monitorar a variabilidade nas condições da superfície da terra, realizando coleta de dados sobre a vegetação, solos e umidade, rios e áreas costeiras, e dados para correção atmosférica (absorção e distorção). Esta missão garante a continuidade dos dados fornecidos pelo SPOT 5 e Landsat 7 (ESA, 2020). Algumas características das imagens obtidas pelo sensor MSI são apresentadas na Tabela 2.

Tabela 2 - Bandas espectrais do visível e infravermelho do sensor MSI do SENTINEL-2

Resolução	Nº da Banda	Descrição	Comprimento de Onda Central (nm)	
			S2A	S2B
10 m	B02	Blue (Azul)	492,7	492,3
	B03	Green (Verde)	559,8	558,9
	B04	Red (Vermelho)	664,6	664,9
	B08	NIR (Infravermelho Próximo)	832,8	832,9
20 m	B05	Red Edge 1	704,1	703,8
	B06	Red Edge 2	740,5	739,1
	B07	Red Edge 3	782,8	779,7
	B08a	Red Edge 4	864,7	864,0
	B11	SWIR 1 (Infravermelho Médio)	1.613,7	1.610,4
	B12	SWIR 2 (Infravermelho Médio)	2.202,4	2.185,7
60 m	B01	Aerossol	442,7	442,3
	B09	Water vapor	945,1	943,2
	B10	Cirrus	1.373,5	1.376,9

Fonte: ESA (2020)

A largura de faixa imageada é de 290 km e o tempo de revisita é de 10 dias no equador com um satélite e 5 dias com 2 satélites sob condições sem nuvens, que resulta em 2-3 dias em latitudes médias. Os limites de cobertura

são entre as latitudes 56° sul e 84° norte. Este satélite transporta uma carga útil de instrumento óptico que mostra 13 bandas espectrais: quatro bandas a 10 m, seis bandas a 20 m e três bandas de 60 m de resolução espacial. A resolução radiométrica das bandas é de 12 bits (ESA, 2020). A grande quantidade de dados do Sentinel-2, cuja continuidade é garantida por décadas, oferece uma oportunidade única para monitorar colheitas em quase tempo real (VELOSO et al., 2017).

O sensoriamento remoto tem sido utilizado com grande êxito na área agrônômica. A premissa do uso de dados digitais no monitoramento de áreas agrícolas é baseada na interação única da biomassa vegetal com a radiação eletromagnética solar. A disponibilidade de pelo menos quatro décadas de dados digitais com várias bandas de ondas do espectro (visível, infravermelho próximo e bandas térmicas) e sua grande cobertura terrestre, torna o sensoriamento remoto superior aos estudos baseados em campo (HEREHER, 2013).

Nesse contexto, várias pesquisas têm sido desenvolvidas com aplicação de dados dos sensores orbitais na agricultura. Como exemplos tem-se as pesquisas de Stepanov et al. (2020), que utilizaram dados do MODIS para monitorar a produtividade da cultura da soja em municípios do território Khabarovsk, localizado no extremo oriente da Rússia, usando campos experimentais como referência, onde obtiveram diferença de 6,3% entre os valores reais e preditos de produtividade para a safra 2019.

Habibi et al. (2021) testaram modelos de previsão da variabilidade espacial da densidade de plantas de soja com imagens do sensor comercial PlanetScope de 3 metros de resolução espacial, a partir de modelos de regressão por Floresta Aleatória, obtendo o valor R^2 0,63. Xin et al. (2013), criaram um modelo para estimar a eficiência de produção de milho e soja com dados do sensor MODIS, onde obtiveram os valores de R^2 0,77 e R^2 0,66, respectivamente.

2.3 Comportamento espectral dos alvos agrícolas

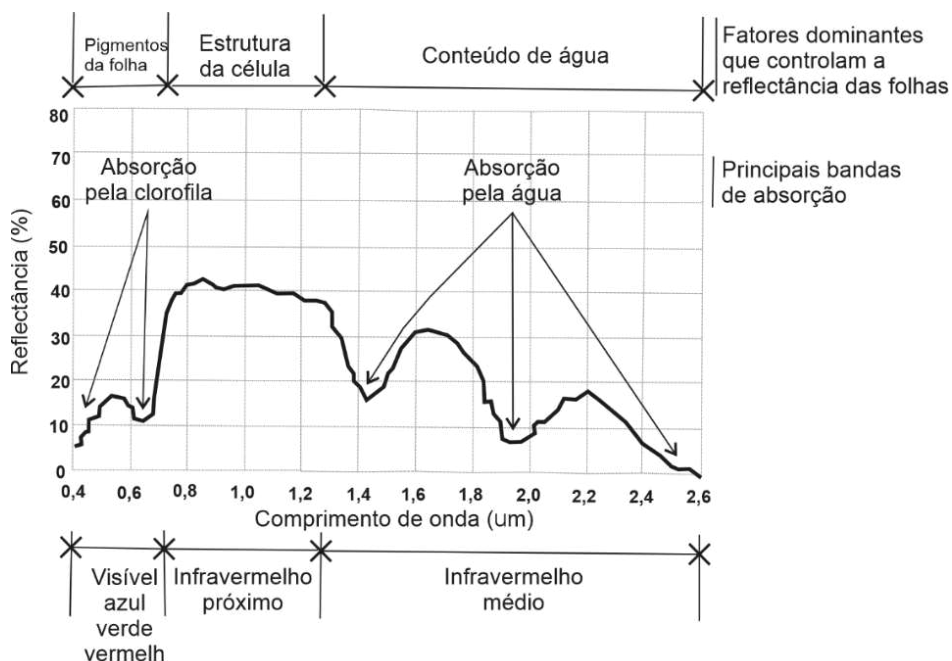
A aplicação das técnicas de Sensoriamento Remoto no estudo da vegetação inclui a necessidade de compreender o processo de interação entre a radiação eletromagnética e os diversos tipos fisionômicos de dosséis (PONZONI et al., 2012), especialmente, o das culturas agrícolas.

A resposta espectral das culturas agrícolas depende de uma série de fatores bioquímicos do vegetal utilizado como alvo, além das características físicas do dossel. Esses fatores estão condicionados a arquitetura do dossel, estágios de desenvolvimento das plantas, parâmetros agrônômicos e condições atmosféricas (MARTINS; GALLO, 2015).

Os mecanismos que influenciam na quantidade de energia eletromagnética refletida pelas folhas são: pigmentos, espaços ocupados pela água e pelo ar e suas estruturas celulares. Outros fatores que afetam a energia refletida pelas folhas estão relacionados com o conteúdo de água, maturação ou idade da folha, posição nodal, condição de iluminação, pubescência e senescência (GAUSMAN, 1985).

A reflectância da vegetação fotossinteticamente ativa apresenta sua curva característica (Figura 2). E a radiação solar contida na faixa de comprimento entre 400 a 3.000 nm, é denominada de espectro refletivo. Este intervalo de comprimento de onda é dividido entre região do visível e infravermelho.

Figura 2 - Curva média de reflectância da vegetação fotossinteticamente ativa



Fonte: Araújo (1999)

Os fatores que estão diretamente envolvidos na interceptação e na absorção da radiação fotossinteticamente ativa podem ser classificados em

morfológicos e fisiológicos. Os fatores morfológicos estão relacionados com a organização espacial da planta, ou seja, a densidade de cobertura vegetal, distribuição horizontal e vertical das folhas e o ângulo de inserção foliar (BERNARDES, 1987).

O parâmetro utilizado para estimar a cobertura vegetal é o Índice de Área Foliar (IAF), conforme equação 1.

$$IAF = \frac{AF}{AS} \quad (1)$$

Em que:

AF - Área foliar;

AS - Área de solo.

Este parâmetro, é dado pela área foliar existente em relação à superfície do solo ocupada pela planta. Em espécies de plantas de ciclo curto, como das culturas agrícolas, a taxa de produção de fotossintetizados aumenta em função do aumento do IAF. Quanto maior o IAF de um dossel, espera-se que sua refletância seja menor na região do visível e maior no infravermelho próximo (PONZONI et al., 2012).

No que compete aos fatores fisiológicos ou funcionais da planta, a taxa fotossintética declina fortemente da plântula até sua maturação, sendo que a radiação refletida pela folha em senescência na região do visível é maior do que a folha jovem. As plantas que sofrem estresse hídrico fecham seus estômatos para perder menos água e conseqüentemente diminuem a taxa fotossintética devido a menor absorção de CO₂. A deficiência de nutrientes na planta é outro fator fisiológico que altera a resposta espectral. Em plantas cujas folhas apresentam deficiência em fósforo e cálcio, a absorção de energia na região do infravermelho é menor do que as folhas sem deficiência, entretanto, em folhas deficientes em enxofre, magnésio, potássio e nitrogênio, a absorção é maior do que em folhas normais (BERNARDES, 1987).

De acordo com Karthikeyan et al. (2020) o rendimento de uma determinada cultura agrícola é dependente de seu estágio de crescimento, de seu tipo e saúde das plântulas. Essas características refletem em suas

propriedades espectrais. Uma vegetação é considerada muito reflexiva e absorviva na região do infravermelho próximo e no vermelho do espectro eletromagnético, respectivamente. Portanto, alguma combinação dessas refletividades das duas bandas será sensível à dinâmica da vegetação.

2.4 Índices de vegetação

Os índices de vegetação foram criados para diminuir o trabalho de análise de dados orbitais, por meio da maximização de informações espectrais no menor número de bandas de operação dos sensores. O estudo comportamental de alvos agrícola, através do índice de vegetação é o mais indicado, porque controla os erros introduzidos nas estimativas da radiância pela resposta lambertiana dos alvos (MOREIRA, 2005). Esses, são transformações lineares da reflectância obtidas em duas ou mais bandas, mediante soma, razão, diferença, ou qualquer outra combinação entre bandas.

Entre os diferentes índices de vegetação, o NDVI - Índice de Vegetação da Diferença Normalizada (Normalized Difference Vegetation Index) e o EVI - Índice de Vegetação Melhorado (Enhanced Vegetation Index), são frequentemente usados como parâmetros de sensoriamento remoto em estudos do crescimento das culturas agrícolas e pesquisas relacionadas à produção (HUANG et al., 2019; LIU et al., 2020; SHAMMI; MENG, 2021). Destaca-se ainda, a utilização do NDMI – Índice de Umidade de Diferença Normalizada (Normalized Difference Moisture Index), como complementar ao NDVI, pois permite identificar o conteúdo de água na vegetação (GAO, 1996).

O NDVI é uma ferramenta para monitoramento da vegetação sendo utilizado para construção de perfis temporais, para detectar atividades sazonais e fenológicas, durante o período de crescimento, pico verde, mudanças fisiológicas das folhas e períodos de senescência (PONZONI et al., 2012).

A principal desvantagem do NDVI é a insensibilidade às variações da vegetação sobre certas condições de cobertura da terra. Este não considera o espectro de dependências dos efeitos atmosféricos, as interações do fundo do dossel e anisotropias de reflectância bidirecional do dossel, particularmente aquelas associadas ao sombreamento (HUETE et al., 1999).

O EVI foi desenvolvido para otimizar o sinal da vegetação, melhorando a sensibilidade de sua detecção em regiões com maior densidade de biomassa e para reduzir a influência do sinal do solo e da atmosfera sobre a resposta do dossel (JUSTICE et al., 1998).

Este índice tem melhorado a sensibilidade em regiões de alta biomassa, o que permite melhor desempenho do monitoramento da vegetação, pois reduz a influência da atmosfera e do substrato abaixo do dossel (JENSEN, 2009). Enquanto o NDVI, normalmente, se baseia no recurso de absorção de pigmento nas regiões do vermelho e do infravermelho próximo, o EVI baseia-se nas regiões vermelha, azul e infravermelho próximo do espectro eletromagnético. Comparado ao NDVI, o EVI é menos sensível a diferentes origens do solo (HUETE et al., 2002).

Shammi e Meng (2021), utilizaram os índices de vegetação EVI e NDVI do sensor MODIS como métricas para monitorar o crescimento e gerar modelos preditivos de rendimento da soja em condados no Delta do Mississippi, EUA, onde obtiveram acurácia de 95% para prever a produtividade. Liu et al. (2020), aplicaram o NDVI do sensor MODIS para estimar rendimentos de grãos de cevada, canola e trigo de primavera nas pradarias canadenses por regressão múltipla e, obtiveram no pico de crescimento das culturas, valores de R^2 entre 0,54 a 0,81 em regiões úmidas.

O NDMI é proposto para sensoriamento remoto para monitorar o teor de água das folhas através dos canais do infravermelho próximo e médio (GAO, 1996). Este índice permite identificar numa cultura agrícola as áreas com problemas de estresse hídrico (TALLOOR; MANHAS; KOTHIARY, 2021).

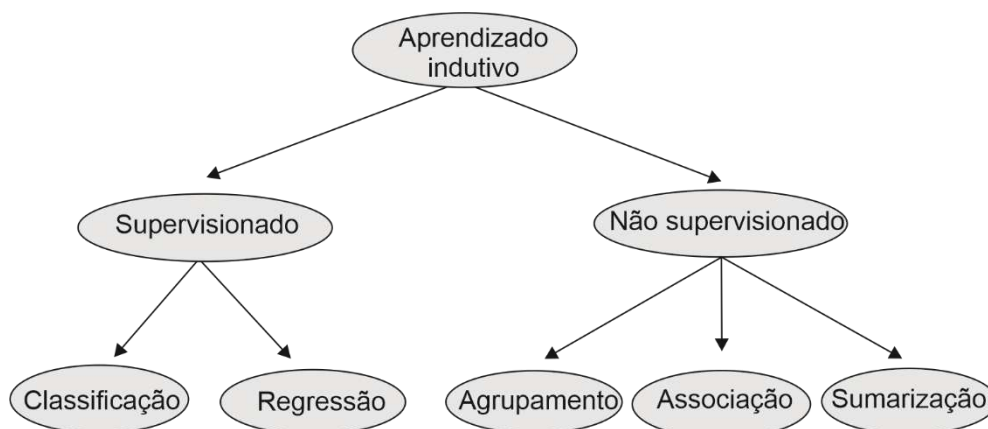
Hendawy et al. (2017), destacam o NDMI como um preditor rápido e de baixo custo para selecionar genótipos de trigo com rendimento superior sob diferentes regimes hídricos. Mimić et al. (2022), utilizaram o valor médio do NDMI para gerar um novo índice de estresse hídrico das culturas de milho e girassol na Sérvia, o que possibilitou avaliar os efeitos da seca na queda do rendimento agrícola.

2.5 Aprendizado de Máquina (Machine Learning) e suas aplicações na agricultura

Nas últimas décadas, a crescente complexidade de problemas a serem tratados e resolvidos computacionalmente e o grande volume de dados gerados por diferentes setores, fez-se necessário a criação de ferramentas computacionais mais sofisticadas, que fossem mais autônomas e, dessa forma, reduzir a intervenção humana e a dependência de especialistas. Para isso, essas técnicas vêm sendo capazes de identificar padrões em seus conjuntos de dados, a partir de experiência passada (MONARD; BARANAUSKAS, 2003). Neste contexto, surge o Aprendizado de Máquina (AM) – *Machine Learning*, naturalmente associado à Inteligência Artificial, que conforme Luger (2013), envolve generalização a partir da experiência.

Algoritmos de AM têm sido utilizados em diversas tarefas que podem ser organizadas de acordo com diferentes critérios, e podem ser aplicados com fins preditivos e descritivos. Em tarefas de previsão, a meta é encontrar uma função, denominada de modelo ou hipótese a partir de dados de treinamento que possa ser utilizada para prever um rótulo ou valor. Em tarefas de descrição, a meta é explorar ou descrever um conjunto de dados, como encontrar grupos de objetos semelhantes (FACELI et al., 2011). A Figura 3 apresenta a hierarquia de aprendizado, em AM.

Figura 3 - Hierarquia de aprendizado



Fonte: Faceli et al. (2011)

As tarefas descritivas, denominadas de não-supervisionadas, são genericamente divididas em agrupamentos, em que os dados são agrupados de acordo com sua similaridade. Sumarização, por outro lado, se caracteriza por uma descrição compacta e de associação, que consiste em encontrar padrões frequentes de associações entre atributos de um conjunto de dados (FACELI et al., 2011). O aprendizado preditivo, também denominado de supervisionado, e as tarefas, se distinguem pelo tipo do rótulo de dados: discreto no caso de classificação e contínuo, no caso de regressão (CERRI; CARVALHO, 2017).

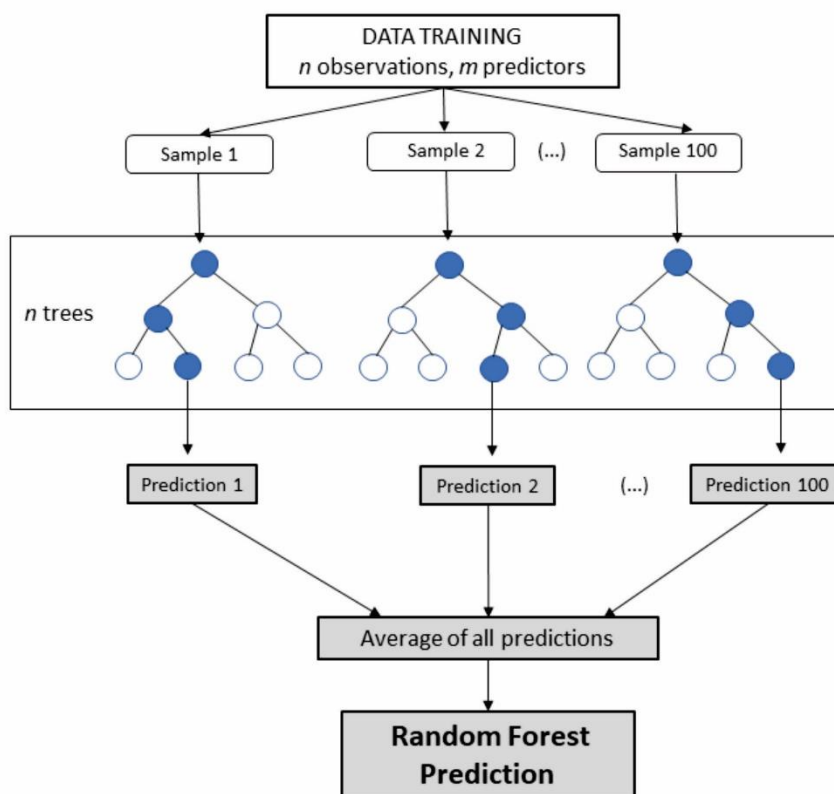
No contexto agrícola, o AM surgiu em conjunto com as tecnologias de BIG DATA e de alto desempenho computacional, que permitiu criar novas oportunidades a fim de desvendar, quantificar e compreender processos intensivos de dados em ambientes operacionais (LIAKOS et al., 2018). Desde então, diversos estudos têm sido realizados aplicando os algoritmos de AM para estimativa de produtividade.

2.5.1 Árvores de decisão

A árvore de decisão constitui um tipo de algoritmo de AM. Essas são representadas como um conjunto de regras que tem seu início na raiz da árvore e caminha até uma das suas folhas. Esta consiste num grafo acíclico direcionado em que cada nó folha corresponde a uma classe ou um nó de decisão, que contém um teste sobre algum atributo (MONARD; ARANAUSKAS, 2003).

O algoritmo de Floresta Randômica na regressão constrói uma infinidade de árvores de decisão no momento do treinamento das amostras, através de uma predição média das árvores individuais (Figura 4). Para Breiman (2001), as florestas aleatórias ou randômicas para regressão são formadas pelo crescimento de árvores dependendo de um vetor aleatório, de modo que o preditor da árvore assuma valores numéricos em oposição aos rótulos de classe.

Figura 4 – Esquema representativo do algoritmo da Regressão por *Random Forest*



Fonte: Adaptado de Rodriguez-Galiano et al. (2016)

James et al. (2013), ressaltam que o *Random Forest* constrói árvores de decisão cada vez que a divisão em uma árvore é considerada, fazendo com que uma amostra aleatória de n preditores são escolhidos como candidatos, divididos do conjunto completo de preditores.

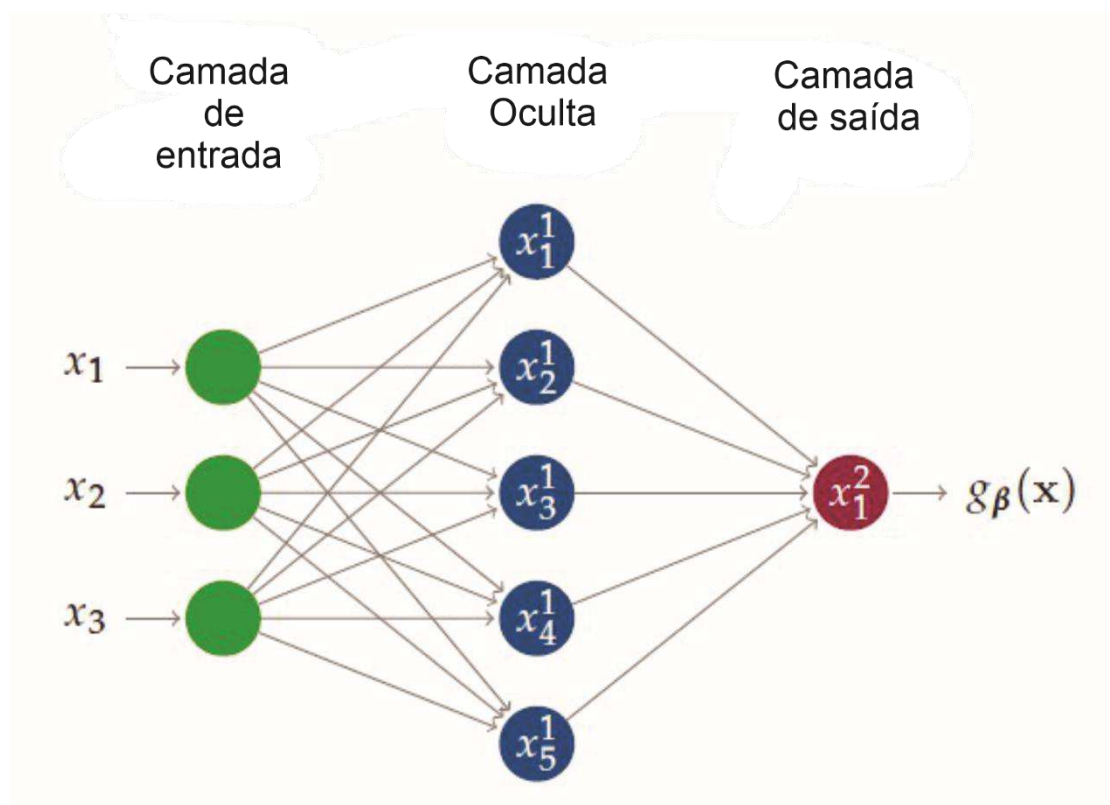
2.5.2 Redes Neurais Artificiais

As Redes Neurais Artificiais se caracterizam por apresentar um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência por meio de técnica computacional. Os procedimentos nestes sistemas são distribuídos por meio de conjuntos ou camadas de neurônios. A solução de um problema é paralela, pois todos os neurônios processam as suas entradas simultaneamente e

independentemente. As informações, que servem como entrada de dados são distribuídas através de nós e camadas de rede (LUGER, 2013).

Na regressão por Redes Neurais Artificiais (RNAs), os modelos lineares adotam como elemento básico uma combinação linear das variáveis predictoras e estas são transformadas por meio de uma transformação não linear. Entretanto, ao invés de usar apenas um desses elementos, eles usam várias camadas de muitos desses elementos. Os resultados de uma camada servem como entradas para a próxima camada. Nesta próxima camada, as entradas são combinadas exatamente da mesma maneira, onde cada elemento forma uma soma ponderada que é transformada de forma não linear (HAND et al., 2001). A estrutura de uma rede neural é apresentada na Figura 5.

Figura 5 - Exemplo de rede neural com uma camada oculta



Fonte: IZBICKI; SANTOS (2020)

A estimação não linear $r(\mathbf{x})$ da rede neural é configurada com entradas ao lado esquerdo da Figura 5, que correspondem as variáveis. Na sequência, existem os chamados nós da camada oculta da rede e cada flecha representa

um peso (parâmetro) b . Os nós nessa camada, representam uma transformação dos nós das variáveis da camada anterior (IZBICKI; SANTOS, 2020).

2.5.3 Máquina de Vetores de Suporte

A metodologia de Máquina de Vetores de Suporte (SVM) é fundamentada na proposição de aprendizado estatístico que estabelece entradas que devem ser seguidas na obtenção de classificadores com boa capacidade de generalização. Essas são consideradas robustas diante de objetos de grande dimensão e na convexidade do problema de otimização formulado durante o treinamento, que implica a existência de um mínimo global (FACELI et al., 2011).

A regressão por SVM é também denominada de SVR (Support Regression Vector), produz um modelo pela classificação do vetor de suporte, dependendo de apenas um subconjunto dos dados de treinamento, pois não se preocupa com pontos que estão além da margem para a previsão do modelo (OLSON; DELEN, 2008).

As SVMs são equivalentes a uma rede neural de duas camadas e possuem uma função de kernel sigmoide. Essa função realiza treinamentos polinomiais alternativos com base radial e os pesos da rede são identificados através de uma programação quadrática com restrições lineares. A entrada de valores apropriados nos parâmetros do SVM tem grande influência em seu desempenho e, portanto, na precisão de sua previsão (BESALATPOUR et al., 2014).

O funcionamento da SVM acontece quando a variável de entrada é primeiramente mapeada para um espaço de recursos de alta dimensão usando a função Kernel, e na sequência, ocorre o tratamento de problemas não lineares em um espaço linear (SALAZAR-ROJAS; CEJUDO-RUIZ; CALVO-BRENES, 2022). A função linear da SVM é dada pela seguinte expressão (LI et al., 2017):

$$y = f(x, \omega) = \sum_{i=1}^m \omega_i \phi(x_i) + b \quad (2)$$

Em que:

$\phi(x_i)$ - Variáveis de entrada após a transformação de Kernel.

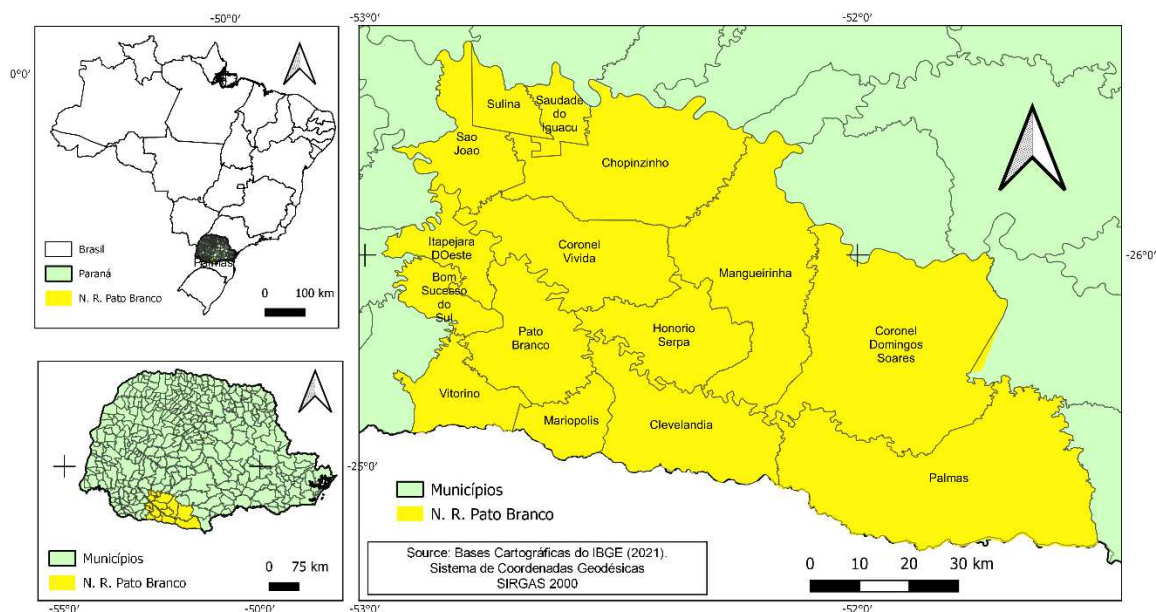
ω_i e b - Coeficientes que são estimados minimizando a função de risco regularizada.

3. MATERIAL E MÉTODOS

3.1 Caracterização da área de estudo

O presente trabalho foi realizado no Núcleo Regional de Pato Branco -PR, divisão estabelecida pela Secretaria de Estado da Agricultura e do Abastecimento – SEAB, formado por 15 municípios pertencentes às mesorregiões sudoeste e centro-sul paranaense (Figura 6).

Figura 6 - Localização do Núcleo Regional Pato Branco - PR



Fonte: Autoria própria (2023)

Estes municípios somam uma área de 9.293,95 Km², onde predomina os climas *Cfa* e *Cfb*, conforme classificação de Köppen. A precipitação média anual histórica no período de 1977 a 2006 variou de 1800 a 2000 mm (ANA, 2013).

3.2 Mapeamento das áreas com a cultura da soja

Para o mapeamento das áreas cultivadas com soja, referente aos anos safras 2019/2020, 2020/2021 e 2021/2022, foram obtidas imagens na faixa do visível e infravermelho dos satélites SENTINEL-2A e 2B (L2A), com resolução espacial de 10 metros e corrigidas no topo da atmosfera. Foram selecionadas imagens geradas entre os meses de setembro e março, período que representa

o intervalo entre a de semeadura à colheita da soja.

Referente ao ano da safra 2019/2020, foi possível obter dez imagens sem presença de nuvens, correspondente às datas de 10/10/2019, 25/10/2019, 19/11/2019, 29/11/2019, 14/12/2019, 29/12/2019, 18/01/2020, 12/02/2020, 22/02/2020 e 13/03/2020. Para a safra 2020/2021, foram obtidas nove imagens, das datas de 24/09/2020, 04/10/2020, 19/10/2020, 08/11/2020, 28/11/2020, 12/01/2021, 06/02/2021, 21/02/2021 e 18/03/2021. Não foi possível obter imagens de qualidade do mês de dezembro de 2020 devido às recorrentes precipitações diárias na região de estudo. Para a safra 2021/2022, foram obtidas dez imagens sem nuvens, nas datas de 09/09/2021, 24/10/2021, 03/11/2021, 13/11/2021, 13/12/2021, 28/12/2021, 12/01/2022, 22/01/2022, 11/02/2022 e 08/03/2022.

Para o mapeamento da área cultivada de soja no Núcleo Pato Branco, foi adaptada a metodologia proposta por Johann et al. (2012), onde produziram-se índices de vegetação EVI e NDVI para as safras de 2019/2020, 2020/2021 e 2021/2022, de modo que contemplasse todo ciclo de desenvolvimento da cultura. As cenas de cada safra foram agrupadas temporalmente, o que permitiu criar um perfil espectral.

Os índices NDVI e EVI são formados por meio das equações 3 e 4:

$$NDVI = \frac{IVP-V}{IVP+V} \quad (3)$$

Em que:

IVP - Infravermelho próximo;

V - Vermelho.

$$EVI = \frac{G(NIR-V)}{(L+NIR+C1 V - C2 A)} \quad (4)$$

Em que:

G - Fator de ganho;

L - Fator de ajuste para o solo;

NIR - Infravermelho próximo;

V - Vermelho;

A - Azul;

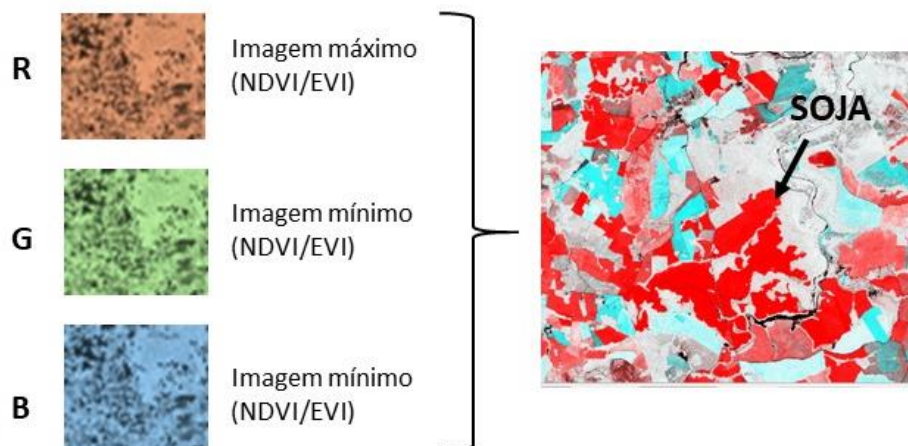
C1 e C2 - Coeficientes de ajuste para o efeito de aerossóis da atmosfera.

Os valores dos coeficientes adotados pelo algoritmo do EVI são: $L = 1$, $C1 = 6$, $C2 = 7,5$ e $G = 2,5$.

Devido à grande variabilidade de datas de semeadura, utilizaram-se imagens de EVI e NDVI para gerar a imagem de mínimo EVI/NDVI e a imagem de máximo EVI/NDVI. Foram utilizadas as cenas de 10/10/2019, 25/10/2019, 19/11/2019, 29/11/2019 e 14/12/2019 da safra 2019/2020, as cenas de 24/09/2020, 04/10/2020, 19/10/2020, 08/11/2020 e 28/11/2020 da safra 2020/2021 e as cenas de 09/09/2021, 24/10/2021, 03/11/2021, 13/11/2021 e 13/12/2021 para a imagem de mínimo EVI/NDVI (menores índices de vegetação da cena), que correspondessem às fases de semeadura e estágio inicial da cultura da soja. As imagens de 29/12/2019, 18/01/2020, 12/02/2020, 22/02/2020 e 13/03/2020 da safra 2019/2020, as cenas de 12/01/2021, 06/02/2021, 21/02/2021 e 18/03/2021 da safra 2020/2021 e as cenas 28/12/2021, 12/01/2022, 22/01/2022, 11/02/2022 e 08/03/2022, utilizou-se para geração da imagem de máximo EVI (maiores valores de EVI dentre as cenas utilizadas), que consideraram o pico vegetativo (final de desenvolvimento, florescimento e formação de colheita).

O Software QGIS 3.10, foi utilizado para gerar uma composição RGB (*Red-Green-Blue*), em que a imagem de máximo EVI/NDVI foi colocada no filtro R (vermelho) e a imagem de mínimo EVI/NDVI foi colocada nos filtros G (verde) e B (azul), tendo-se gerado uma composição colorida na qual as áreas em vermelho representam a cultura da soja e as demais cores, os outros alvos (Figura 7).

Figura 7 - Composição RGB espectro-temporal para a cultura da soja



Fonte: Autoria própria (2023)

Para facilitar o processo de classificação supervisionada de imagem, foram aplicadas máscaras de uso permanente, tais como hidrografia, pastagem, áreas urbanas e cobertura florestal, obtidas no IAT (Instituto de Água e Terra) do estado do Paraná. Essas máscaras (vetores shapefiles) são provenientes do Mapa de uso e cobertura da terra do Estado do Paraná, produzido a partir de imagens WorldView 2 e Pleiades 1A e 1B, na escala 1:25.000, com padrão de exatidão cartográfica PEC A, referente aos anos 2012 e 2013.

Após a aplicação das máscaras de uso permanente sobre o RGB dos índices de vegetação, foram criadas duas classes - classe 1: Soja, e classe 2: Outros. Em seguida, foram gerados dez polígonos para cada classe e 100 pontos aleatórios no interior de cada polígono, totalizando 1000 pontos para cada classe. O conjunto de dados foi dividido em 800 pontos por classe para treinamento e 200 pontos por classe para validação.

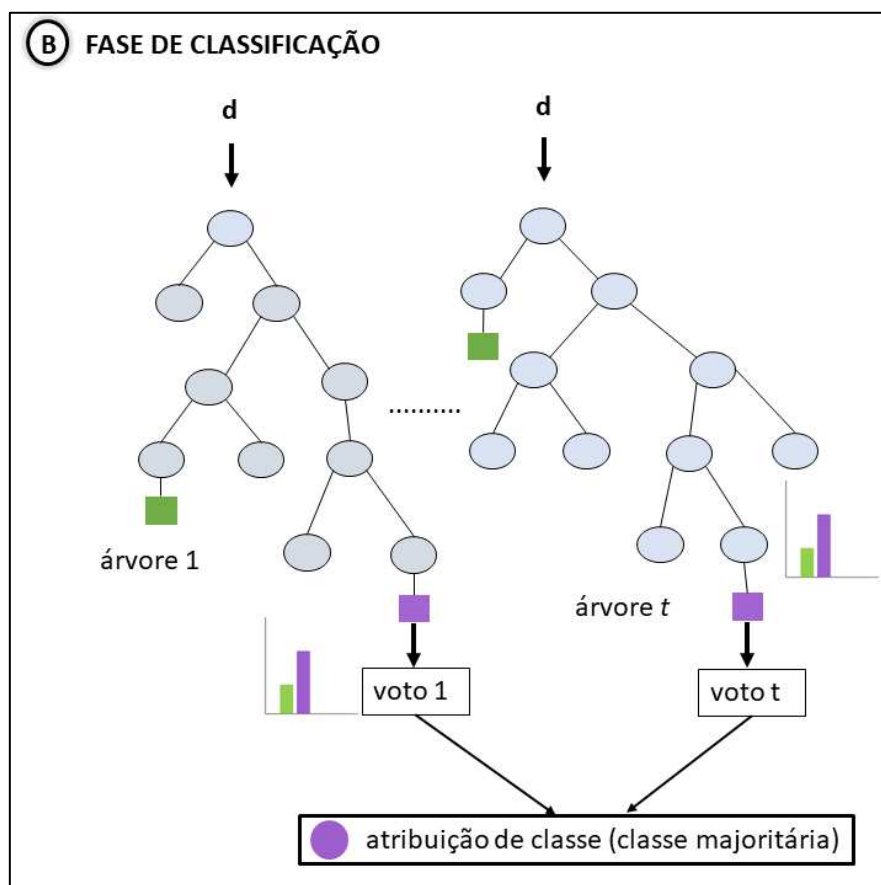
Para a classificação supervisionada de imagem, utilizou-se o algoritmo Random Forest no aplicativo R Studio (R Core Team, 2022), sendo criadas 500 árvores (ntrees) e duas variáveis testadas em cada divisão. Os pacotes utilizados foram: "randomForest", "sf", "tidyverse", "raster", "rgdal", "sp", "magrittr", "dplyr" e "rgeos". O desempenho do algoritmo de classificação foi mensurado pela matriz confusão, que para cada classe, computa a fração de amostras que são corretamente e erroneamente detectadas (ZANOTTA; FERREIRA; ZORTEA, 2019).

3.3 Modelo de classificação Random Forest

Random Forest se tornou um algoritmo bastante difundido na área de sensoriamento remoto devido à precisão de suas classificações. Este possui bom desempenho quando se utiliza base de dados com alta dimensionalidade e multicolinearidade, sendo rápido e insensível ao *overfitting* (BELGIU; DRĂGU, 2016).

O classificador supervisionado Random Forest cria inúmeras árvores independentes, gerando subconjuntos de amostras de treinamento por meio de substituição (uma abordagem de ensacamento) e de variáveis de entrada em cada nó da árvore. A classificação final acontece através de uma abordagem de conjunto por meio de votação majoritária (BREIMAN, 2001), como pode ser visualizado na Figura 8.

Figura 8 - Classificação por Random Forest



Fonte: BELGIU; DRĂGU (2016)

Durante a classificação por Random Forest, quando ocorre o treinamento de cada árvore por um subconjunto aleatório do conjunto de dados original,

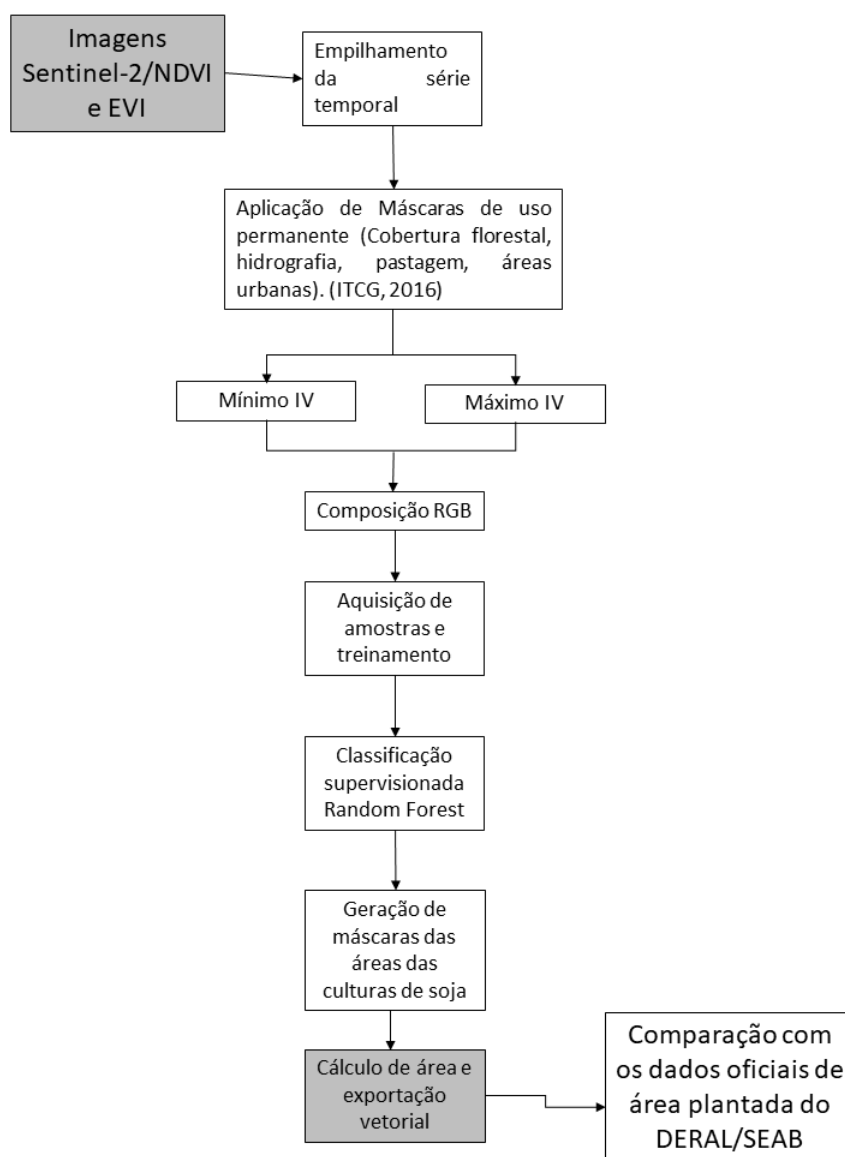
aproximadamente um terço dos elementos são deixados de fora do conjunto de treinamento. Esses elementos externos são denominados *Out Of Bag* (OOB) e correspondem a estimativa da taxa de erro, o que permite avaliar a precisão da classificação (ZHONG; GONG; BIGANDO, 2014).

3.4 Análise de comparação de áreas

Após realizada a classificação, as imagens *raster* foram convertidas em vetores, onde foram extraídas na forma de polígonos (*shapefiles*), e assim, obtidas as máscaras da cultura da soja. A partir destas, calculou-se a área em hectares e realizou-se a comparação com os dados oficiais de área produzida referente aos anos safras 2019/2020, 2020/2021 e 2021/2022, obtidos da SEAB/PR.

O fluxograma da Figura 9 apresenta a síntese desta primeira etapa da metodologia, que visa mapear as áreas com a cultura da soja.

Figura 9 - Fluxograma da metodologia do mapeamento das áreas com a cultura da soja



Fonte: Autoria própria (2023)

3.5 Modelo de estimativa de produtividade

Para a criação de um modelo de estimativa de produtividade, utilizou-se índices de vegetação EVI, NDVI e NDMI gerados a partir de imagens dos sensores MSI do SENTINEL-2 e MODIS do AQUA e TERRA, correspondentes aos meses de setembro a março de cada ano safra, de modo que contemplasse o período entre a semeadura da cultura da soja até o máximo vigor vegetativo. Ressalta-se que para a geração do índice NDMI, ocorreu restauração dos pixels

das imagens do Infravermelho médio tanto para o MODIS quanto ao MSI, a fim compatibilizar a resolução espacial. O índice NDMI é gerado pela equação a seguir:

$$NDMI = \frac{IVP - IVM}{IVP + IVM} \quad (5)$$

Em que:

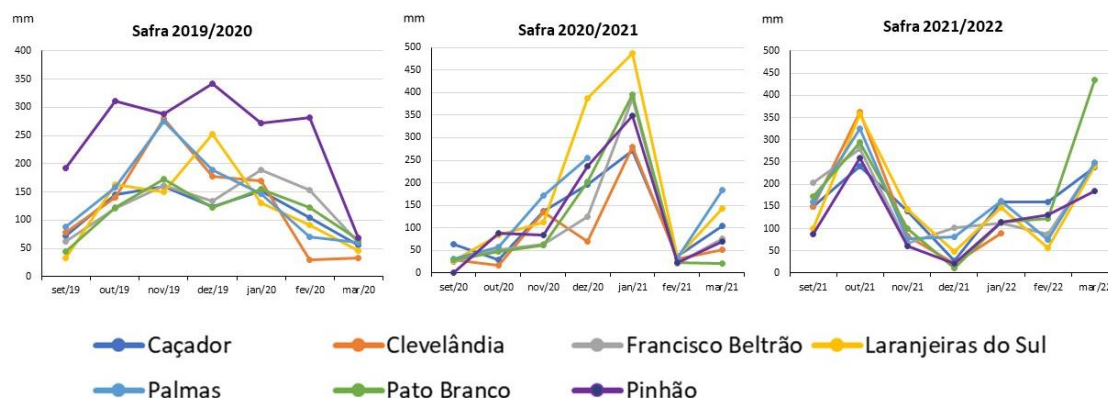
IVP - Infravermelho próximo;

IVM – Infravermelho médio.

Obteve-se também, dados de precipitação, provenientes de sete estações meteorológicas inseridas e próximas ao Núcleo Pato Branco, duas do INMET (Instituto Nacional de Meteorologia), localizadas nos municípios de Clevelândia/PR e Caçador/SC; e cinco do SIMEPAR (Sistema Meteorológico do Paraná), localizadas nos municípios de Francisco Beltrão/PR, Laranjeiras do Sul/PR, Pinhão/PR, Palmas/PR e Pato Branco/PR. Os valores de média diária de chuva de cada estação foram calculados de forma acumulada conforme a data das imagens de satélite, abrangendo dados de setembro a março para cada ano safra. Na sequência, os valores foram interpolados pelo Inverso da Distância ao Quadrado (IDW), no aplicativo QGIS 3.10 e gerado imagens *raster* para extração dos dados precipitação para toda a região de estudo.

A inserção de dados de precipitação no modelo de regressão para estimativa de produtividade foi realizada para avaliar a importância dessa variável nos modelos preditivos, pois a distribuição das chuvas ocorreu de forma irregular quando comparado os períodos das três safras da cultura da soja (Figura 10).

Figura 10 - Distribuição das chuvas nas estações meteorológicas inseridas próximas ao Núcleo Regional Pato Branco



Fonte: INMET (2022); SIMEPAR (2022)

Outros dados, que fizeram também parte do modelo, foram: os tipos de solos e geologia, sendo obtidos no IAT (Instituto de Águas e Terras) do Paraná e dados de altitude e declividade provenientes e disponíveis pelo INPE (Instituto Nacional de Pesquisas Espaciais) através do banco de dados geomorfométricos do Brasil, denominado TOPODATA.

Esses conjunto maior de dados permitiu maior caracterização física da região estudada, que possui predominância de solos do tipo Latossolos, Nitossolos, Chernossolos e Neossolos em regiões com maior declividade (EMBRAPA, 2006); geologia com predominância de derrames basálticos e alguns locais com rochas ácidas do tipo Riólitos e Riodacitos (MINEROPAR, 2006), sendo derivadas da formação Serra Geral na bacia sedimentar do Paraná (NARDY; MACHADO; OLIVEIRA, 2008); e variações de altitude que vão de 400 a 1.300 metros.

Na sequência, obteve-se dados reais referentes aos nove talhões da safra 2019/2020, 24 talhões da safra 2020/2021 e 15 talhões da safra 2021/2022, com áreas que variam de 4 a 199 ha, localizados nos municípios de Pato Branco, Bom Sucesso do Sul, Mangueirinha, Clevelândia e Palmas (Tabela 3). A localização das áreas de estudos e dos respectivos talhões é apresentada na Figura 11.

Tabela 3 - Descrição dos talhões utilizados no estudo

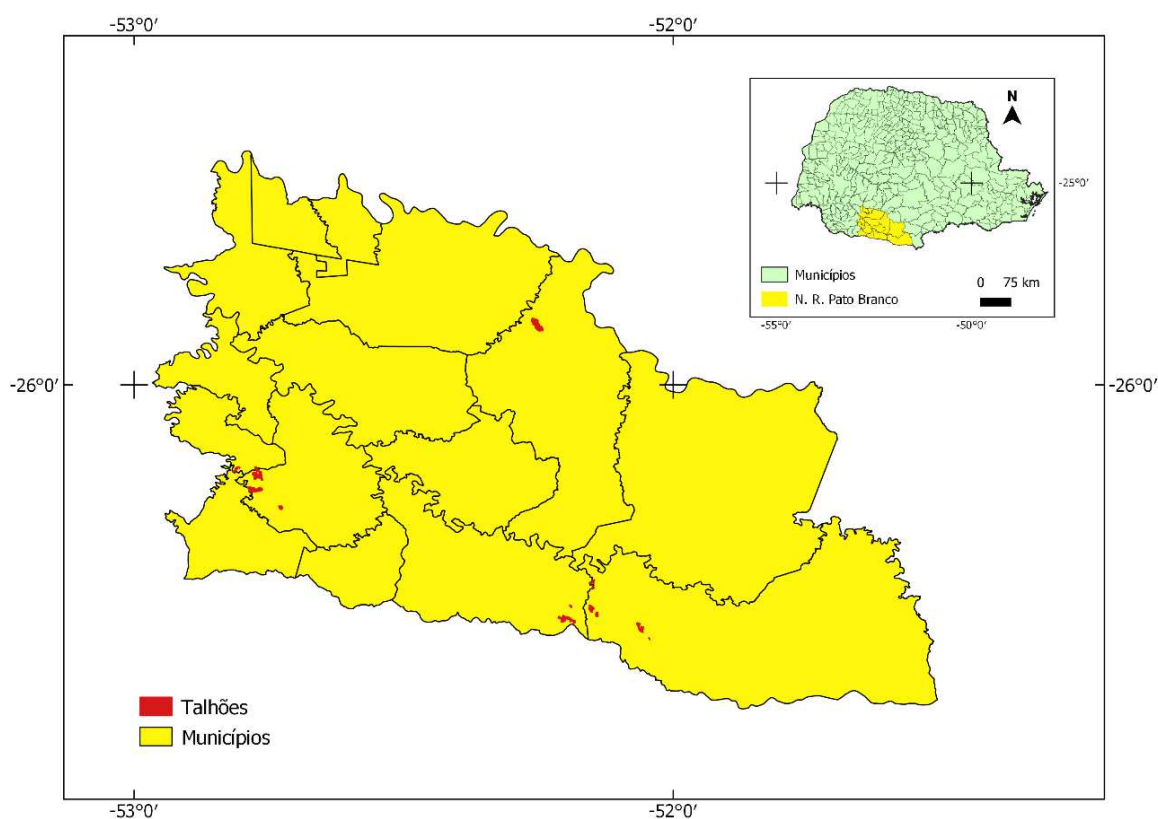
Safra 2019/2020			
Talhão	Área (ha)	Data da colheita	Localização
Talhão 1	65,3	20/01/2020	Manguerinha
Talhão 2 - A	154,2	05/03/2020	Manguerinha
Talhão 2 - B	85,4	22/03/2020	Manguerinha
Talhão 3	15,5	13/03/2020	Pato Branco
Talhão 4	104,7	20/01/2020	Pato Branco
Talhão 5	21,7	20/01/2020	Pato Branco
Talhão 6	47,9	14/01/2020	Pato Branco
Talhão 7	107,7	11/03/2020	Pato Branco
Talhão 8	12,1	13/01/2020	Bom Sucesso do Sul
Safra 2020/2021			
Talhão 1 - A	199,6	28/01/2021	Pato Branco
Talhão 1 - B	39,9	04/02/2021	Pato Branco
Talhão 2 - A	14,9	07/03/2021	Manguerinha
Talhão 2 - B	94,0	22/02/2021	Manguerinha
Talhão 2 - C	37,2	07/02/2021	Manguerinha
Talhão 2 - D	31,6	11/03/2021	Manguerinha
Talhão 3	143,9	15/02/2021	Pato Branco
Talhão 4	106,6	12/02/2021	Pato Branco
Talhão 5	10,7	11/02/2021	Pato Branco
Talhão 6	13,6	07/02/2021	Pato Branco
Talhão 7	55,6	01/04/2021	Palmas
Talhão 8 - A	15,4	18/03/2021	Clevelândia
Talhão 8 - B	10,8	18/03/2021	Clevelândia
Talhão 8 - C	40,6	05/04/2021	Clevelândia
Talhão 9 - A	4,1	15/03/2021	Palmas
Talhão 9 - B	10,9	15/03/2021	Palmas
Talhão 9 - C	7,8	15/03/2021	Palmas
Talhão 10 - A	37,2	08/03/2021	Palmas
Talhão 10 - B	16,6	09/03/2021	Palmas
Talhão 11	8,0	07/04/2021	Clevelândia
Talhão 12 - A	32,5	15/02/2021	Pato Branco
Talhão 12 - B	28,8	07/02/2021	Bom Sucesso do Sul
Talhão 13 - A	7,2	03/02/2021	Pato Branco
Talhão 13 - B	26,1	06/02/2021	Bom Sucesso do Sul

Tabela 3 - Descrição dos talhões utilizados no estudo Safra 2021/2022

(conclusão)

Talhão 1 - A	106,8	21/01/2022	Pato Branco
Talhão 1 - B	38,4	13/01/2022	Pato Branco
Talhão 2	116,9	03/04/2022	Pato Branco
Talhão 3	54,7	02/04/2022	Pato Branco
Talhão 4	13,5	16/01/2022	Pato Branco
Talhão 5	12,1	21/01/2022	Pato Branco
Talhão 6 - A	6,9	14/02/2022	Clevelândia
Talhão 6 - B	4,9	12/02/2022	Clevelândia
Talhão 6 - C	107,8	14/02/2022	Clevelândia
Talhão 6 - D	34,1	14/02/2022	Clevelândia
Talhão 6 - E	14,5	14/02/2022	Clevelândia
Talhão 6 - F	8,1	14/02/2022	Clevelândia
Talhão 6 - G	4,3	14/02/2022	Clevelândia
Talhão 7	16,9	16/01/2022	Pato Branco
Talhão 8	34,0	29/02/2022	Palmas

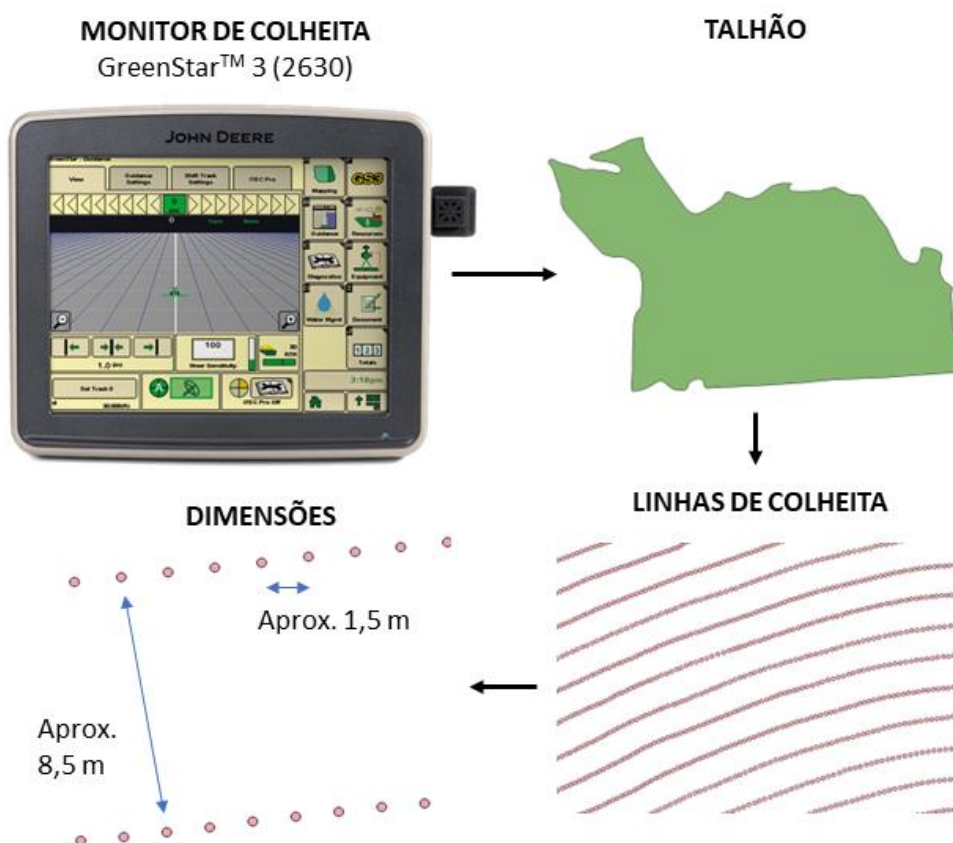
Fonte: Autoria própria (2023)

Figura 11 - Localização dos talhões utilizados para obtenção dos dados reais de produtividade de soja

Fonte: Autoria própria (2023)

Para o levantamento dos dados de produção de soja foram utilizadas três colhedoras em todas as safras, as quais estavam equipadas com monitor de colheita GS3 da John Deere®, calibrados periodicamente pela própria concessionária regional, o que possibilitou obter dados numa dimensão aproximada de 1,5 x 8,5 metros (Figura 12).

Figura 12 - Esquema geral de obtenção dos dados de colheita da soja nos talhões



Fonte: Adaptação de John Deere® (2022)

A partir desses dados, utilizou-se o processo KDD (*Knowledge-Discovery in Databases*) para realizar a mineração de dados por Aprendizado de Máquina.

3.6 Mineração de dados por Aprendizado Máquina

O processo KDD foi utilizado por meio das etapas de pré-processamento, mineração e pós-processamento de dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Com relação a etapa de pré-processamento, a partir dos dados pontuais de produtividade dos talhões, realizou-se o ajuste de granularidade de acordo com as resoluções espaciais das imagens MODIS e MSI, sendo calculado a média para cada pixel de 250 e 10 metros, respectivamente. Após o ajuste, ocorreu a extração dos valores dos índices de vegetação EVI e NDVI de cada sensor, através da ferramenta Estatísticas Zonais do software QGIS 3.4

Isso possibilitou a criação de seis conjuntos de dados, sendo o primeiro do sensor MSI, com 45 atributos e 48.744 instâncias da safra 2019/2020; 41 atributos e 66.390 instâncias da safra 2020/2021 e 45 atributos e 44.995 instâncias da safra 2021/2022; e do sensor MODIS, com 88 atributos e 152 instâncias da safra 2019/2020; 85 atributos e 254 instâncias da safra 2020/2021 e 83 atributos e 223 instâncias da safra 2021/2022. Os *outliers* e extremos valores foram identificados e excluídos no software WEKA 3.8.5 e as variáveis categóricas foram convertidas em *dummies*.

Na sequência, o número de variáveis de cada conjunto de dados foi reduzido a partir da identificação da colinearidade através do pacote *olsrr* do R Studio. Bocard et al. (2011) o fator de inflação de variáveis, denominado VIF (*Variance Inflation Fator*), mensura a proporção pela qual a variância de um coeficiente de regressão é inflada na presença de outras variáveis explicativas. Os autores consideram que VIFs acima de 20 indicam forte multicolinearidade, o que deve ser examinado e evitado se possível. Essa indicação foi aplicada em todos os conjuntos de dados, o que reduziu significativamente o número de variáveis.

3.7 Regressão por Aprendizado de Máquina

A etapa de mineração dos dados foi realizada no R-Studio, onde ocorreu, primeiramente, a divisão dos conjuntos, sendo 70% para treinamento e 30% para teste, através dos pacotes “e1071”, “h2o” e “randomForest”. Os algoritmos de regressão por AM testados foram Floresta Randômica, Máquina de Vetores de Suporte e Redes Neurais, a partir da variável resposta Massa de Produção e das variáveis preditoras tais como os índices de vegetação, solo, geologia, altitude, declividade e precipitação.

Com relação aos hiperparâmetros de configuração dos algoritmos de regressão, em Floresta Randômica foi determinado *n*tree equivalente a 100 árvores e as variáveis preditoras foram de acordo com conjunto de dados. Para Redes Neurais, especificou-se 500 épocas e uma camada oculta e, na Regressão por Máquinas de Vetores de Suporte, utilizou-se o parâmetro *Epsilon* e o kernel do tipo linear.

3.8 Aplicação do modelo para estimativa da produção no Núcleo Regional Pato Branco

Após a testagem dos algoritmos de Aprendizado de Máquina, o modelo de regressão gerado com maior precisão foi aplicado para estimar a produção no Núcleo Regional e nos municípios, a partir do conjunto teste dos dados regionais. Isso permitiu realizar a comparação da produção alcançada pelo modelo com os dados oficiais referente aos anos safra 2019/2020, 2020/2021 e 2021/2022 obtidos da SEAB/PR.

O fluxograma da Figura 13 apresenta a síntese desta segunda etapa da metodologia, que visa estimar a produção de soja.

Figura 13 - Fluxograma da metodologia da estimativa de produtividade de soja para o Núcleo Regional Pato Branco



Fonte: Autoria própria (2023)

4. RESULTADOS E DISCUSSÃO

4.1 Mapeamento das áreas

A classificação das áreas da cultura da soja (2.000 instâncias), foi avaliada pela matriz de confusão e pela precisão, tanto para o conjunto de treinamento, quanto para o conjunto de teste. Os índices de vegetação NDVI e EVI apresentaram resultados similares na classificação das áreas de cultivo da soja.

Na safra 2019/2020, a taxa de erro foi de 0,06% para o NDVI e de 0% para o EVI. Para a safra de 2020/2021 os valores obtidos para a taxa de erro de classificação foram 0% tanto para EVI, quanto NDVI e, para a safra 2021/2022, a taxa de erro foi de 0% para NDVI e de 0,5% para o EVI, todas para o conjunto treinamento. Esses resultados podem ser visualizados de forma detalhada nas Tabelas 4, 5 e 6.

Tabela 4 - Matriz de confusão e porcentagem de erros para conjunto treinamento e teste das máscaras de NDVI e EVI da cultura da soja, Núcleo Pato Branco - Safra 2019/2020. UTFPR, Pato Branco - PR, 2023

Estimativa OOB da taxa de erro						
Estimativa	NDVI			EVI		
	Soja	Outros	Erro	Soja	Outros	Erro
Soja	799	1	0,00125%	800	0	0%
Outros	0	800	0%	0	800	0%
Taxa de erro	0,06%			0%		
Taxa de erro do conjunto teste						
Estimativa	NDVI			EVI		
	Soja	Outros	Erro	Soja	Outros	Erro
Soja	199	1	0,005%	200	0	0%
Outros	0	200	0%	0	200	0%
Taxa de erro	0,25%			0%		

Fonte: Autoria própria (2023)

Tabela 5 - Matriz de confusão e porcentagem de erros para conjunto treinamento e teste das máscaras de NDVI e EVI da cultura da soja, Núcleo Pato Branco - Safra 2020/2021. UTFPR, Pato Branco - PR, 2023

Estimativa OOB da taxa de erro						
Estimativa	NDVI			EVI		
	Soja	Outros	Erro	Soja	Outros	Erro
Soja	800	0	0%	800	0	0%
Outros	0	800	0%	0	800	0%
Taxa de erro	0%			0%		

Taxa de erro do conjunto teste						
Estimativa	NDVI			EVI		
	Soja	Outros	Erro	Soja	Outros	Erro
Soja	200	0	0%	200	0	0%
Outros	0	200	0%	0	200	0%
Taxa de erro	0%			0%		

Fonte: Autoria própria (2023)

Tabela 6 - Matriz de confusão e porcentagem de erros para conjunto treinamento e teste das máscaras de NDVI e EVI da cultura da soja, Núcleo Pato Branco - Safra 2020/2021. UTFPR, Pato Branco - PR, 2023

Estimativa OOB da taxa de erro						
Estimativa	NDVI			EVI		
	Soja	Outros	Erro	Soja	Outros	Erro
Soja	800	0	0%	797	3	0,00375%
Outros	0	800	0%	5	795	0,00625%
Taxa de erro	0%			0,5%		

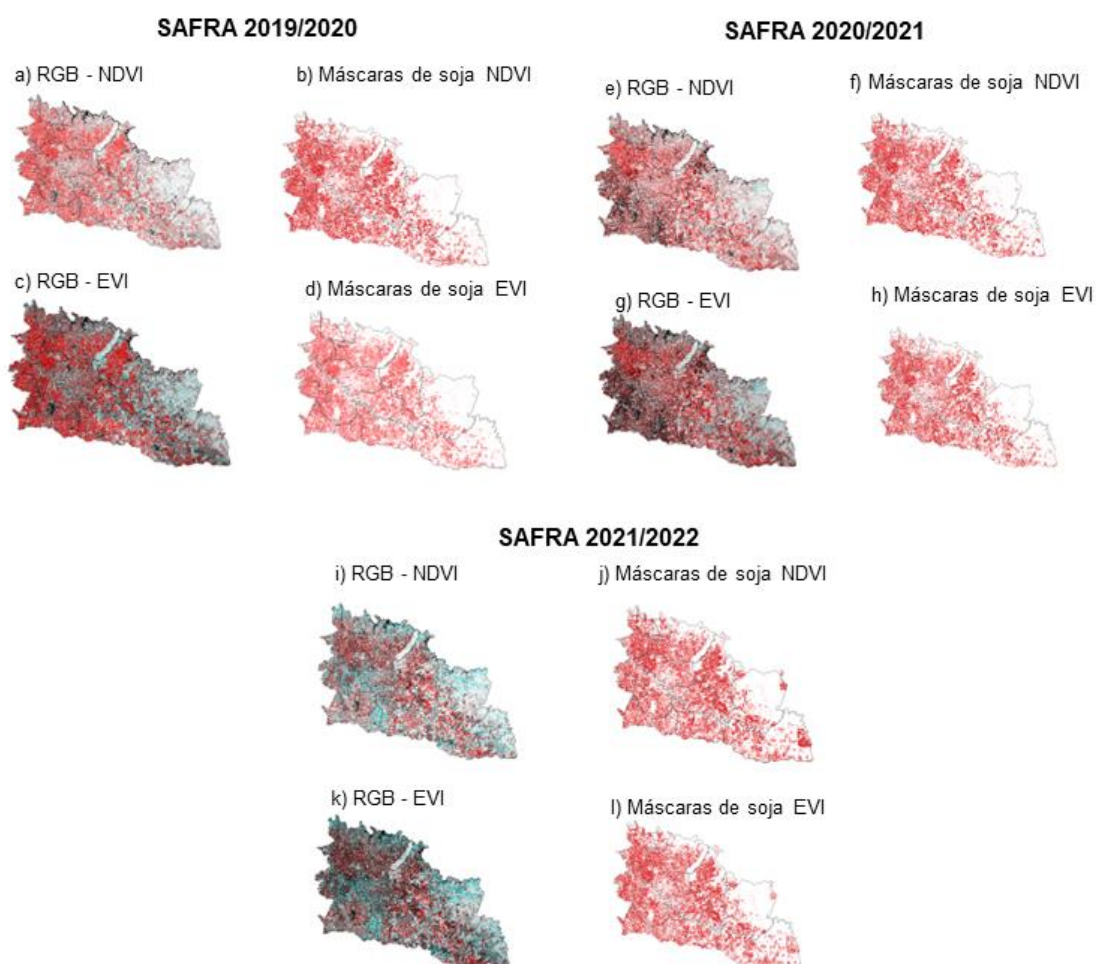
Taxa de erro do conjunto teste						
Estimativa	NDVI			EVI		
	Soja	Outros	Erro	Soja	Outros	Erro
Soja	200	0	0%	195	5	0,025
Outros	0	200	0%	2	198	0,010
Taxa de erro	0%			1,75%		

Fonte: Autoria própria (2023)

Wardlow e Egbert (2010), compararam esses mesmos índices obtidos do sensor MODIS para mapear as culturas agrícolas das planícies centrais dos Estados Unidos, onde obtiveram resultados semelhantes com os resultados levantados neste estudo.

A partir da classificação das áreas-alvos utilizando-se os canais RGBs e índices de vegetação EVI e NDVI, foram geradas máscaras nas imagens para identificação da cultura da soja no Núcleo de Pato Branco (Figuras 14b, 14d, 14f, 14h, 14j e 14l).

Figura 14 - Distribuição espacial da cultura da soja no Núcleo Pato Branco. UTFPR, Pato Branco - PR, 2023



Fonte: Autoria própria (2023)

A análise entre os dados da área estimada (máscara) e da área observada (DERAL/SEAB) foi realizada para as três safras estudadas (Tabela 7).

Tabela 7 - Informações de área oficial (DERAL/SEAB) e obtidas pela classificação para o Núcleo Pato Branco - PR. UTFPR, Pato Branco - PR, 2023

IV	Variáveis	2019/2020	2020/2021	2021/2022
NDVI	Área Deral/SEAB (ha)	324.950	328.080	340.480
	Área classificação (ha)	317.418	317.361	328.346
	Diferença de área (ha)	-7.532	- 10.719	12.134
	Diferença de área (%)	-2,31%	- 3,26%	- 3,69%
EVI	Área Deral/SEAB (ha)	324.950	328.080	340.480
	Área classificação (ha)	305.814	294.854	327.032
	Diferença de área (ha)	-19.136	-33.226	-13.448
	Diferença de área (%)	-5,88%	-10,12%	- 3,94%

Fonte: DERAL/SEAB (2023) e Autoria própria (2023)

Para a cultura da soja, as áreas obtidas pelas máscaras foram menores que às áreas oficiais do DERAL/SEAB. Essa diferença pode estar relacionada à questão temporal das máscaras de uso permanente do solo que foram aplicadas sobre as imagens, as quais foram produzidas pelo ITCG com dados dos anos 2012 e 2013.

As máscaras da cultura da soja obtidas a partir da classificação espectro-temporal de imagens do satélite SENTINEL-2/NDVI para o Núcleo Pato Branco, apresentaram resultado melhor do que as imagens do SENTINEL-2/EVI, quando comparadas as suas áreas com os dados oficiais do DERAL/SEAB. Resultado similar foi encontrado na pesquisa de Feyisa et al. (2020), em que o índice do MODIS/NDVI apresentou um desempenho superior ao MODIS/EVI no mapeamento das áreas agrícolas do Nianmar, com precisão de 95% através do classificador Random Forest. Ao comparar esses índices, os autores ressaltam que o EVI é projetado para suprimir sinais de superfícies sem vegetação, o que leva a diminuir sua precisão nas classes não agrícolas. Johann et al. (2012), também realizaram o mapeamento das culturas de verão do estado do Paraná com imagens MODIS/EVI, e as diferenças entre os dados oficiais e os classificados ficaram em 11,48% para a safra 2004/2005 e 10,8% para 2005/2006.

4.2 Modelos de estimativa de produtividade de soja

A Tabela 8 apresenta as métricas de desempenho dos algoritmos de Aprendizado de Máquina denominados *Random Forest*, *LibSVM* e *MultiLayer Perceptron*. Estes foram aplicados aos conjuntos de dados a partir da divisão aleatória, sendo 70% treinamento e 30% teste. O modelo de regressão foi gerado a partir do conjunto treinamento e avaliado o desempenho sobre o conjunto teste (BARI; CHAOUCHI, JUNG, 2019), na sequência, esse modelo foi utilizado para realizar as predições sobre os conjuntos de dados regionais.

Tabela 8 - Regressão com os algoritmos de Aprendizado de Máquina. UTFPR, Pato Branco - PR, 2023

Algoritmo	Sensores	R ² Treinamento			R ² Teste		
		Safras					
		19/20	20/21	21/22	19/20	20/21	21/22
Random Forest	MSI	0,91	0,91	0,92	0,60	0,58	0,62
	MODIS	0,84	0,95	0,91	0,33	0,62	0,53
LibSVM	MSI	0,35	0,19	0,29	0,34	0,19	0,27
	MODIS	0,22	0,58	0,47	0,16	0,40	0,27
MultiLayer Perceptron	MSI	0,32	0,20	0,26	0,32	0,20	0,26
	MODIS	0,41	0,64	0,52	0,10	0,27	0,10

Fonte: Autoria própria (2023)

O algoritmo de melhor desempenho no modelo de regressão para estimativa de produtividade foi o *Random Forest*, tanto para os índices de vegetação do sensor MODIS quanto para o sensor MSI do SENTINEL-2 (Tabela 8), pois apresentam o intervalo de R² 0,84 a 0,95 no conjunto treinamento e o intervalo R² 0,33 a 0,62 no conjunto teste. James et al. (2013), enfatizam que é difícil determinar o que é um bom valor de R² numa regressão por AM, e que em geral, isso vai depender da aplicação, como por exemplo, em áreas relacionadas à biologia, o modelo linear é apenas uma aproximação, pois alguns fatores de impossível mensuração costumam afetar o modelo.

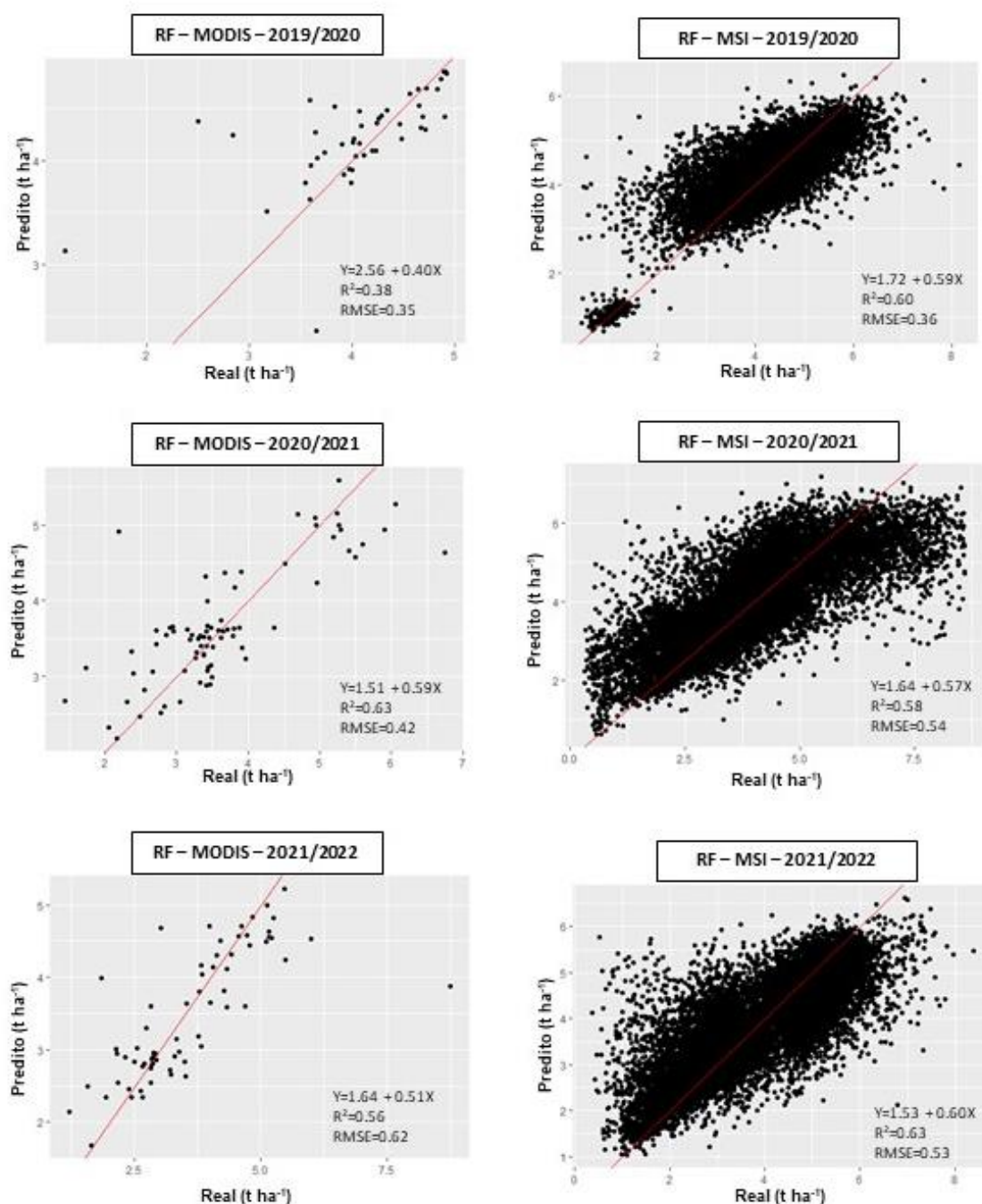
As pesquisas de Alabi et al. (2022) e Khanal et al. (2018) também demonstram a melhor performance do algoritmo RF em relação aos demais nas estimativas de rendimento agrícola. Dentre as possibilidades dos demais

algoritmos não terem tido bom desempenho, Chergui (2022) cita que as redes neurais possuem a desvantagem de não suportar problemas mínimos locais no momento de determinar a estrutura adequada da rede, além do alto custo computacional; e as máquinas de vetores de suporte, de terem sensibilidade a ruídos e não ser adequada a grandes conjuntos de dados.

Rodriguez-Galiano et al. (2016) afirmam que a robustez da regressão por Random Forest se dá pela capacidade de identificar e modelar as complexas relações não lineares, pois lidam com um grande número de preditores, além de determinar o grau de importância de cada um na construção do modelo. Jeong et al. (2016), destacam o potencial desse algoritmo de regressão para a estimativa de safras globais e regionais.

De posse do conjunto de dados, foi possível gerar modelos lineares de dispersão do conjunto teste e comparar a produtividade real dos talhões com a estimada pelo RF, como demonstra a Figura 15.

Figura 15 - Dispersão da produtividade real e predita de soja por Random Forest. UTFPR, Pato Branco - PR, 2023



Fonte: Autoria própria (2023)

Os modelos gerados com os índices de vegetação do sensor MSI do SENTINEL-2, apresentaram R² 0,60 para a safra 2019/2020, R² 0,58 para a safra 2020/2021 e R² 0,63 para a safra 2021/2022. Para os dados do sensor MODIS, os coeficientes de correlação de 0,38 para a safra 2019/2020, R² 0,63

para 2020/2021 e R^2 0,56 para a safra 2021/2022. Os valores da raiz quadrada do erro médio (*Root Mean Squared Deviation* - RMSE) demonstram que os modelos gerados a partir de dados do MSI foram parecidos aos obtidos pelo sensor MODIS.

Esses resultados demonstraram previsibilidade similares que alguns estudos anteriores relacionados ao NDVI e EVI no rendimento de safras. Por exemplo, Johnson (2014), apresentou um modelo de estimativa para as safras de 2006 a 2012 com um R^2 que variou de 0,46 a 0,71 na cultura da soja. Liu et al. (2020), aplicaram o NDVI do MODIS para estimar rendimentos de grãos de cevada, canola e trigo, onde obtiveram no pico de crescimento das culturas, valores de R^2 entre 0,54 a 0,81 em regiões úmidas através de regressões lineares múltiplas. Em contrapartida, Shami e Meng (2021), alcançaram um R^2 de 0,95 ao estimar o rendimento da soja no delta do Mississippi quando utilizaram métricas de crescimento derivadas do NDVI e EVI do MODIS, porém a variável resposta produção empregada foi em nível de condado.

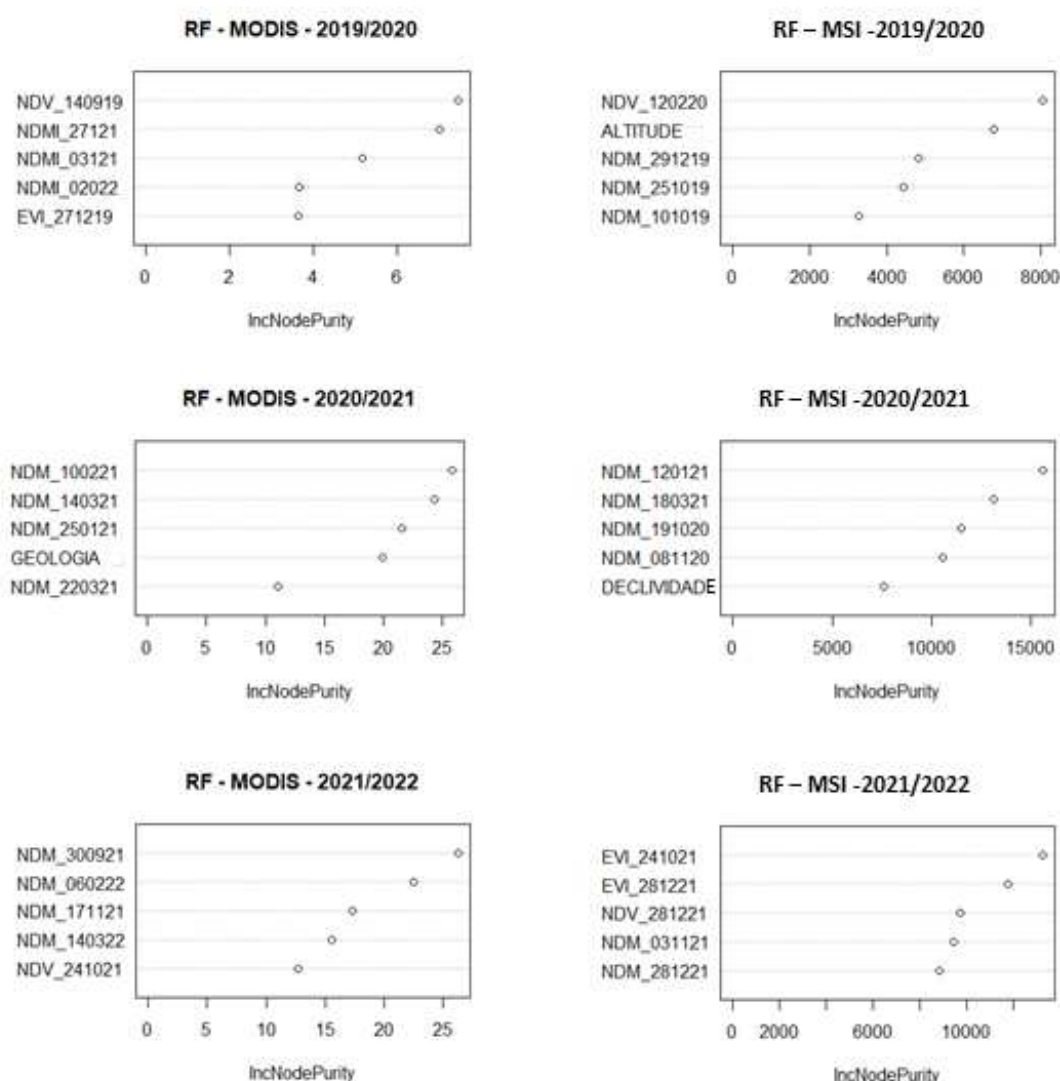
Dado et al. (2020) testaram a estimativa de rendimento de soja através do *Random Forest* para o Meio-Oeste dos Estados Unidos durante o período de 2008-2018, onde utilizaram dados dos satélites SENTINEL-2 e Landsat, dados climáticos e mapas de produção obtidos de monitores de colheita. Entretanto, os resultados obtidos foram inferiores ao encontrados na presente pesquisa, pois os autores obtiveram um R^2 de 0,45.

Nesse sentido, ainda cabe destacar a recente utilização dos Veículos Aéreos Não-Tripulados (VANTs) no monitoramento de culturas agrícolas, que permitem maior autonomia no momento de obtenção de imagens e principalmente sua alta resolução espacial. Entretanto, no estudo realizado por ALABI et al. (2022), no qual monitoraram um talhão de soja localizado na Nigéria, com imagens de 12 cm de resolução espacial e algoritmos como Random Forest e modelo cúbico, os autores obtiveram um R^2 de 0,89, que é superior ao obtido nesse trabalho. Isso permite concluir que a alta resolução espacial da imagem determina a obtenção de um modelo mais preciso.

Além dos dados gerados, também foi possível avaliar quais foram as cinco variáveis que mais contribuíram para gerar os modelos de regressão por RF para a estimativa de produtividade. O ranqueamento da importância das variáveis extraído pelo algoritmo RF é calculado *IncNodePurity*, correspondente à

diminuição total nas impurezas do nó a partir da divisão nos preditores, usando a função de importância no pacote '*randomForest*'. Um aumento no valor de *IncNodePurity* implica uma diminuição no erro quadrático médio, o que significa que os maiores valores representam as variáveis mais importantes para a resposta (HABIBI et al., 2021). Os resultados obtidos para as variáveis utilizadas neste estudo são demonstrados na Figura 16.

Figura 16 – Ranqueamento das variáveis preditoras para cada modelo gerado pelos sensores MODIS e MSI, para predição da estimativa de produtividade da cultura da soja por meio do algoritmo RF. UTFPR, Pato Branco - PR, 2023



Fonte: Autoria própria (2023)

Nota-se que os índices de vegetação EVI, NDVI e NDMI foram importantes para gerar os modelos de estimativa de produtividade da soja. Os

dados tipos de solo e precipitação foram excluídos do modelo na etapa do pré-processamento, pois foram enquadrados como *outliers* e como variáveis inflacionadas, respectivamente.

A partir dos dados processados, comparou-se a produtividade estimada através do conjunto teste com a real dos próprios talhões, por meio de estatísticas descritivas nas Tabelas 9 e 10.

Tabela 9 - Estatísticas descritivas da produtividade de soja ($t\ ha^{-1}$) real dos talhões e da estimativa do Núcleo Regional Pato Branco, por Aprendizado de Máquina com dados do sensor MSI. UTFPR, Pato Branco - PR, 2023

Estatísticas	Produtividade Real ($t\ ha^{-1}$)			Produtividade Estimada por Random Forest ($t\ ha^{-1}$)		
	Safras					
	19/20	20/21	21/22	19/20	20/21	21/22
Mínima	0,453	0,316	0,358	0,695	0,619	1,033
1º quartil	3,828	3,139	2,762	3,931	3,318	3,085
Mediana	4,347	3,678	4,043	4,290	3,653	3,978
3º quartil	3,828	4,363	5,016	3,931	4,270	4,752
Máxima	8,150	8,593	8,403	6,477	7,195	6,626
Média	4,298	3,835	3,881	4,300	3,843	3,900
DP	0,900	1,269	1,366	0,692	0,950	1,047
CV	0,209	0,331	0,352	0,160	0,247	0,268

Fonte: Autoria própria (2023)

Tabela 10 - Estatísticas descritivas da produtividade de soja ($t\ ha^{-1}$) real dos talhões e da estimativa do Núcleo Regional Pato Branco, por Aprendizado de Máquina com dados do sensor MODIS. UTFPR, Pato Branco - PR, 2023.

Estatísticas	Produtividade Real			Produtividade Estimada por		
	(t ha ⁻¹)			Random Forest (t ha ⁻¹)		
	19/20	20/21	21/22	19/20	20/21	21/22
Mínima	1,205	1,439	1,205	2,371	2,176	1,680
1º quartil	3,731	3,045	2,712	4,036	3,141	2,809
Mediana	4,085	3,426	3,381	4,279	3,544	3,184
3º quartil	4,650	3,904	4,371	4,481	4,236	4,308
Máxima	4,927	6,751	8,753	4,848	5,591	5,222
Média	4,073	3,634	3,587	4,215	3,680	3,487
DP	0,689	1,040	1,270	0,450	0,780	0,867
CV	0,169	0,286	0,354	0,106	0,212	0,248

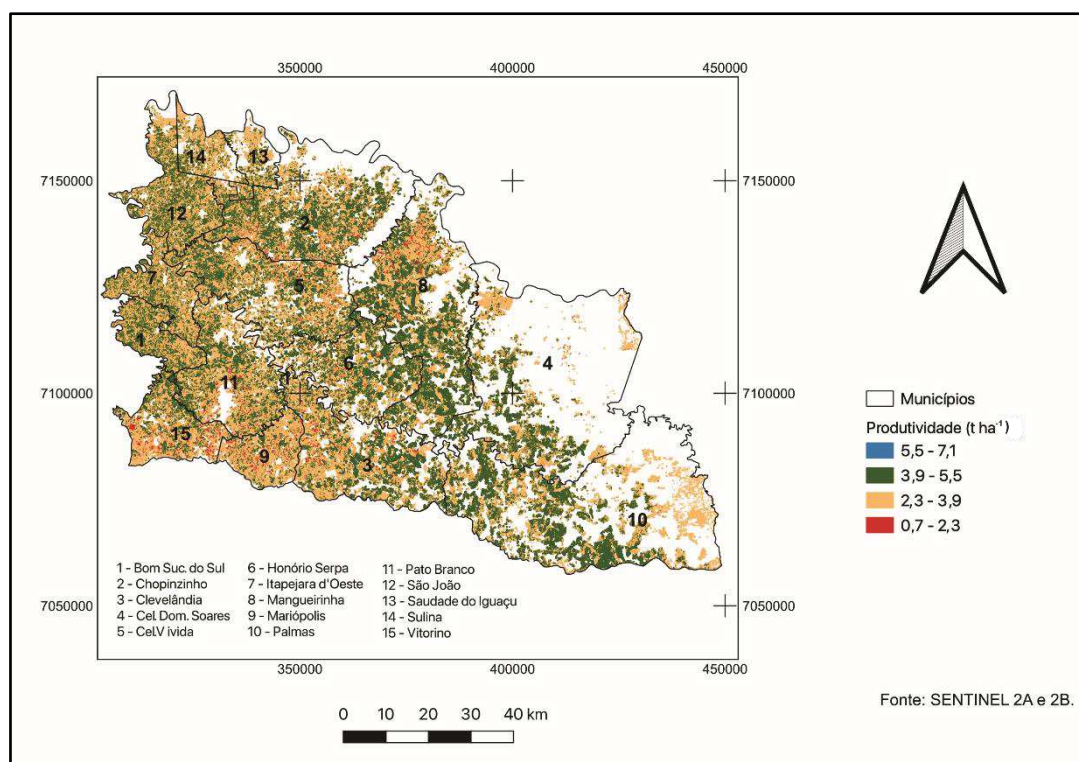
Fonte: Autoria própria (2023)

A média de produtividade real e estimada para os talhões avaliados apresentou diferença de 2 kg ha⁻¹ para a safra 2019/2020, 8 kg ha⁻¹ para a safra 2020/2021 e 19 kg ha⁻¹ para a safra 2021/2022 com os dados do sensor MSI do SENTINEL-2. Para os dados do sensor MODIS as diferenças das médias encontradas foram de 142 kg ha⁻¹ para a safra 2019/2020, 43 kg ha⁻¹ para a safra 2020/2021 e 100 kg ha⁻¹ para a safra 2021/2022. A estimativa gerada pelo RF apresentou diferença nas médias de produtividade entre os sensores, indicando maior precisão com o sensor MSI.

4.3 Estimativa da produtividade no Núcleo Regional Pato Branco

A partir dos modelos obtidos com os índices de vegetação do sensor MSI do SENTINEL-2, foi possível estimar a produtividade no Núcleo Regional de Pato Branco mediante a aplicação dos mesmos sobre as máscaras de áreas obtidas por NDVI. Os resultados para as safras 2019/2020, 2020/2021 e 2021/2022, podem ser visualizados nas Figuras 17, 18 e 19 e foram estimados para cada pixel com área de 100 m².

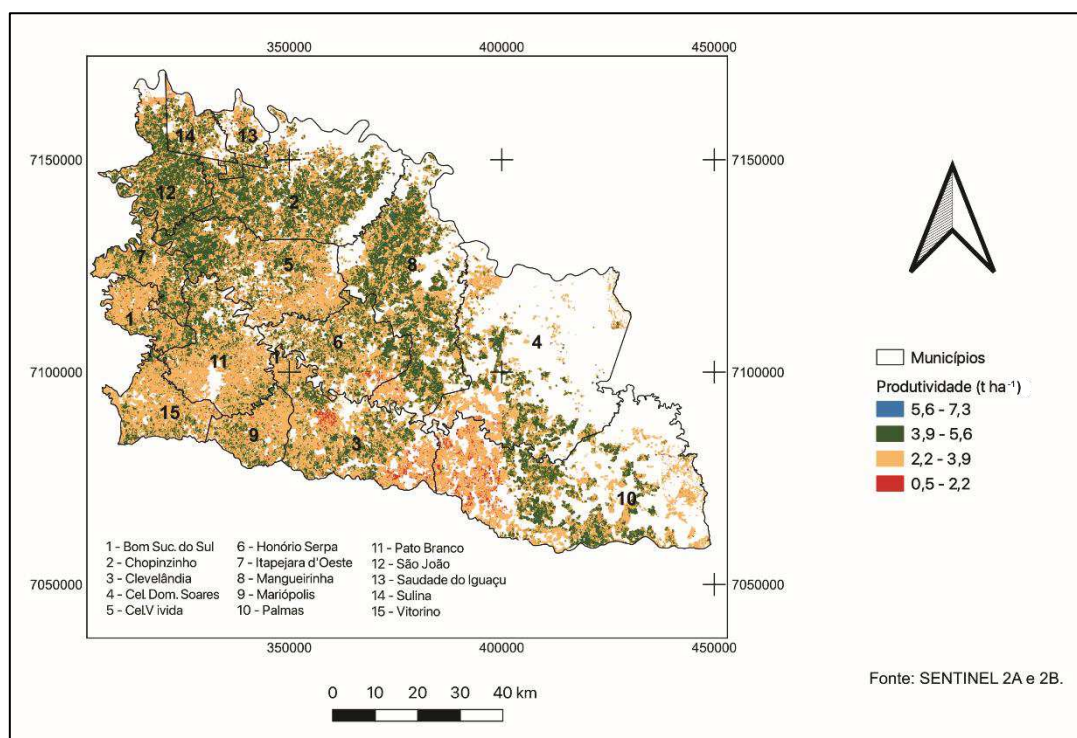
Figura 17 - Mapa da distribuição da estimativa de produtividade da soja ($t\ ha^{-1}$) no Núcleo Regional de Pato Branco – Safra 2019/2020. UTFPR, Pato Branco - PR, 2023



Fonte: Autoria própria (2023)

Para a safra 2019/2020, predominou a produtividade média acima de $3,9\ t\ ha^{-1}$ para a maioria dos municípios do Núcleo Regional de Pato Branco (Figura 17). Enquanto, para a safra 2020/2021, a produtividade foi inferior a $2,2\ t\ ha^{-1}$ em regiões localizadas nos municípios de Clevelândia, Manguaçu e Palmas (Figura 18).

Figura 18 - Mapa da distribuição da estimativa de produtividade da soja ($t\ ha^{-1}$) no Núcleo Regional de Pato Branco – Safra 2020/2021. UTFPR, Pato Branco - PR, 2023



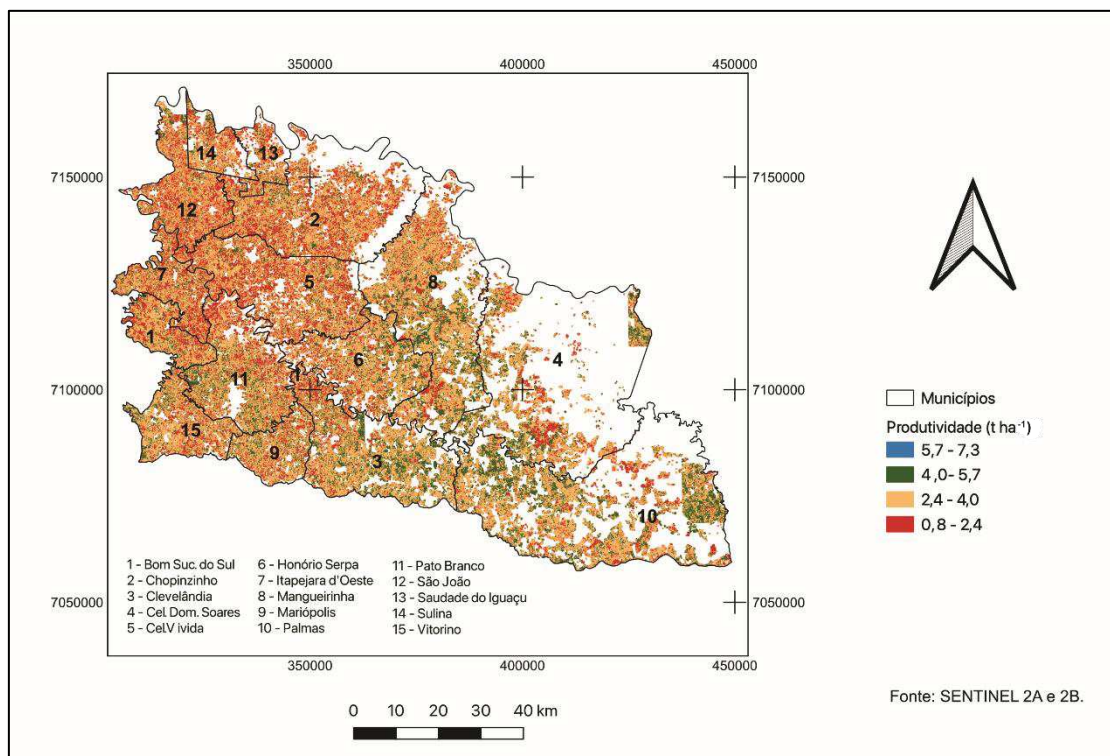
Fonte: Autoria própria (2023)

Para a safra 2021/2022, devido às fortes estiagens ocasionadas pelo fenômeno *La Niña*, a maioria dos municípios do Núcleo Pato Branco apresentou produtividade inferior a $2,4\ t\ ha^{-1}$, com ocorrência de precipitações somente na época de plantio e emergência da cultura, sendo o restante do ciclo marcado pela falta de chuvas conforme Figura 10.

O fenômeno *La Niña* quando comparado com o *El Niño*, afeta o Brasil de maneiras opostas, pois as chuvas e temperaturas no sul do Brasil são reduzidas, fazendo com que as massas de ar mais frias e secas cheguem à região com maior frequência. Em contrapartida, nas regiões norte e nordeste do país, as chuvas ocorrem de forma distribuídas, o que permite o desenvolvimento mais favorável para a agricultura (HINNAH et al., 2020).

Lizume et al. (2014), destacam os impactos negativos que o fenômeno *La Niña* causa na produção de soja e milho na América do Sul, pois estão associados às condições mais quentes e secas, causando perdas de até 13% da área colhida.

Figura 19 - Mapa da distribuição da estimativa de produtividade da soja ($t\ ha^{-1}$) no Núcleo Regional de Pato Branco – Safra 2021/2022. UTFPR, Pato Branco - PR, 2023



Fonte: Autoria própria (2023)

A produção oficial informada para o Núcleo Pato Branco foi de 1.285.770 toneladas para a safra 2019/2020, 1.223.946 toneladas para a safra 2020/2021 e apenas 820.882 toneladas para a safra 2021/2022, o que acarretou uma diferença entre dados reais e estimados de -6,9%, -8,8% e +24,1%, respectivamente (Tabela 11).

Tabela 11 - Comparação entre produção oficial e estimada de soja (t) por município do Núcleo Regional de Pato Branco. UTFPR, Pato Branco - PR, 2023

Municípios	Produção (t) oficial (SEAB/PR)			Produção (t) Estimada (Random Forest)			Diferença %		
	19/20	20/21	21/22	19/20	20/21	21/22	19/20	20/21	21/22
Bom S. do Sul	53.750	52.080	32.310	50.797	39.763	29.740	-5,5	-23,7	-8,0
Chopininho	112.000	109.150	50.100	118.985	116.830	90.052	6,2	7,0	79,7
Clevelândia	124.800	136.875	89.147	108.883	101.482	98.598	-12,8	-25,9	10,6
Cel. D. Soares	61.380	62.640	68.400	72.428	64.486	77.800	18,0	2,9	13,7
Coronel Vivida	106.668	102.120	50.575	114.616	109.758	85.906	7,5	7,5	69,9
Honório Serpa	77.040	79.180	60.500	81.035	76.026	72.521	5,2	-4,0	19,9
Itap. D'Oeste	60.000	57.190	13.410	53.441	51.198	35.284	-10,9	-10,5	163,1
Mangueirinha	184.800	164.775	152.460	164.834	163.682	149.345	-10,8	-0,7	-2,0
Mariópolis	53.095	55.440	33.605	41.063	41.958	27.883	-22,7	-24,3	-17,0
Palmas	123.200	113.400	125.800	124.190	104.924	146.620	0,8	-7,5	16,6
Pato Branco	117.000	95.200	50.200	93.773	79.983	67.781	-19,9	-16,0	35,0
São João	85.234	84.101	42.050	79.615	81.448	58.537	-6,6	-3,2	39,2
S. do Iguaçu	11.340	12.000	2.325	14.488	12.052	13.233	27,8	0,4	469,2
Sulina	18.500	19.795	9.720	21.481	18.906	18.113	16,1	-4,5	86,3
Vitorino	96.960	80.000	40.280	58.015	53.883	47.452	-40,2	-32,6	17,8
Núcleo Pato Branco	1.285.770	1.223.946	820.882	1.197.644	1.116.379	1.018.865	-6,9	-8,8	24,1

Fonte: DERAL/SEAB (2023); Autoria própria (2023).

Apesar dos resultados discrepantes entre a produção oficial e estimada na safra 2021/2022 para os municípios de Saudade do Iguaçu e Itapejara d'Oeste, os quais indicam oficialmente uma quebra de safra superior a 75% quando comparados com a safra anterior, ressalta-se que a mesma situação não condiz com a produção total do Núcleo Pato Branco, onde a redução foi de apenas 33% em relação à safra 2020/2021. Ainda nesse sentido, cabe destacar que nos municípios vizinhos aos supracitados, tais como Sulina e Pato Branco, respectivamente, que possuem características físicas de solos e de topografia similares, não acompanham a mesma redução da produção agrícola da soja.

Os resultados acima descritos na Tabela 12, se diferem parcialmente do apresentado por Richetti et al. (2018), que buscaram estimar a produção da soja para todo o estado do Paraná, na safra 2013/2014, utilizando-se de preditores por índice EVI do sensor MODIS, com resolução espacial de 250 metros, onde houve uma superestimativa quando comparada com dados oficiais. Essa diferença pode ser justificada pelo fato que nesse trabalho utilizou-se dados de terreno, precipitação e três tipos de índices de vegetação como preditores, sendo

os índices gerados a partir de imagens do sensor MSI do SENTINEL-2, com 10 metros de resolução espacial, onde foi possível estimar a produção a cada 100 m², diferentemente do MODIS, que são estimados a cada 6,25 hectares.

5. CONCLUSÕES

Os resultados obtidos a partir dos modelos de estimativas de produção demonstraram similaridades aos encontrados atualmente na literatura.

A utilização de índices de vegetação obtidos por imagens de alta resolução espacial, como o MSI do SENTINEL-2, permite estimar a produção da soja de modo igual e superior as do sensor MODIS que possuem alta temporalidade quando associadas com algoritmos de Aprendizado Máquina e quando consideram a variabilidade espacial do talhão na variável resposta da regressão.

O índice NDVI proporcionou maior acurácia na determinação das áreas de cultivo da soja no Núcleo Regional Pato Branco quando comparado ao EVI.

O algoritmo de *AM Random Forest* gera um modelo de estimativa de produção com maior precisão quando comparado com os de Redes Neurais e Máquinas de Vetores de Suporte.

Os índices de vegetação NDMI, EVI e NDVI se configuraram como importantes variáveis preditoras na construção dos modelos.

Os dados do sensor MSI associados com o algoritmo *Random Forest* permitiram estimar a produção regional.

6. CONSIDERAÇÕES FINAIS

Os resultados obtidos nessa pesquisa trazem como possibilidade, a inclusão de dados orbitais processados por técnicas computacionais de *Machine Learning*, na metodologia de coleta de dados para as estatísticas agrícolas municipais. Considerando que as estatísticas oficiais são geradas de forma subjetiva e não quantificam os erros envolvidos, essa proposta se configura como uma alternativa complementar para determinação de áreas e de produção agrícola.

A realização de novos testes por meio do algoritmo *Random Forest* em regiões agrícolas com uniformidade nas suas características físicas tais como declividade, altitude, geologia e solos, associados a dados meteorológicos mais precisos, podem aprimorar as estimativas de safras.

Ressalta-se ainda, para trabalhos futuros, a utilização de imagens orbitais de maior resolução espacial e *Machine Learning* em estudos de monitoramento de safras e em projetos de melhoramento genético de espécies agrícolas de interesse. Essa prática servirá como apoio para tomada de decisões e no gerenciamento de sistemas agrícolas.

REFERENCIAS BIBLIOGRÁFICAS

ANA - AGÊNCIA NACIONAL DE ÁGUAS. Pluviometria. In: **Atlas Geográfico de Recursos Hídricos do Brasil**. 2013. Disponível em: <<http://www2.snirh.gov.br/atlasrh2013/>>

ALABI, T. R.; ABEBE, A. T.; CHIGEZA, G.; FOWOBAJE, K. R.; Estimation of soybean grain yield from multispectral high-resolution UAV data with machine learning models in West Africa. **Remote Sensing Applications: Society and Environment**, v. 27, p.1-17. 2022.

ARAÚJO, L. E. O. C. **Análise de cobertura vegetal e de biomassa em áreas de contato Floresta/savana a partir de dados TM/Landsat e JERS-1**. Dissertação - (Mestrado em Sensoriamento Remoto). Instituto Nacional de Pesquisas Espaciais. São José dos Campos. 1999.

BALBINOT JUNIOR, A. A.; NEPOMUCENO, A. L.; HIRAKURI, M. H.; FRANCHINI, J. C.; DEBIASI, H.; COELHO, A. E. Análise da produção, área cultivada, produtividade e estabilidade produtiva da soja nos estados brasileiros em 25 safras (1996/1997 a 2020/2021). **Boletim de Pesquisa e Desenvolvimento 27**. Londrina: Embrapa Soja, 2022.

BARI, A.; CHAOUCHI, M.; JUNG, T. **Análise preditiva para leigos**. Tradução de Wendy Campos. 2ª Edição. Rio de Janeiro: Alta Books, 2019.

BECK, P. S. A.; ATZBERGER, C.; HOGDA, K. A.; JOHANSEN, B.; SKIDMORE, A. K. Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. **Remote Sensing of Environment**, v. 100, n. 3, p. 321-334, 2006.

BELGIU, M.; DRAGUT, L. Random forest in remote sensing: A review of applications and future directions. **Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24-31, 2016.

BERNARDES, M. S. Fotossíntese no dossel de plantas cultivadas. In: CASTRO. P.R.C.; FERREIRA. S.O.; YAMADA. T. (org.). **Ecofisiologia da produção agrícola**. Piracicaba: Associação Brasileira de Potassa e do Fosfato. 1987. p. 13-48.

BESALATPOUR, A. A.; AYOUBI, S.; HAJABBASI, M. A.; JAZI, A.; JASI Y. A.; GHARIPOUR, A. Feature selection using parallel genetic algorithm for the prediction of geometric mean diameter of soil aggregates by machine learning methods. **Arid Land Research and Management**, v. 28, p. 383-394, 2014.

BOCARD, D.; GILLET, F.; LEGENDRE, P. Numerical Ecology with R. USA: SPRINGER, 2011.

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

CERRI, R.; CARVALHO, A. C. P. L. F. Aprendizado de máquina: breve introdução e aplicações. **Cadernos de Ciência & Tecnologia**, v. 34, n. 3, p. 297-313, 2017.

CHERGUI, N. Durum wheat yield forecasting using machine learning. **Artificial Intelligence in Agriculture**, v. 6, p.156-166, 2022.

CHLINGARYAN, A.; SUKKARIEH, S.; WHELAN, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. **Computers and Electronics in Agriculture**, v. 151, n. 8, p. 61-69, 2018.

COMPANHIA NACIONAL DE ABASTECIMENTO – CONAB. **Acompanhamento da Safra Brasileira de Grãos**. Brasília, DF, v. 7 - Safra 2019/20 - Décimo primeiro levantamento, Brasília, p. 1-62, setembro 2020. Disponível em: <https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos/item/download/45062_b938afb12d52edd613d312dc4390da24>. Acesso: 23 jan. 2023.

COMPANHIA NACIONAL DE ABASTECIMENTO - CONAB. **Acompanhamento da Safra Brasileira de Grãos**. Brasília, DF, v. 8, safra 2020/21, n. 10 décimo primeiro levantamento, setembro 2021. Disponível em: <https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos/item/download/45076_d4d0046db86c6b909212c302cdee3955>. Acesso: 23 jan. 2023.

COMPANHIA NACIONAL DE ABASTECIMENTO - CONAB. **Acompanhamento da Safra Brasileira de Grãos**. Brasília, DF, v. 9, safra 2021/22, n. 11 décimo primeiro levantamento, agosto 2022. Disponível em: <https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos/item/download/44171_1d9f893d78f593b07d41887104acc43f>. Acesso: 23 jan. 2023.

DADO, W. T.; DEINES, J. M.; PATEL, R.; LIANG, S. Z.; LOBELL, D. B. High-Resolution Soybean Yield Mapping Across the US Midwest Using Subfield Harvester Data. **Remote Sensing**, v. 12, p. 1-22, 2020.

DERAL/SEAB. **Boletins conjunturais**. [Curitiba: DERAL/SEAB]. 2020. Disponível em: <https://www.agricultura.pr.gov.br/sites/default/arquivos_restritos/files/documento/2020-12/Prognostico%20Soja%20-%202020_21.pdf>. Acesso: 10 dez. 2020.

DERAL/SEAB. **Metodologia de coleta de dados da Previsão Subjetiva de Safras (PSS)**. [Curitiba: DERAL/SEAB]. 2019 Disponível em: <https://www.agricultura.pr.gov.br/sites/default/arquivos_restritos/files/documento/2019-09/metodologia_pss.pdf>. Acesso: 15 jun. 2021.

DERAL/SEAB. **Estimativa de safra**. [Curitiba: DERAL/SEAB]. 2023. Disponível em: <<https://www.agricultura.pr.gov.br/deral/safras>>. Acesso: 24 jan. 2023.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. **História da soja**. 2020. Disponível em: <<https://www.embrapa.br/web/portal/soja/cultivos/soja1/historia>>. Acesso em: 23 jan. 2023.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. **Levantamento de reconhecimento dos solos do estado do Paraná**. Rio de Janeiro. 2006. Escala 1:250.000.

EUROPEAN SPACIAL AGENCY. **SENTINEL-2**. 2020. Disponível em <<https://sentinel.esa.int/web/sentinel/missions/sentinel-2/overview>>. Acesso em: 15 abr. 2021.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. **Inteligência Artificial: uma abordagem aprendizagem de máquina**. Rio de Janeiro: LTC. 2011.

FAO – Food and Agriculture Organization of United Nations. **Commodities and Development Report 2017**. New York and Geneva. 2017. Disponível em: <<http://www.fao.org/3/I7937EN/i7937en.pdf>>. Acesso em: 26 jul. 2021.

FEHR, W. R.; CAVINESS, C. E. **Stages of soybean development**. Ames: Iowa State University of Science and Technology. 1977.

FEYISA, G. L.; PALAO, L. K.; NELSON, A.; GUMMA, M. K.; PALIWALA, A.; WIN, K. T.; NGE, K. H.; JOHNSON, D. E. Characterizing and mapping cropping patterns in a complex agro-ecosystem: An iterative participatory mapping procedure using machine learning algorithms and MODIS vegetation indices. **Computers and Electronics in Agriculture**, v. 175, p. 1-11. 2020.

GALLEGO, J.; CARFAGNA, E.; BARUTH, B. Accuracy, objectivity and efficiency of remote sensing for agricultural statistics. BENEDETTI, R.; BEE, M.; ESPA, G.; PIERSIMONI, F. (org.). **Agricultural Survey Methods**. West Sussex: John Wiley & Sons Ltd. 2010. p.193-211.

GANDHI, N.; ARMSTRONG L.; PETKAR, O.; TRIPATHY, A. K. Rice crop yield prediction in India using support vector machines. *In*: 13th International Joint Conference on Computer Science and Software Engineering (JCSSE). 2016. **Anais** [...]. Khon Kaen: IEEE. 2016. p.1-5.

GAO, Bo-cai. NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. **Remote Sensing of Environment**, v. 58, n. 3, p. 257-266, 1996.

GAUSMAN, H. W. Leaf reflectance of near-infrared. **Photogrammetric Engineering and Remote Sensing**, v. 51, n. 11, p. 1725-1734, 1985.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining: conceitos, técnicas, algoritmos, orientações e aplicações**. 2ª Ed. Rio de Janeiro: Elsevier, 2015.

HABIBI, L. N.; WATANABE, T.; MATSUI, T.; TANAKA, T. Machine learning techniques to predict soybean plant density using UAV and satellite-based remote sensing. **Remote sensing**, v. 13, n. 13, p. 1-20, 2021.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. Massachusetts: Library of Congress Cataloging-in-Publication Data. 2001.

HENDAWY, S. E.; HASSAN, W. M.; SUHAIBANI, N. A.; SCHMIDHALTER, U.

Spectral assessment of drought tolerance indices and grain yield in advanced spring wheat lines grown under full and limited water irrigation. **Agricultural Water Management**, v. 182, p. 1-12, 2017.

HEREHER, M. E. The status of Egypt's agricultural lands using MODIS Aqua data. **The Egyptian Journal of Remote Sensing and Space Science**, v. 16, n. 1, p. 83-89, 2013.

HINNAH, F. D.; SENTELHAS, P. C.; GLEASON, M. L.; DIXON, P. M.; ZHANG, X. Assessing Biogeography of Coffee Rust Risk in Brazil as Affected by the El Niño Southern Oscillation. **Plant disease**, v. 104, n. 4, p. 1013-1018, 2020.

HUANG, X.; LIU, J.; ZHU, W.; ATZBERGER, C.; LIU, Q. The optimal threshold and vegetation index time series for retrieving crop phenology based on a modified dynamic threshold method. **Remote Sens**, v. 11, n. 23, p. 1-20, 2019.

HUETE, A.; JUSTICE, C.; LEEUWEN, W. **Modis Vegetation Index (MOD 13): algorithm theoretical basis**. NASA. 1999. Disponível em: <https://modis.gsfc.nasa.gov/data/atbd/atbd_mod13.pdf>. Acesso em: 15 jan. 2021.

HUETE, A.; DIDAN, K.; MIURA, T.; RODRIGUEZ, E. P.; GAO, X.; FERREIRA, L. G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, v. 83, n.1-2, p.195-213, 2002.

INSTITUTO NACIONAL DE METEOROLOGIA - INMET. **Tabela de dados das estações**. Disponível em: < <https://tempo.inmet.gov.br/TabelaEstacoes/A001>>. Acesso em: 14 dez. 2022.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. São Carlos, SP: Rafael Izbicki, 2020.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to Statistical Learning** – with application in R. London: Springer. 2013.

JEONG, J. H.; RESOP, J.; MUELLER, N. D.; FLEISHER, D. H.; YUN, K.; BUTLER, E. E.; TIMLIN, D. J.; SHIM, K.; GERBER, J. S.; REDDY, V. R.; KIM, S. Random forests for global and regional crop yield predictions. **PLoS ONE**, v. 11, n. 6, p. 1-15, 2016.

JENSEN, J. R. **Sensoriamento Remoto do Ambiente: uma perspectiva em Recursos Terrestres**. Tradução de José Carlos Neves Epiphânio. São José dos Campos: Parênteses. 2009. 598p.

JOHANN, J. A.; ROCHA, J. V.; DUFT, D. G.; LAMPARELLI, R. A. C. Estimativa de áreas com culturas de verão no Paraná. por meio de imagens multitemporais EVI/Modis. **Pesquisa Agropecuária Brasileira**, v. 47, n. 9, p.1295-1306, 2012.

JOHNSON, D. M. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. **Remote Sensing of Environment**, v. 141, p.116-128, 2014.

JOHN DEERE. **Monitor GreenStar™3 2630**. 2022. Available at: <<https://www.deere.com.br/pt/agricultura-de-precis%C3%A3o/receptores-monitores-e-rtk/monitor-greenstar-gs2-2630/>>. Access in: August. 01. 2022.

JUSTICE, C. O.; VERMOTE, E.; TOWNSHEND, J. R. G.; DEFRIES, R. S. The moderate resolution imaging spectroradiometer (Modis): land remote sensing for global change research. **IEEE Transactions on Geoscience and Remote Sensing**, v. 36, n. 4, p. 1228-1249, 1998.

KHANAL, S.; FULTON, J.; KLOPFENSTEIN, A.; DOURIDAS, N.; SCOTT, A. S. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. **Computers and Electronics in Agriculture** v. 153, p. 213-225, 2018.

KARTHIKEYAN, L.; CHAWLA, I.; MISHRA, A. K. A review of remote sensing applications in agriculture for food security: Crop growth and yield. irrigation. and crop losses. **Journal of Hydrology**, v. 586, p. 1-22, 2020.

LIAKOS, K. G.; BUSATO, P.; MOSHOU, D.; PEARSON, S.; BOCHTIS, D. Machine learning in agriculture: a review. **Sensors**, v. 18, p. 1-29, 2018.

LIZUMI, T., LUO, J. J.; CHALLINOR, A.; SAKURAI, G.; YOKOZAWA, M.; SAKUMA, H.; BROWN, M. E.; YAMAGATA, T. Impacts of El Niño Southern Oscillation on the global yields of major crops. **Nat Commun**, v. 5, p. 1-7, 2014.

LUCIANO, A. C. S.; PICOLI, M. C. A.; DUFT, D. G.; ROCHA, J. V.; LEAL, M. R. L. V.; LE MAIRE, G. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. **Computers and Electronics in Agriculture**, v. 184, p. 1-10, 2021.

LI, H.; WANG, J.; WANG, Q.; TIAN, C.; QIAN, X.; LENG, X. Magnetic properties as a proxy for predicting fine-particle-bound heavy metals in a support vector machine approach. **Environmental, Science & Technology**, v. 51, n. 12, p. 6927-6935, 2017.

LI, L.; WANG, B.; FENG, P.; LIU, L. D.; HE, Q.; ZHANG, Y.; WANG, Y.; LI, S.; LU, X.; YUE, C.; Yi Li, Y.; HE, J.; FENG, H.; YANG, G.; YU, Q. Developing machine learning models with multi-source environmental data to predict wheat yield in China. **Computers and Electronics in Agriculture**, v. 194, p. 1-12, 2022.

LIU, J.; HUFFMAN, T.; QIAN, B.; SHANG, J.; LI, Q.; DONG, T.; DAVIDSON, A.; JING, Q. Crop Yield Estimation in the Canadian Prairies Using Terra / MODIS-Derived Crop Metrics. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 13, p.1-13, 2020.

LUGER, G. F. **Inteligência Artificial**. 6ª ed. São Paulo: Pearson Education do Brasil. 2013.

MARTINS, G. D.; GALO, M. L. B. T. Caracterização espectral da cana-de-açúcar infectada por nematoides e *migdolus fryanus* por espectrorradiometria de campo. **Boletim de Ciências Geodésicas**, v. 21, n. 4, p. 783-796, 2015.

MIMIĆ, G.; ŽIVALJEVIĆ, B.; BLAGOJEVIĆ, D.; PEJAKE, B.; BRDAR, S.

Quantifying the Effects of Drought Using the Crop Moisture Stress as an Indicator of Maize and Sunflower Yield Reduction in Serbia. **Atmosphere**, v. 13, n. 11, p. 1-12, 2022.

MINEROPAR. **Cartas geológicas do Paraná**. Curitiba. 2006. Escala 1:250.000.

MOLIN, J. P.; AMARAL, L. R.; COLAÇO, A. F. **Agricultura de precisão**. São Paulo: Oficina de Textos. 2015.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. **Sistemas Inteligentes Fundamentos e Aplicações**. Barueri: Manole Ltda. 2003.

MOREIRA, Maurício Alves. **Fundamentos do Sensoriamento Remoto e metodologias de aplicação**. 3ª ed. atual. ampl. Viçosa: Ed. UFV. 2005.

NARDY, A. J. R.; MACHADO, F. B. OLIVEIRA, M. A. F. As rochas vulcânicas mesozóicas ácidas da Bacia do Paraná litoestratigrafia e considerações geoquímico-estratigráficas. **Geociências**, São Paulo, v. 38, n. 1, p. 178-195, 2008.

NASA LP DAAC. **MODIS overview**. Disponível em: <https://lpdaac.usgs.gov/data/get-started-data/collection-overview/missions/modis-overview/>. Acesso em: 18 maio 2021.

NEUMAIER, N.; FARIAS, J. R. B.; NEPOMUCENO, A. L.; HENNING, L. M. M.; FOLONI, J. S. S.; MORAES, L. A. C.; GONÇALVES, S. L. Ecofisiologia da soja. *In*: SEIXAS, C. D. S.; NEUMAIER, N.; JUNIOR, A. A. B.; KRZYZANOWSKI, F. C.; LEITE, R. M. V. B. C. (org.). **Tecnologias de produção de soja**. Londrina: Embrapa Soja. 2020. p. 33-54.

OLIVEIRA JUNIOR, A.; CASTRO, C.; OLIVEIRA, F. A.; KLEPKER, D. Fertilidade do solo e avaliação do estado nutricional da soja. *In*: SEIXAS, C. D. S.; NEUMAIER, N.; JUNIOR, A. A. B.; KRZYZANOWSKI, F. C.; LEITE, R. M. V. B. C. (org.). **Tecnologias de produção de soja**. Londrina: Embrapa Soja. 2020. p. 133-184.

OLSON, D. L.; DELEN, D. **Advanced Data Mining Techniques**. Berlin: Springer. 2008.

PANTAZI, X.; MOSHOU, D.; ALEXANDRIDIS, T.; WHETTON, R. L.; MOUAZEN, A. M. Wheat yield prediction using machine learning and advanced sensing techniques. **Computers and Electronics in Agriculture**, v. 121, p. 57-65, 2016.

PONZONI, F. J.; SHIMABUKURO, Y. E.; KUPLICH, T. M. **Sensoriamento Remoto da vegetação**. 2ª ed. atual. Ampl. São Paulo: Oficina de Textos. 2012.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing. 2022. Vienna, Austria. Available in: <http://www.r-project.org/index.html>. Access in: May 19, 2023.

RICHETTI, J.; JUDGE, J.; BOOTE, K. J.; JOHANN, J. A.; OPAZO, A. U.; BECKER, W. R.; PALUDO, A.; SILVA, L. C. A. Using phenology-based enhanced

vegetation index and machine learning for soybean yield estimation in Paraná State. Brazil. **Journal of Applied Remote Sensing**, v. 12, n. 2, p.1-16. 2018.

RODRIGUEZ-GALIANO, V. F.; SANCHEZ-CASTILLO, M.; DASH, J.; ATKINSON, P. M.; OJEDA-ZUJAR, J. Modelling interannual variation in the spring and autumn land surface phenology of the European forest. **Biogeosciences**, v.13, p.3305-3317. 2016.

SALAZAR-ROJAS, T.; CEJUDO-RUIZ, F. R.; CALVO-BRENES, G. Comparison between machine linear regression (MLR) and support vector machine (SVM) as model generators for heavy metal assessment captured in biomonitors and road dust. **Environmental Pollution**, v. 314, p. 1-11, 2022.

SEIXAS, C. D. S.; SOARES, R. M.; GODOY, C. V.; MEYER, M. C.; COSTAMILAN, L. M.; DIAS, W. P.; ALMEIDA, A. M. R. Manejo de doenças. *In*: SEIXAS, C. D. S.; NEUMAIER, N.; JUNIOR, A. A. B.; KRZYZANOWSKI, F. C.; LEITE, R. M. V. B. C. (org.). **Tecnologias de produção de soja**. Londrina: Embrapa Soja. 2020. p. 227-263.

SENTELHAS, P. C.; BATTISTI, R.; CÂMARA, G. M. S.; FARIAS, J. R. B.; HAMPF, A. C.; NENDEL, C. The soybean yield gap in Brazil - magnitude, causes and possible solutions for sustainable production. **Journal of Agricultural Science**, v. 1, p. 1-18, 2015.

SHAMMI, S. A.; MENG, Q. Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling. **Ecological Indicators**, v. 121, p. 1-12, 2021.

SISTEMA METEOROLÓGICO DO PARANA – SIMEPAR. **Dados das estações**. Curitiba: SIMEPAR, 2022.

SINGH, P.; PANDEY, P. C.; PETROPOULOS, G. P.; PAVLIDES, A.; SRIVASTAVA, P. K.; KOUTSIAS, N.; DENG., K. A. K.; BOA, Y. Hyperspectral remote sensing in precision agriculture: present status, challenges, and future trends. *In*: PANDEY, P. C.; SRIVASTAVA, P. K.; BALZTER, H.; BHATTACHARYA, B.; PETROPOULOS, G. P. (org.). **Earth Observation - Hyperspectral Remote Sensing**. Elsevier. 2020. p. 121-146.

STEPANOV, A.; DUBROVIN, K.; SOROKIN, A.; ASEEVA, T. Predicting soybean yield at the regional scale using remote sensing and climatic data. **Remote Sensing**, v. 12, p. 1-23, 2020.

TALLOOR, A. K.; MANHAS, D. S.; KOTHIARY, G. C. Retrieval of land surface temperature, normalized difference moisture index, normalized difference water index of the Ravi basin using Landsat data. **Applied Computing and Geosciences**, v. 9, p. 1-11, 2021.

VELOSO, A.; MERMOZ, S.; BOUVET, A.; TOAN, T. L.; PLANELLS, M.; DEJOUX, J. F.; CESCHIA, E. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. **Remote Sensing of Environment**, v. 199, p.1-12, 2017.

WARDLOW, B. D.; EGBERT, S. L. A comparison of MODIS 250-m EVI and NDVI

data for crop mapping: a case study for southwest Kansas. **International Journal of Remote Sensing**, v. 31, n. 3, p. 805-830, 2010.

WEISS, M.; JACOB, F.; DUVEILLER, G. Remote sensing for agricultural applications: A meta-review. **Remote Sensing of Environment**, v. 236, p.1-69, 2020.

XIN, Q.; GONG, P.; YU, C.; YU, L.; BROICH, M.; SUYKER, A. E.; MYNENI, R. B. A production efficiency, model-based method for satellite estimates of corn and soybean yields in the midwestern US. **Remote Sensing**. v. 5, n. 11, 2013.

ZANOTTA, D. C.; FERREIRA, M. P.; ZORTEA, M. **Processamento de imagens de satélite**. São Paulo: Oficina de Textos, 2019.

ZHONG, L.; GONG, P.; BIGING, G. S. Efficient corn and soybean mapping with temporal extendability: a multi-year experiment using Landsat imagery. **Remote Sensing of Environment**. v. 140, p. 1-13, 2014.

APÊNDICE A – Conjuntos de dados SEM aplicação de filtros tais como outliers, extremos valores e VIFs do sensor MSI

SENSOR MSI – SENTINEL 2						
	Safra 2019/2020	Tipo de variável	Safra 2019/2020	Tipo de variável	Safra 2019/2020	Tipo de variável
1	PROD_2019	Numérica	PROD_2020	Numérica	PROD_2022	Numérica
2	EVI_10102019	Numérica	EVI_24092020	Numérica	EVI_09092021	Numérica
3	EVI_25102019	Numérica	EVI_04102020	Numérica	EVI_24102021	Numérica
4	EVI_19112019	Numérica	EVI_19102020	Numérica	EVI_03112021	Numérica
5	EVI_29112019	Numérica	EVI_08112020	Numérica	EVI_13112021	Numérica
6	EVI_14122019	Numérica	EVI_28112020	Numérica	EVI_03122021	Numérica
7	EVI_29122029	Numérica	EVI_12012021	Numérica	EVI_28122021	Numérica
8	EVI_18012020	Numérica	EVI_06022021	Numérica	EVI_12012022	Numérica
9	EVI_12022020	Numérica	EVI_21022021	Numérica	EVI_22012022	Numérica
10	EVI_22022020	Numérica	EVI_18032021	Numérica	EVI_11022022	Numérica
11	EVI_13032020	Numérica	NDVI_24092020	Numérica	EVI_08032022	Numérica
12	NDVI_10102019	Numérica	NDVI_04102020	Numérica	NDVI_09092021	Numérica
13	NDVI_25102019	Numérica	NDVI_19102020	Numérica	NDVI_24102021	Numérica
14	NDVI_19112019	Numérica	NDVI_08112020	Numérica	NDVI_03112021	Numérica
15	NDVI_29112019	Numérica	NDVI_28112020	Numérica	NDVI_13112021	Numérica
16	NDVI_14122019	Numérica	NDVI_12012021	Numérica	NDVI_03122021	Numérica
17	NDVI_29122029	Numérica	NDVI_06022021	Numérica	NDVI_28122021	Numérica
18	NDVI_18012020	Numérica	NDVI_21022021	Numérica	NDVI_12012022	Numérica
19	NDVI_12022020	Numérica	NDVI_18032021	Numérica	NDVI_22012022	Numérica
20	NDVI_22022020	Numérica	NDMI_24092020	Numérica	NDVI_11022022	Numérica
21	NDVI_13032020	Numérica	NDMI_04102020	Numérica	NDVI_08032022	Numérica
22	NDMI_10102019	Numérica	NDMI_19102020	Numérica	NDMI_09092021	Numérica
23	NDMI_25102019	Numérica	NDMI_08112020	Numérica	NDMI_24102021	Numérica
24	NDMI_19112019	Numérica	NDMI_28112020	Numérica	NDMI_03112021	Numérica
25	NDMI_29112019	Numérica	NDMI_12012021	Numérica	NDMI_13112021	Numérica
26	NDMI_14122019	Numérica	NDMI_06022021	Numérica	NDMI_03122021	Numérica
27	NDMI_29122029	Numérica	NDMI_21022021	Numérica	NDMI_28122021	Numérica
28	NDMI_18012020	Numérica	NDMI_18032021	Numérica	NDMI_12012022	Numérica
29	NDMI_12022020	Numérica	PREC_24092020	Numérica	NDMI_22012022	Numérica
30	NDMI_22022020	Numérica	PREC_04102020	Numérica	NDMI_11022022	Numérica
31	NDMI_13032020	Numérica	PREC_19102020	Numérica	NDMI_08032022	Numérica
32	PREC_10102019	Numérica	PREC_08112020	Numérica	PREC_09092021	Numérica
33	PREC_25102019	Numérica	PREC_28112020	Numérica	PREC_24102021	Numérica
34	PREC_19112019	Numérica	PREC_12012021	Numérica	PREC_03112021	Numérica
35	PREC_29112019	Numérica	PREC_06022021	Numérica	PREC_13112021	Numérica
36	PREC_14122019	Numérica	PREC_21022021	Numérica	PREC_03122021	Numérica
37	PREC_29122029	Numérica	PREC_18032021	Numérica	PREC_28122021	Numérica
38	PREC_18012020	Numérica	ALTITUDE	Numérica	PREC_12012022	Numérica
39	PREC_12022020	Numérica	DECLIVIDADE	Numérica	PREC_22012022	Numérica
40	PREC_22022020	Numérica	GEOLOGIA	Catagórica	PREC_11022022	Numérica
41	PREC_13032020	Numérica	SOLOS	Catagórica	PREC_08032022	Numérica
42	ALTITUDE	Numérica	--	--	ALTITUDE	Numérica
43	DECLIVIDADE	Numérica	--	--	DECLIVIDADE	Numérica
44	GEOLOGIA	Catagórica	--	--	GEOLOGIA	Catagórica
45	SOLOS	Catagórica	--	--	SOLOS	Catagórica

APÊNDICE B – Conjunto de dados SEM aplicação de filtros tais como outliers, extremos valores e VIFs do sensor MODIS

SENSOR MODIS – ACQUA E TERRA

	Safra 2019/2020	Tipo de variável	Safra 2019/2020	Tipo de variável	Safra 2019/2020	Tipo de variável
1	PROD_2019	Numérica	PROD_2020	Numérica	PROD_2022	Numérica
2	EVI_22092019	Numérica	EVI_13092020	Numérica	EVI_06092021	Numérica
3	EVI_08102019	Numérica	EVI_21092020	Numérica	EVI_14092021	Numérica
4	EVI_16102019	Numérica	EVI_29092020	Numérica	EVI_22092021	Numérica
5	EVI_24102019	Numérica	EVI_07102020	Numérica	EVI_30092021	Numérica
6	EVI_01112019	Numérica	EVI_23102020	Numérica	EVI_08102021	Numérica
7	EVI_09112019	Numérica	EVI_31102020	Numérica	EVI_16102021	Numérica
8	EVI_17112019	Numérica	EVI_08112020	Numérica	EVI_01112021	Numérica
9	EVI_25112019	Numérica	EVI_16112020	Numérica	EVI_17112021	Numérica
10	EVI_03122019	Numérica	EVI_24112020	Numérica	EVI_03122021	Numérica
11	EVI_11122019	Numérica	EVI_02122020	Numérica	EVI_11122021	Numérica
12	EVI_19122019	Numérica	EVI_10122020	Numérica	EVI_19122021	Numérica
13	EVI_27122019	Numérica	EVI_18122020	Numérica	EVI_27122021	Numérica
14	EVI_01012020	Numérica	EVI_26122020	Numérica	EVI_01012022	Numérica
15	EVI_09012020	Numérica	EVI_01012021	Numérica	EVI_17012022	Numérica
16	EVI_17012020	Numérica	EVI_09012021	Numérica	EVI_25012022	Numérica
17	EVI_25012020	Numérica	EVI_17012021	Numérica	EVI_02020222	Numérica
18	EVI_02022020	Numérica	EVI_25012021	Numérica	EVI_10020222	Numérica
19	EVI_10022020	Numérica	EVI_10022021	Numérica	EVI_18020222	Numérica
20	EVI_1802020	Numérica	EVI_18022021	Numérica	EVI_26020222	Numérica
21	EVI_28022020	Numérica	EVI_26022021	Numérica	EVI_14032022	Numérica
22	EVI_05032020	Numérica	EVI_06032021	Numérica	EVI_22032022	Numérica
23	EVI_13032020	Numérica	EVI_14062021	Numérica	NDVI_14092021	Numérica
24	EVI_21032020	Numérica	EVI_22032021	Numérica	NDVI_22092021	Numérica
25	NDVI_14092019	Numérica	NDVI_13092020	Numérica	NDVI_30092021	Numérica
26	NDVI_22092019	Numérica	NDVI_21092020	Numérica	NDVI_08102021	Numérica
27	NDVI_30092019	Numérica	NDVI_29092020	Numérica	NDVI_16102021	Numérica
28	NDVI_08102019	Numérica	NDVI_07102020	Numérica	NDVI_24102021	Numérica
29	NDVI_16102019	Numérica	NDVI_23102020	Numérica	NDVI_01112021	Numérica
30	NDVI_24102019	Numérica	NDVI_31102020	Numérica	NDVI_17112021	Numérica
31	NDVI_01112019	Numérica	NDVI_08112020	Numérica	NDVI_03122021	Numérica
32	NDVI_09112019	Numérica	NDVI_16112020	Numérica	NDVI_11122021	Numérica
33	NDVI_17112019	Numérica	NDVI_24112020	Numérica	NDVI_19122021	Numérica
34	NDVI_25112019	Numérica	NDVI_02122020	Numérica	NDVI_27122021	Numérica
35	NDVI_03122019	Numérica	NDVI_10122020	Numérica	NDVI_01012022	Numérica
36	NDVI_11122019	Numérica	NDVI_18122020	Numérica	NDVI_09012022	Numérica
37	NDVI_19122019	Numérica	NDVI_26122020	Numérica	NDVI_17012022	Numérica
38	NDVI_27122019	Numérica	NDVI_01012021	Numérica	NDVI_25012022	Numérica
39	NDVI_01012020	Numérica	NDVI_09012021	Numérica	NDVI_02020222	Numérica
40	NDVI_09012020	Numérica	NDVI_17012021	Numérica	NDVI_10020222	Numérica
41	NDVI_17012020	Numérica	NDVI_25012021	Numérica	NDVI_18020222	Numérica
42	NDVI_25012020	Numérica	NDVI_10022021	Numérica	NDVI_26020222	Numérica
43	NDVI_02022020	Numérica	NDVI_18022021	Numérica	NDVI_14032022	Numérica
44	NDVI_10022020	Numérica	NDVI_26022021	Numérica	NDVI_22032022	Numérica
45	NDVI_1802020	Numérica	NDVI_06032021	Numérica	NDMI_14092021	Numérica
46	NDVI_28022020	Numérica	NDVI_14062021	Numérica	NDMI_23092021	Numérica

47	NDVI_05032020	Numérica	NDVI_22032021	Numérica	NDMI_30092021	Numérica
48	NDVI_13032020	Numérica	NDMI_13092020	Numérica	NDMI_08102021	Numérica
49	NDVI_21032020	Numérica	NDMI_21092020	Numérica	NDMI_16102021	Numérica
50	NDMI_14092019	Numérica	NDMI_29092020	Numérica	NDMI_24102021	Numérica
51	NDMI_22092019	Numérica	NDMI_07102020	Numérica	NDMI_01112021	Numérica
52	NDMI_30092019	Numérica	NDMI_15102020	Numérica	NDMI_09112021	Numérica
53	NDMI_08102019	Numérica	NDMI_22102020	Numérica	NDMI_17112021	Numérica
54	NDMI_16102019	Numérica	NDMI_31102020	Numérica	NDMI_25112021	Numérica
55	NDMI_24102019	Numérica	NDMI_07112020	Numérica	NDMI_03122021	Numérica
56	NDMI_01112019	Numérica	NDMI_16112020	Numérica	NDMI_11122021	Numérica
57	NDMI_09112019	Numérica	NDMI_24112020	Numérica	NDMI_19122021	Numérica
58	NDMI_17112019	Numérica	NDMI_02122020	Numérica	NDMI_27122021	Numérica
59	NDMI_25112019	Numérica	NDMI_10122020	Numérica	NDMI_01012022	Numérica
60	NDMI_03122019	Numérica	NDMI_18122020	Numérica	NDMI_09012022	Numérica
61	NDMI_11122020	Numérica	NDMI_26122020	Numérica	NDMI_17012022	Numérica
62	NDMI_19122019	Numérica	NDMI_01012021	Numérica	NDMI_25012022	Numérica
63	NDMI_27122019	Numérica	NDMI_09012021	Numérica	NDMI_02020222	Numérica
64	NDMI_01012020	Numérica	NDMI_17012021	Numérica	NDMI_06020222	Numérica
65	NDMI_09012020	Numérica	NDMI_25012021	Numérica	NDMI_10020222	Numérica
66	NDMI_17012020	Numérica	NDMI_02022021	Numérica	NDMI_18020222	Numérica
67	NDMI_25012020	Numérica	NDMI_10022021	Numérica	NDMI_26020222	Numérica
68	NDMI_02022020	Numérica	NDMI_18022021	Numérica	NDMI_14032022	Numérica
69	NDMI_10022020	Numérica	NDMI_26022021	Numérica	NDMI_22032022	Numérica
70	NDMI_18022020	Numérica	NDMI_06032021	Numérica	PREC_09092021	Numérica
71	NDMI_28022020	Numérica	NDMI_14032021	Numérica	PREC_24102021	Numérica
72	NDMI_05032020	Numérica	NDMI_22032021	Numérica	PREC_03112021	Numérica
73	NDMI_13032020	Numérica	PREC_24092020	Numérica	PREC_13112021	Numérica
74	NDMI_21032020	Numérica	PREC_04102020	Numérica	PREC_03122021	Numérica
75	PREC_10102019	Numérica	PREC_19102020	Numérica	PREC_28122021	Numérica
76	PREC_25102019	Numérica	PREC_08112020	Numérica	PREC_12012022	Numérica
77	PREC_19112019	Numérica	PREC_28112020	Numérica	PREC_22012022	Numérica
78	PREC_29112019	Numérica	PREC_12012021	Numérica	PREC_11020222	Numérica
79	PREC_14122019	Numérica	PREC_06022021	Numérica	PREC_08032022	Numérica
80	PREC_29122019	Numérica	PREC_21022021	Numérica	DECLIVIDADE	Numérica
81	PREC_18012019	Numérica	PREC_18032021	Numérica	ALTITUDE	Numérica
82	PREC_12022020	Numérica	DECLIVIDADE	Numérica	SOLOS	Categórica
83	PREC_22022020	Numérica	ALTITUDE	Numérica	GEOLOGIA	Categórica
84	PREC_13032020	Numérica	SOLOS	Categórica	-	-
85	DECLIVIDADE	Numérica	GEOLOGIA	Categórica	-	-
86	ALTITUDE	Numérica	-	-	-	-
87	SOLOS	Categórica	-	-	-	-
88	GEOLOGIA	Categórica	-	-	-	-

APÊNDICE C – Variáveis utilizadas COM aplicação de filtros tais como outliers, extremos valores e VIFs para os conjuntos do sensor MSI

SENSOR MSI – SENTINEL 2

	Safra 2019/2020	Tipo de variável	Safra 2020/2021	Tipo de variável	Safra 2021/2022	Tipo de variável
1	PROD_2019	Numérica	PROD_2020	Numérica	PROD_2022	Numérica
2	NDMI_10102019	Numérica	NDMI_19102020	Numérica	EVI_24102021	Numérica
3	NDMI_25102019	Numérica	NDMI_08112020	Numérica	NDMI_03112021	Numérica
4	NDMI_29122019	Numérica	NDMI_12012021	Numérica	EVI_28122021	Numérica
5	NDVI_12022020	Numérica	NDVI_18032021	Numérica	NDVI_28122021	Numérica
6	ALTITUDE	Numérica	EVI_18032021	Numérica	NDMI_28122021	Numérica
7	DECLIVIDADE	Numérica	NDMI_18032021	Numérica	DECLIVIDADE	Numérica
8	SOLOS	Categórica	DECLIVIDADE	Numérica	SOLOS	Categórica
9	-	-	GEOLOGIA	Categórica	-	-

APÊNDICE D – Variáveis utilizadas COM aplicação de filtros tais como outliers, extremos valores e VIFs para os conjuntos do sensor MODIS

SENSOR MODIS – ACQUA E TERRA

	Safra 2019/2020	Tipo de variável	Safra 2020/2021	Tipo de variável	Safra 2021/2022	Tipo de variável
1	PROD_2019	Numérica	PROD_2020	Numérica	PROD_2022	Numérica
2	NDVI_14092019	Numérica	EVI_13092020	Numérica	EVI_06092021	Numérica
3	NDMI_16102019	Numérica	NDMI_13092020	Numérica	NDMI_30092021	Numérica
4	NDMI_24102019	Numérica	NDMI_21092020	Numérica	NDMI_08102021	Numérica
5	NDMI_01112019	Numérica	NDMI_29092020	Numérica	NDMI_16120021	Numérica
6	NDMI_03122019	Numérica	NDM_07102020	Numérica	NDVI_24102021	Numérica
7	EVI_11122019	Numérica	NDMI_22102020	Numérica	NDMI_24102021	Numérica
8	NDVI_11122019	Numérica	NDMI_31102020	Numérica	NDMI_01112021	Numérica
9	EVI_19122019	Numérica	NDMI_07112020	Numérica	NDMI_09112021	Numérica
10	NDMI_19122019	Numérica	NDMI_16112020	Numérica	NDMI_17112021	Numérica
11	NDVI_19122019	Numérica	EVI_16112020	Numérica	NDMI_25112021	Numérica
12	EVI_27122019	Numérica	NDMI_24112020	Numérica	NDVI_11122021	Numérica
13	NDVI_27122019	Numérica	EVI_02122020	Numérica	EVI_11122021	Numérica
14	NDMI_27122019	Numérica	NDVI_02122020	Numérica	NDMI_11122021	Numérica
15	NDMI_01012020	Numérica	NDM_02122020	Numérica	NDMI_19122021	Numérica
16	NDMI_09012020	Numérica	EVI_10122020	Numérica	NDVI_19122021	Numérica
17	NDMI_02022020	Numérica	NDMI_10122020	Numérica	EVI_01012022	Numérica
18	NDMI_18022020	Numérica	EVI_18122020	Numérica	NDVI_09012022	Numérica
19	DECLIVIDADE	Numérica	NDMI_18122020	Numérica	NDMI_09012022	Numérica
20	SOLOS	Categórica	NDVI_18122020	Numérica	NDMI_02022022	Numérica
21	GEOLOGIA	Categórica	EVI_26122020	Numérica	NDMI_06022022	Numérica
22	-	-	NDVI_26122020	Numérica	NDMI_10022022	Numérica
23	-	-	NDMI_26122020	Numérica	NDMI_18022022	Numérica
24	-	-	NDVI_01012021	Numérica	NDMI_26022022	Numérica
25	-	-	NDM_01012021	Numérica	NDMI_14032022	Numérica
26	-	-	EVI_09012021	Numérica	NDMI_22032022	Numérica
27	-	-	NDVI_09012021	Numérica	DECLIVIDADE	Categórica
28	-	-	NDMI_09012021	Numérica	SOLOS	Categórica
29	-	-	NDMI_25012021	Numérica	-	-
30	-	-	NDMI_10022021	Numérica	-	-
31	-	-	NDMI_14032021	Numérica	-	-
32	-	-	NDMI_22032021	Numérica	-	-
33	-	-	SOLOS	Categórica	-	-
34	-	-	GEOLOGIA	Categórica	-	-