MAPEAMENTO DA CULTURA DA SOJA POR MEIO DE IMAGENS SENTINEL-2 E MACHINE LEARNING

Danielli Batistella[1], Alcir José Modolo[2], José Ricardo da Rocha Campos[3], Vanderlei Aparecido de Lima[4]

[1] Departamento de Agrimensura - UTFPR, batistella@utfpr.edu.br; [2]Departamento de Agronomia - UTFPR, alcir@utfpr.edu.br; [3]Departamento de Agronomia; jrcampos@utfpr.edu.br and [4]Departamento de Química - UTFPR – valima@utfpr.edu.br

RESUMO

O presente estudo teve como objetivo mapear áreas de cultivo de soja utilizando imagens SENTINEL-2, utilizando o algoritmo de classificação supervisionada *Random Forest*. O estudo foi realizado no núcleo regional de Pato Branco (PR), composto por 15 municípios, localizado no Sudoeste do estado do Paraná. Para determinar as áreas de cultivo de soja, foram criados perfis espectrotemporais RGB a partir da geração de imagens máximas e mínimas de NDVI (*Normalized Difference Vegetation Index*) e EVI (*Enhanced Vegetation Index*). Os resultados das áreas demonstraram similaridade com os dados reais quando comparados com as estatísticas oficiais registradas pelo Departamento de Economia Rural – DERAL.

Palavras-chave — *Soja, SENTINEL-2, Random Forest.*

ABSTRACT

The present study aimed to map soybean cultivation areas using SENTINEL-2 images, using the Random Forest supervised classification algorithm. the study was carried out in the regional center of Pato Branco (PR), composed of 15 municipalities, located in the southwest of the state of Paraná. To determine the soybean cultivation areas, a RGB Spectro temporal profiles were created from the generation of maximum and minimum NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index) images. The results of the areas demonstrated similarity with the real data when compared with the official statistics recorded by the Department of Rural Economy – DERAL.

key words — *Soybean, SENTINEL-2, Random Forest.*

1. INTRODUÇÃO

O Sensoriamento Remoto orbital permite operacionalizar os programas de previsão e monitoramento de safras [8], além de adquirir informações sobre grandes extensões geográficas em tempo hábil e caracterizar a fenologia de culturas ou da vegetação, de forma geral.

A estimativa de áreas agrícolas associada com dados de Sensoriamento Remoto orbital tem sido amplamente utilizada em conjunto com ferramentas de apoio para tomada de decisões. Entretanto, isso requer o processamento de enorme quantidade de dados de diferentes plataformas. Para tais propósitos, métodos de Aprendizado de Máquina ligados a área de Inteligência Artificial, têm sido amplamente utilizados para geração de modelos de classificação e preditivos, em função de sua capacidade de processar um grande número de entradas de dados. Essas novas abordagens farão parte da prática da agricultura de precisão num futuro próximo [4].

Dados de satélite disponíveis gratuitamente, mesmo fornecendo menor resolução espacial e temporal, têm sido utilizados para monitorar a vegetação diversas culturas agrícolas. Apesar de alguns sensores orbitais apresentarem 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 [10].

Dentre os produtos de orbitais de imageamento terrestre gratuitos e que podem ser aplicados na agricultura, está o sensor MSI do satélite SENTINEL-2, com resolução espacial de 10 metros e temporal de 5 dias [6].

Nesse contexto, o presente trabalho tem como finalidade mapear as áreas de produção da cultura da soja a partir da criação de um perfil-espectro temporal de índices de vegetação NDVI e EVI, utilizando-se imagens do sensor MSI do SENTINEL-2 e por meio do algoritmo de Aprendizado de Máquina, *Random Forest*.

2. MATERIAL E MÉTODOS

A pesquisa foi realizada 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 1).

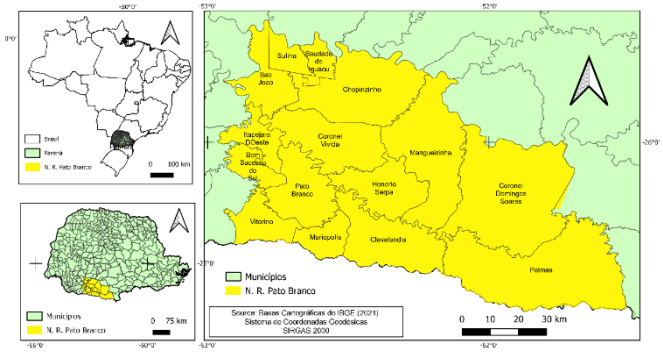

Figura 1 **- Localização do Núcleo Regional Pato Branco - PR**

Os municípios do núcleo regional de Pato Branco, PR 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 [1].

Para o mapeamento das áreas cultivadas com soja, referente aos anos safras 2021/2022, foram obtidas imagens na faixa do visível e infravermelho do satélite 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.

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 a safra de 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.

Foram utilizadas as cenas de 09/09/2021, 24/10/2021, 03/11/2021, 13/11/2021e 13/12/2021para 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 28/12/2021, 12/01/2022, 22/01/2022, 11/02/2022e 08/03/2022, foram utilizadas para geração da imagem de máximo EVI/NDVI (maiores valores de EVI/NDVI 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, 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 2).

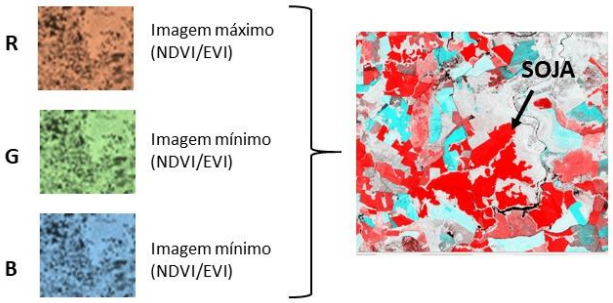

Figura 2 - Composição RGB do espectro-temporal para a cultura da soja

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 Águas e Terras) 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 Worlview 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, sendo criadas 500 árvores (ntrees) e duas variáveis testadas em cada divisão. Os pacotes utilizados foram: "Random Forest", "sf", tidyverse", raster", rgdal", “sp", magrittr", dplyr" e “rgeos".

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 dedados com alta dimensionalidade e multicolinearidade, sendo rápido e insensível ao *overfitting* [2].

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 [3].

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 [11].

3. RESULTADOS

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 ao ano safra 2021/2022, obtidos do DERAL.

A classificação das áreas da cultura da soja (2.000 instâncias), foi avaliada pela matriz confusão e pela precisão, tanto para o conjunto de treinamento, quanto para o conjunto deteste. Os resultados da classificação podem ser visualizados na Tabela 1.

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%		

Tabela 1 - 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, PR.

Os resultados demonstram que na safra 2021/2022, a taxa de erro foi de 0% para NDVI e de 0,5% para o EVI, para o conjunto treinamento, enquanto para o conjunto teste, o erro foi de 0% para NDVI e 1,75% para EVI.

A partir da classificação das áreas-alvos utilizando-se os canais RGB e índices de vegetação EVI e NDVI, foram geradas máscaras nas imagens para identificação da cultura da soja no Núcleo de Pato Branco (Figura 3 – a, b, c e d).

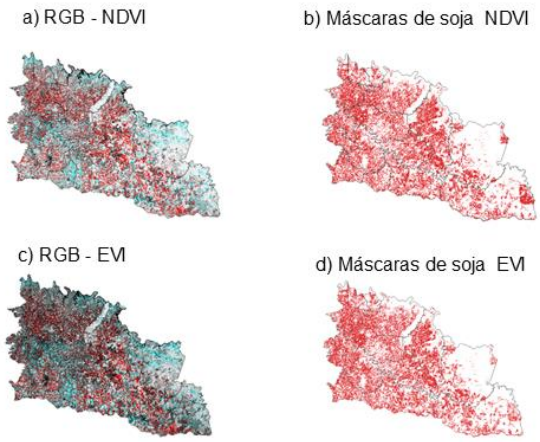

Figura 3 - Distribuição espacial da cultura da soja no Núcleo Pato Branco

As máscaras das áreas estimada para cultura da soja foram quantificadas no aplicativo QGIS e, na sequência, comparada com a área observada a região pelos órgãos oficiais. Os resultados podem ser conferidos na Tabela 2.

IV	**Variáveis**	**2021/2022**
NDVI	Área Deral/SEAB (ha)	340.480
	Área classificação (ha)	328.346
	Diferença de área (ha)	12.134
	Diferença de área (%)	- 3,69%
EVI	Área Deral/SEAB (ha)	340.480
	Área classificação (ha)	327.032
	Diferença de área (ha)	-13.448
	Diferença de área (%)	- 3,94%

Tabela 2- Informações de área oficial (DERAL/SEAB) e obtidas pela classificação para o Núcleo Pato Branco – PR.

Para a cultura da soja, as áreas obtidas pelas máscaras foram menores que às áreas oficiais do DERAL/SEAB, sendo -3,69% para NDVI e -3,94% para EVI. Essa diferença pode estar relacionada ao fator 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.

4. DISCUSSÃO

As máscaras das imagens da cultura da soja, obtidas

a partir da classificação espectro-temporal de imagens do satélite SENTINEL-2/NDVI para a região do Núcleo Pato Branco, apresentaram melhor desempenho em comparação às imagens do SENTINEL-2/EVI, quando suas áreas foram comparadas aos dados oficiais do DERAL. Um resultado semelhante foi observado no estudo de [7], onde o índice MODIS/NDVI mostrou desempenho superior ao MODIS/EVI no mapeamento de áreas agrícolas em Myanmar, alcançando 95% de precisão com o classificador Random Forest. Ao comparar esses índices, os autores destacam que o EVI foi projetado para suprimir sinais de superfícies não vegetadas, o que pode reduzir sua precisão em classes não agrícolas.

5. CONCLUSÃO

Este trabalho contribuiu para o mapeamento das áreas agrícolas de cultivo de soja no Núcleo Regional de Pato Branco, referente à safra 2021/2022. Os resultados obtidos mostraram boa concordância com os dados oficiais do DERAL. As imagens do satélite SENTINEL-2, apesar de não possuírem alta frequência temporal, revelaram-se uma alternativa eficaz para o mapeamento das áreas de cultivo de soja. O algoritmo Random Forest demonstrou eficiência na classificação supervisionada dessas áreas. Conclui-se que a integração de técnicas de Machine Learning com dados orbitais de sensoriamento remoto é uma excelente alternativa para o monitoramento de safras.

6. REFERÊNCIAS

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

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

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

[4] A. CHLINGARYAN, S. SUKKARIEH and B. WHELAN. 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.

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

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

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

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

[9] J. A. JOHANN, J. V. ROCHA, D. G. DUFT and R. A. C. LAMPARELLI. 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.

[10] P. SINGH, P. C. PANDEY, G. P. PETROPOULOS, A. PAVLIDES, P. K. SRIVASTAVA, N. KOUTSIAS, K. A. DENGK and Y. BOA. 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.

[11] L. ZHONG, P. GONG and G. S BIGING. 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.