COMPARAÇÃO DE ÍNDICES DE VEGETAÇÃO DERIVADOS DO SENTINEL-2 PARA ESTIMATIVA DE PRODUTIVIDADE DE SOJA

Ester de Carvalho Pereira [1], Geovani Pereira dos Santos [2], Daniel S. Kaster [2], Gabriela Cristina Salgado[3], Cynthia Junqueira[3], Adilson Chinatto[3], Ana Cláudia dos Santos Luciano[1]

[1]Universidade de São Paulo, Escola de Agricultura Luiz de Queiroz - ESALQ, Piracicaba, SP, Brasil; esterpereira@usp.br, analuciano@usp.br

[2]Universidade Estadual de Londrina, Londrina, PR, Brasil; geovanimoc@gmail.com, dskaster@uel.br

[3]Espectro LTDA, Campinas, SP, Brasil; salgadogc@gmail.com, cyjunqueira@uol.com.br, chinatto@gmail.com

RESUMO

O objetivo deste trabalho foi comparar o uso de diferentes índices de vegetação e suas combinações no desempenho de modelos empíricos para estimar a produtividade de soja a nível municipal na região Centro-Oeste do Brasil. Foram utilizadas imagens do satélite Sentinel-2 e dados históricos do IBGE, com as áreas de soja identificadas pelo Mapbiomas para as safras de 2019/2020 a 2021/2022. Os índices NDVI, EVI, NDRE e NDMI foram calculados, e o modelo de produtividade foi desenvolvido com o algoritmo Random Forest, considerando cinco cenários. O R^2 variou de 0,49 a 0,65 e o RMSE de 336,44 a 407,64 kg/ha. Os cenários com EVI e NDRE apresentaram R^2 de 0,56 e 0,57, respectivamente, enquanto a combinação dos índices resultou em R^2 de 0,65 e RMSE de 336,44 kg/ha, evidenciando o potencial dos índices para a estimativa de produtividade.

Palavras-chave **—** NDRE, EVI, modelos empíricos, séries temporais, produção de soja.

ABSTRACT

The objective of this study was to compare the use of different vegetation indices and their combinations in the performance of empirical models to estimate soybean yield at the municipal level in the Central-West region of Brazil. Sentinel-2 satellite images and historical yield data from IBGE were used, with soybean areas identified by Mapbiomas for the 2019/2020 to 2021/2022 seasons. The NDVI, EVI, NDRE, and NDMI indices were calculated, and the yield model was developed using the Random Forest algorithm, considering five scenarios. The R^2 ranged from 0.49 to 0.65, and the RMSE from 336.44 to 407.64 kg/ha. Scenarios with EVI and NDRE presented R^2 values of 0.56 and 0.57, respectively, while the combination of indices resulted in an R^2 of 0.65 and an RMSE of 336.44 kg/ha, demonstrating the potential of these indices for yield estimation.

Key words **—** *NDRE, EVI, empirical models, time series, soybean yield.*

1. INTRODUÇÃO

A soja, principal grão produzido e exportado pelo Brasil, teve uma produção de 147,68 milhões de toneladas na safra 2023/2024, em uma área de 45,73 milhões de hectares. A região Centro-Oeste representa 45% dessa produção, sendo a principal região produtora de grãos no país. A cultura é essencial para a indústria alimentícia, biodiesel e ração animal e, consiste em uma importante fonte de proteína vegetal. O monitoramento da produção e das áreas cultivadas da soja é fundamental para prever a oferta e demanda da cultura, bem como orientar negociações e garantir a segurança alimentar. No Brasil, o Instituto Brasileiro de Geografia e Estatística (IBGE) e a Companhia Nacional de Abastecimento (Conab) realizam o monitoramento de área e produção por meio de questionários e entrevistas.

Nos últimos anos, o acesso a dados de satélite possibilitou o acompanhamento do uso da terra, áreas desmatadas e cultivos agrícolas, permitindo ainda correlacionar essas informações com a produtividade. Com o avanço dos algoritmos de aprendizado de máquina, surgem modelos de estimativa de produtividade mais robustos para diversas culturas [1], [2].

Índices espectrais como o *Normalized Difference Vegetation Index* (NDVI) e o *Enhanced Vegetation Index* (EVI) são amplamente utilizados no desenvolvimento de modelos de estimativa de produtividade [3]. Outra variável relevante nesses modelos é o histórico de produtividade, que reflete o potencial produtivo das áreas agrícolas [4].

Novos sensores, como o Multispectral Instrument a bordo do Sentinel-2 (S2/MSI), apresentam melhorias na resolução espectral, incluindo as bandas na região do red-

edge e, resolução espacial de 10 metros. Estudos recentes têm mostrado resultados satisfatórios para a estimativa de produtividade de soja a nível local com o uso de dados Sentinel-2, apresentando coeficientes de determinação (R^2) acima de 0,75 [5], [6]. Assim, verifica-se que o uso dos dados do S2/MSI possui grande potencial para o desenvolvimento de modelos empíricos em escalas regionais e globais.

Neste contexto, o objetivo deste trabalho foi comparar o uso de diferentes índices de vegetação e a combinação destes no desempenho de modelos empíricos de estimativa de produtividade de soja a nível municipal. Para isso, foram utilizadas imagens Sentinel-2 e dados históricos de produtividade.

2. MATERIAL E MÉTODOS

A área de estudo está localizada na região Centro-Oeste brasileira (Figura 1). A região é composta pelos estados de Goiás (GO), Mato Grosso (MT), Mato Grosso do Sul (MS) e o Distrito Federal (DF). A soja é plantada na região entre setembro e dezembro, com a colheita iniciando em janeiro e se estendendo até abril [7].

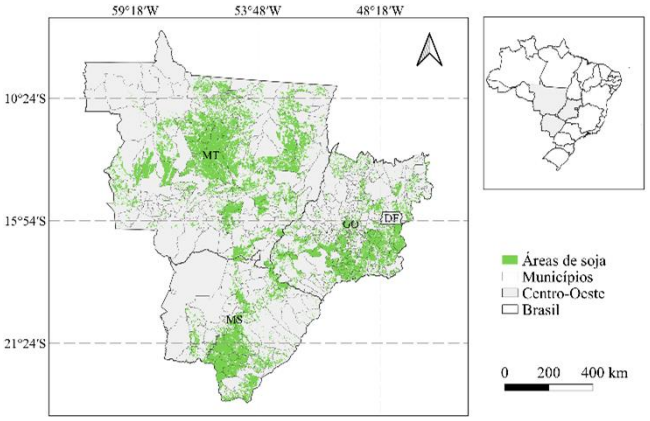


Figura 1 Áreas de soja no Centro-Oeste do Brasil na safra 2022/2023. Fonte: MapBiomas (2024).

Os dados de produtividade utilizados no estudo foram adquiridos através do IBGE (https://sidra.ibge.gov.br/pesquisa/pam/tabelas). Foram obtidos dados de produtividade das safras 2014/2015 a 2021/2022. Os dados de produtividade obtidos do IBGE foram analisados para a remoção de falhas e inconsistências. Para as safras de 2019/2020, 2020/2021 e 2021/2022, houve registros de produtividade de soja para 437 municípios da região. Foram removidos da análise os municípios com pelo menos um registro faltante de produtividade e os municípios com dados de produtividade idênticos para anos consecutivos. Assim, foram selecionados 323 municípios para análise, resultando em um total de 969 registros de produtividade ao longo do período (2019/2020 a 2021/2022), com valores variando de 420 kg/ha a 5000 kg/ha. Com base nos dados de produtividade do IBGE, calculou-se a média histórica dos últimos cinco anos para cada município.

As áreas de soja na região foram identificadas com base nos mapas da Coleção 8 do projeto MapBiomas [8]. Foram utilizados dados do Sentinel-2 (produto Level-2), produto de reflectância de superfície, para cálculo dos índices de vegetação [9]. Foram selecionados os seguintes índices de vegetação: NDVI [10], EVI [11], *Normalized Difference red-edge Index* (NDRE) [12] e Normalized Difference Moisture Index (NDMI) [13] (Tabela 1). Em seguida, foram selecionadas imagens disponíveis entre 1º de outubro e 31 de março das safras 2019/2020, 2020/2021 e 2021/2022. Por fim, foi aplicada uma máscara para remoção de pixels com nuvens e sombras, e os valores medianos dos índices de vegetação foram coletados para cada município nas áreas correspondentes à soja. Para os dados faltantes devido à presença de nuvens e sombras, foi realizada uma interpolação linear. Os dados do período de estudo foram agregados em intervalos de dez dias por meio do cálculo da média.

Índice	Fórmula
NDVI	(NIR - red) / (NIR + red)
EVI	2.5 * (NIR - red) / NIR + (6 * red) - (7.5 * blue) + 1)
NDMI	(NIR – SWIR 1) / (NIR + SWIR 1)
NDRE	(NIR – red-edge 3) / (NIR + red-edge 3)

Tabela 1 Índices de vegetação utilizados no estudo com as respectivas fórmulas. *red* = vermelho, NIR=infravermelho próximo, SWIR = infravermelho médio, *blue* = azul.

Para desenvolver o modelo de produtividade de soja, foi utilizado o algoritmo *Random Forest* [14]. Para todos os modelos, o conjunto de dados foi dividido em 70% para treinamento e 30% para teste, garantindo que ambas as partes mantivessem as proporções de dados para cada estado da região Centro-Oeste e safra. Os hiperparâmetros do modelo *Random Forest* foram otimizados utilizando os dados de teste e validação cruzada com 5 fold, através da ferramenta *GridSearchCV* da biblioteca *scikit-learn* do *Python* [15]. Foram comparados cinco cenários considerando a média de produtividade dos últimos cinco anos e diferentes índices de vegetação como apresentado na Tabela 2. Os modelos foram avaliados utilizando o erro quadrático médio (RMSE) e o R^2. Todas as atividades foram executadas no Google Colab integrado ao ambiente Google Earth Engine utilizando o pacote *geemap*.

Cenário	Variáveis utilizadas
I	NDVI + PROD
II	EVI + PROD
III	NDMI + PROD
IV	NDRE + PROD
V	NDVI + EVI + NDMI + NDRE + PROD

Tabela 2 Diferentes cenários testados para a obtenção dos modelos de produtividade de soja.

3. RESULTADOS E DISCUSSÃO

A Figura 2 apresenta o R^2 e o RMSE dos modelos obtidos para estimar a produtividade de soja em cada cenário. O R^2 variou de 0,49 a 0,65, enquanto o RMSE variou de 336,44 a 407,64 kg/ha. O Cenário I apresentou o menor R^2 (0,49), seguido dos cenários III (R^2 de 0,54), IV (0,56), II (0,57) e V, que apresentou o maior R^2 (0,65). Quanto ao RMSE, o Cenário V apresentou o menor valor, 336,44 kg/ha, seguido do Cenário II (371,84 kg/ha), IV (377,29 kg/ha), III (383,87 kg/ha) e I, com o maior valor de RMSE (407,64 kg/ha).

Ao avaliar os cenários que utilizaram os índices de vegetação separadamente, foi verificado que os cenários que utilizaram os índices EVI (Cenário II) e NDRE (Cenário IV) obtiveram os maiores valores de R^2 e os menores valores de RMSE. O EVI é um índice que utiliza as bandas do *red* e NIR e possui coeficientes para ajustar a influência do solo na resposta do índice, além de utilizar a banda do azul para corrigir os efeitos atmosféricos na banda do vermelho. O EVI apresenta uma linearidade aprimorada na estimativa de parâmetros biofísicos da vegetação, quando comparado ao NDVI, que apresenta uma "saturação" do índice em monitoramentos da vegetação [11]. Outros trabalhos identificaram que o EVI foi uma variável importante para a estimativa de produtividade de soja [6], [16].

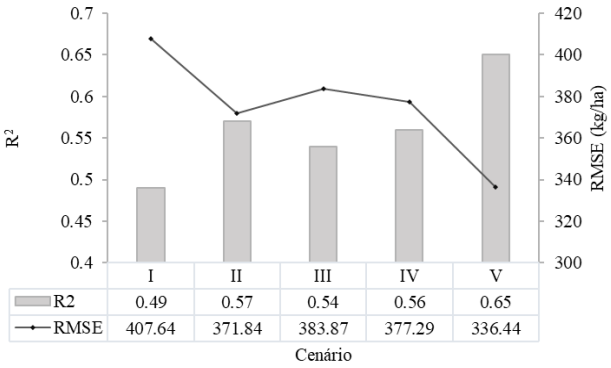


Figura 2 Valores do coeficiente de determinação (R^2) e do erro quadrático médio (RMSE) dos modelos de estimativa de produtividade da soja para diferentes cenários.

Ressalta-se que índices como o NDRE têm demonstrado grande potencial para a estimativa de produtividade, pois consideram tanto a saúde da vegetação e sua capacidade fotossintética, associadas ao *red-edge*, quanto a estrutura celular e a biomassa das plantas, relacionadas ao NIR. O *red-edge* corresponde à região de transição entre o *red* e o NIR, onde ocorre um aumento acentuado na reflectância devido ao espalhamento celular característico da região do NIR [17].

O modelo desenvolvido no Cenário 5 para estimar a produtividade de soja no Centro-Oeste brasileiro, em escala municipal, alcançou um R^2 igual 0,65 e um RMSE de 336,44 kg/ha (Figura 3). Esses resultados são comparáveis aos de outros estudos que estimaram a produtividade de soja no Brasil em nível municipal usando índices espectrais, com RMSE variando entre 320 e 420 kg/ha [2], [18]. Vale destacar que esses estudos utilizaram séries históricas mais longas, superiores a 10 anos, além de dados climáticos adicionais. Pesquisadores utilizaram dados de 19 safras para desenvolver um modelo de estimativa da produtividade de soja no Brasil em nível municipal, combinando variáveis espectrais (MODIS e Landsat) com dados climáticos. Os autores alcançaram um modelo com R^2 de 0,69 e RMSE de 344 kg/ha, resultados semelhantes aos obtidos neste estudo [18].

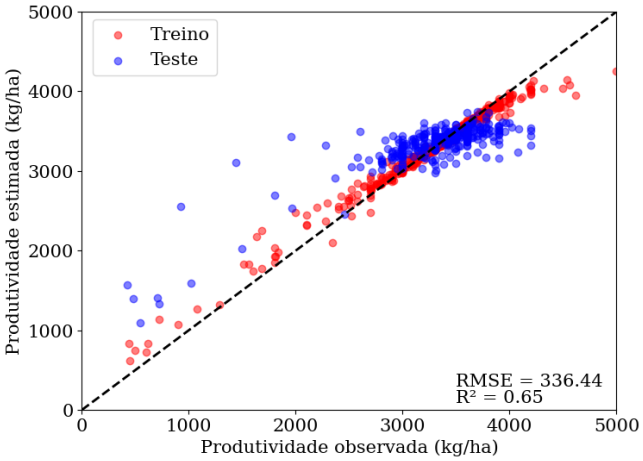


Figura 3. Gráfico de dispersão dos valores observados versus os valores estimados pelo modelo do Cenário 5 para a estimativa da produtividade de soja nas safras de 2019/2020 a 2021/2022.

4. CONCLUSÕES

O uso de imagens do satélite Sentinel-2 combinado com dados históricos de produtividade apresentou potencial para o desenvolvimento de um modelo empírico de estimativa de produtividade de soja a nível municipal para a região Centro-Oeste brasileira (RMSE de 336 kg/ha). O modelo desenvolvido com os índices EVI e NDRE apresentaram melhores resultados (R^2 acima de 0.56), indicando o potencial desses índices para a estimativa da produtividade de soja. No entanto, a combinação de mais índices de vegetação permitiu o desenvolvimento de um modelo de estimativa de produtividade mais robusto (R^2 igual a 0.65).

5. AGRADECIMENTOS

Este trabalho foi apoiado pelo CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico, para o financiamento do projeto PreCISIA (424490/2021-8), por meio do Programa de Treinamento de Recursos Humanos em Áreas Estratégicas (RHAE): bolsas concedidas a Ester de Carvalho Pereira (350860/2022-0), Geovani Pereira dos Santos (350921/2022-9) e Gabriela Cristina Salgado (350923/2022-1). Os autores agradecem ao Instituto da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo projeto financiado 'Imagens de satélite e

aprendizado de máquina para estimativa de produtividade de cana-de-açúcar em regiões do estado de São Paulo' (número da concessão 2021/11183-5).

6. REFERÊNCIAS

[1] Y. Shendryk, R. Davy, and P. Thorburn, Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning, *Field Crops Res.*, vol. 260, p. 107984, Jan. 2021, doi: 10.1016/j.fcr.2020.107984.

[2] R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. V. Prasad, and I. A. Ciampitti, Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil, *Agric. For. Meteorol.*, vol. 284, p. 107886, Apr. 2020, doi: 10.1016/j.agrformet.2019.107886.

[3] T. van Klompenburg, A. Kassahun, and C. Catal, Crop yield prediction using machine learning: A systematic literature review, *Comput. Electron. Agric.*, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[4] Y. Ma, Z. Zhang, Y. Kang, and M. Özdoğan, Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach, *Remote Sens. Environ.*, vol. 259, p. 112408, Jun. 2021, doi: 10.1016/j.rse.2021.112408.

[5] L. G. T. Crusiol *et al.*, Strategies for monitoring within-field soybean yield using Sentinel-2 Vis-NIR-SWIR spectral bands and machine learning regression methods, *Precis. Agric.*, vol. 23, no. 3, pp. 1093–1123, Jun. 2022, doi: 10.1007/s11119-022-09876-5.

[6] B. Pejak *et al.*, Soya Yield Prediction on a Within-Field Scale Using Machine Learning Models Trained on Sentinel-2 and Soil Data, *Remote Sens.*, vol. 14, no. 9, p. 2256, May 2022, doi: 10.3390/rs14092256.

[7] Conab, Conab - Calendário Agrícola (plantio e colheita). Accessed: Jun. 03, 2024. [Online]. Available: http://www.conab.gov.br/institucional/publicacoes/outras-publicacoes/item/15406-calendario-agricola-plantio-e-colheita

[8] MapBiomas Project, Collection 8 of the Annual Land Cover and Land Use Maps of Brazil (1985-2022). MapBiomas Data, 2023. doi: 10.58053/MAPBIOMAS/VJIJCL.

[9] European Space Agency, Sentinel-2 MSI Level-2A BOA Reflectance. 2022. doi: 10.5270/S2_-znk9xsj.

[10] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, Monitoring vegetation systems in the Great Plains with ERTS, Jan. 1974. Accessed: May 16, 2024. [Online]. Available: https://ntrs.nasa.gov/citations/19740022614

[11] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, Overview of the radiometric and biophysical performance of the MODIS vegetation indices, *Remote Sens. Environ.*, vol. 83, no. 1–2, pp. 195–213, Nov. 2002, doi: 10.1016/S0034-4257(02)00096-2.

[12] A. A. Gitelson and M. N. Merzlyak, Signature Analysis of Leaf Reflectance Spectra: Algorithm Development for Remote Sensing of Chlorophyll, *J. Plant Physiol.*, vol. 148, no. 3–4, pp. 494–500, Jan. 1996, doi: 10.1016/S0176-1617(96)80284-7.

[13] B. Gao, NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space, *Remote Sens. Environ.*, vol. 58, no. 3, pp. 257–266, Dec. 1996, doi: 10.1016/S0034-4257(96)00067-3.

[14] L. Breiman, Random Forests, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[15] F. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[16] Y. Li *et al.*, A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering, *Int. J. Appl. Earth Obs. Geoinformation*, vol. 118, p. 103269, Apr. 2023, doi: 10.1016/j.jag.2023.103269.

[17] B. S. Christenson, W. T. Schapaugh Jr., N. An, K. P. Price, V. Prasad, and A. K. Fritz, Predicting Soybean Relative Maturity and Seed Yield Using Canopy Reflectance, *Crop Sci.*, vol. 56, no. 2, pp. 625–643, 2016, doi: 10.2135/cropsci2015.04.0237.

[18] X.-P. Song, H. Li, P. Potapov, and M. C. Hansen, Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning, *Agric. For. Meteorol.*, vol. 326, p. 109186, Nov. 2022, doi: 10.1016/j.agrformet.2022.109186.