Detecção de Soja com U-Net e MapBiomas

Vinicius Pereira Tavares de Sousa
Universidade Tecnológica Federal do Paraná – UTFPR
devviniciustavares@gmail.com

Rafael Gomes Mantovani
Universidade Tecnológica Federal do Paraná – UTFPR
rafaelmantovani@utfpr.edu.br

Daniel Campos
Universidade Tecnológica Federal do Paraná – UTFPR
danielcampos@utfpr.edu.br

Abstract

This work proposes a 3D U-Net architecture for the automatic segmentation of soybean fields using multitemporal and multispectral satellite imagery. A dataset was generated through the Google Earth Engine (GEE), combining spectral bands, vegetation indices (Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI)), and the MapBiomas soybean mask for the 2021/2022 growing season in southern Brazil. The use of vegetation indices is supported by their proven effectiveness in agricultural monitoring [1]. To address the strong class imbalance between soybean and non-soybean regions, a stratified balancing strategy was applied, ensuring a more uniform distribution across different levels of crop coverage. The proposed model integrates temporal information across five months of the crop cycle, enabling the extraction of both spatial and temporal patterns, similar to other works that explore multitemporal deep learning architectures [2]. A hybrid loss function combining balanced Binary Cross-Entropy (BCE) and the Dice coefficient was employed to improve segmentation accuracy, especially in regions with low soybean presence. Experimental results demonstrate that the model achieves stable learning, with early stopping preventing overfitting and the best validation performance occurring at epoch 13. Visual analyses confirm strong spatial consistency between predictions and reference masks. The study highlights the potential of 3D convolutional architectures for large-scale automated crop mapping [3] and suggests future improvements by incorporating additional seasons and broader geographic coverage.

Keywords

Soybean segmentation, multitemporal imagery, 3D U-Net, Google Earth Engine, agricultural remote sensing

1 Introdução

O monitoramento de culturas agrícolas por meio de imagens de satélite tem se tornado cada vez mais importante para apoiar decisões no campo e melhorar o acompanhamento da produção. A soja, em especial, é uma das principais culturas do Brasil, o que torna fundamental o desenvolvimento de métodos automáticos capazes de identificar suas áreas de plantio de forma rápida e confiável.

Com o avanço de plataformas como o *Google Earth Engine* (GEE), tornou-se possível extrair grandes volumes de dados multitemporais de maneira escalável [4]. Ao mesmo tempo, arquiteturas de *Deep Learning*, como a *U-Net*, têm demonstrado excelente desempenho em tarefas de segmentação de imagens devido ao seu uso eficiente de conexões de atalho e preservação de detalhes espaciais [5].

Neste trabalho, utilizamos o GEE para gerar um conjunto de dados contendo bandas espectrais, índices de vegetação (*NDVI* e *NDWI*) e a máscara de soja para cinco meses do ciclo agrícola 2021/2022 nos estados do Paraná e Rio Grande do Sul. O uso desses índices é amplamente validado na literatura para análise fenológica e monitoramento agrícola [1]. Em seguida, treinamos um modelo *U-Net 3D* para segmentar automaticamente áreas de soja a partir dessas imagens, incorporando tanto a variabilidade espacial quanto temporal, como sugerido em estudos de aprendizado profundo multitemporal [2]. O objetivo é avaliar a capacidade da rede de reconstruir a máscara de soja com base em informações espectrais e temporais, oferecendo uma alternativa eficiente e automatizada para o mapeamento agrícola em grande escala.

2 Trabalhos relacionados

Diversos estudos têm explorado o uso de imagens de satélite combinadas com métodos de *Deep Learning* para a detecção e mapeamento de culturas agrícolas. Técnicas tradicionais baseadas em índices de vegetação, como o *NDVI* e o *Enhanced Vegetation Index* (EVI), já foram amplamente utilizadas para monitoramento de lavouras devido à sua sensibilidade à variação espectral ao longo do ciclo fenológico [1]. No entanto, métodos estatísticos e limiares fixos apresentam limitações em cenários heterogêneos, sendo sensíveis a ruídos atmosféricos, nuvens e variações regionais.

Com o avanço das redes neurais convolucionais, arquiteturas como a *U-Net* têm se consolidado como a principal escolha para tarefas de segmentação semântica [5]. A *U-Net* destaca-se por permitir a recuperação de detalhes espaciais por meio de suas conexões de atalho, o que a torna particularmente adequada para segmentação de imagens com padrões espaciais complexos, como áreas agrícolas. Trabalhos recentes têm demonstrado que modelos *U-Net* aplicados a dados multiespectrais conseguem capturar padrões relevantes mesmo em regiões com grande variação temporal [2].

Além disso, abordagens multitemporais têm sido exploradas com o uso de redes 3D ou combinações de convoluções 2D com modelos recorrentes. As convoluções 3D permitem explorar simultaneamente padrões espaçotemporais, proporcionando representações mais ricas da dinâmica da cultura [3]. Em aplicações agrícolas, essas abordagens têm se mostrado eficazes para identificar transições fenológicas e reduzir ambiguidades espectrais [6].

O uso de plataformas como o *Google Earth Engine* tem se tornado um componente essencial na construção de *pipelines* de mapeamento agrícola em larga escala, graças à sua capacidade de processar grandes séries temporais de forma distribuída [4]. Trabalhos recentes demonstram resultados promissores ao combinar o GEE com redes neurais profundas para mapeamento automatizado de soja [7].

Dessa forma, o presente estudo se apoia nessas contribuições ao integrar dados multitemporais, bandas espectrais, índices de vegetação e uma arquitetura *U-Net 3D*, buscando capturar tanto a

estrutura espacial quanto a evolução temporal das áreas de soja no sul do Brasil.

2.1 Dataset

O conjunto de dados foi construído utilizando a plataforma *Google Earth Engine* [4], a partir de imagens multitemporais. Foram selecionados cinco meses do ciclo agrícola, de outubro de 2021 a fevereiro de 2022, nos estados do Paraná e Rio Grande do Sul, abrangendo o período crítico de desenvolvimento da cultura da soja.

Para cada ponto geográfico, foram extraídas as bandas RGB, a banda de infravermelho e dois índices amplamente utilizados no monitoramento agrícola: *NDVI* e *NDWI* [1]. A máscara de soja fornecida pelo MapBiomas foi utilizada como referência binária, sendo 1 para presença de soja (classe positiva) e 0 para ausência.

Define-se como amostra cada recorte espacial fixo correspondente a uma sequência temporal de cinco meses, organizado em um tensor no formato $(5, 128, 128, 6)$, com resolução espacial de 128×128 pixels e seis variáveis espectrais por mês. As máscaras associadas foram estruturadas no formato $(5, 128, 128, 1)$.

Todas as bandas foram normalizadas para o intervalo [0,1], garantindo padronização numérica entre as variáveis e maior estabilidade no processo de treinamento.

Inicialmente, o conjunto total de dados foi dividido em 80% para treinamento, 10% para validação e 10% para teste, correspondendo a 17.600, 2.200 e 2.200 amostras, respectivamente. Os conjuntos de validação e teste permaneceram inalterados ao longo de todo o processo, preservando a distribuição original dos dados.

Devido ao forte desequilíbrio entre regiões com e sem soja, o balanceamento estratificado foi aplicado exclusivamente ao conjunto de treinamento. Para cada amostra de treino, calculou-se o percentual médio de pixels classificados como soja ao longo das cinco imagens temporais, representando a proporção de área ocupada pela cultura naquele recorte.

As amostras foram agrupadas em faixas de proporção variando de 0% a 85%, com incremento de 0,5 ponto percentual. A faixa correspondente a 0% (ausência total de soja) foi tratada separadamente para evitar predominância excessiva de regiões negativas. Para cada faixa, estabeleceu-se um limite máximo de amostras; quando esse limite era excedido, realizava-se amostragem aleatória sem reposição dentro da própria faixa.

Esse procedimento reduziu a predominância das classes majoritárias no conjunto de treinamento, mantendo inalteradas as distribuições dos conjuntos de validação e teste e contribuindo para uma avaliação mais realista da capacidade de generalização do modelo.

2.2 Arquitetura do Modelo

A arquitetura proposta consiste em uma *U-Net 3D* adaptada para séries temporais multiespectrais. O modelo recebe como entrada amostras contendo cinco meses consecutivos de observações, cada um com seis bandas. A rede segue a estrutura *encoder–decoder* tradicional da U-Net, reconhecidamente eficaz em segmentação [5].

O *encoder* reduz a dimensionalidade espacial por meio de convoluções 3D, extraindo padrões complexos que combinam variação espectral, espacial e temporal, como destacado em trabalhos de aprendizado espaçotemporal [3]. O *decoder* reconstrói a máscara final utilizando camadas de *Conv3DTranspose*, normalização em lote e ativação *LeakyReLU*. As *skip connections* permitem que informações espaciais de alta resolução sejam preservadas durante a reconstrução.

A Tabela 1 resume a arquitetura completa do modelo.

Table 1: Arquitetura do modelo 3D proposto para segmentação de soja

Bloco	Camada	Filtros	Saída	Observações
Entrada	–	–	(5, 128, 128, 6)	5 meses, 6 bandas
Encoder	Conv3D + BN + LeakyReLU	32	(5, 64, 64, 32)	Kernel $8 \times 8 \times 8$, stride (1, 2, 2)
Encoder	Conv3D + BN + LeakyReLU	64	(5, 32, 32, 64)	Kernel $4 \times 4 \times 4$, stride (1, 2, 2)
Encoder	Conv3D + BN + LeakyReLU	128	(5, 16, 16, 128)	Kernel $2 \times 2 \times 2$, stride (1, 2, 2)
Latente	Flatten + Dense	–	(256)	Vetor latente
Decoder	Dense + Reshape	–	(5, 16, 16, 128)	Reconstrução inicial
Decoder	Conv3DTranspose + BN + LeakyReLU	128	(5, 32, 32, 128)	*Skip connection*
Decoder	Conv3DTranspose + BN + LeakyReLU	64	(5, 64, 64, 64)	*Skip connection*
Decoder	Conv3DTranspose + BN + LeakyReLU	32	(5, 128, 128, 32)	*Skip connection*
Saída	Conv3DTranspose (*sigmoid*)	1	(5, 128, 128, 1)	Máscara final

2.3 Função de Perda

A função de perda combina dois componentes fundamentais: a *Binary Cross-Entropy* (BCE) balanceada e o coeficiente de *Dice*. A BCE balanceada utiliza pesos ajustados dinamicamente de acordo com a proporção de pixels positivos e negativos em cada *batch*, reduzindo o impacto do desequilíbrio entre classes.

O coeficiente de *Dice* mede a similaridade entre a máscara prevista e a real, favorecendo a sobreposição das regiões segmentadas. A combinação destas métricas segue práticas comuns em segmentação biomédica e remota.

2.4 Configurações de Treinamento

O modelo foi treinado utilizando o otimizador *Adam*, com taxa de aprendizado inicial de 0,001. O treinamento utilizou *batches* de tamanho 64, com até 500 épocas.

Foi aplicado *early stopping* monitorando a perda de validação, interrompendo o treinamento quando não houve melhora por 10 épocas consecutivas e restaurando automaticamente os melhores pesos.

3 Experimentos

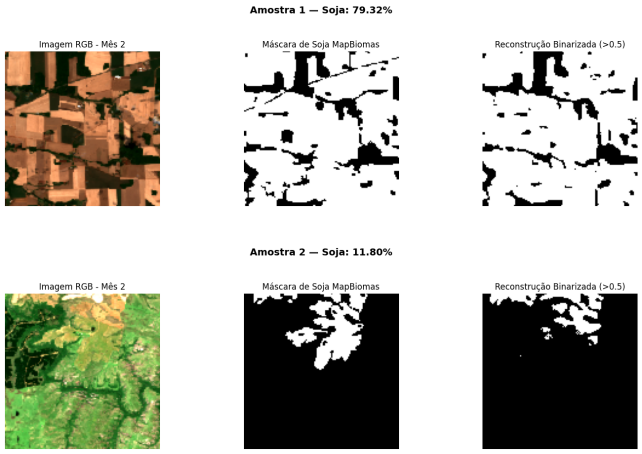
3.1 Avaliação do Treinamento

A Tabela 2 apresenta a evolução da perda ao longo do treinamento. A menor perda de validação ocorreu na época 13.

Table 2: Resumo dos principais resultados durante o treinamento

Época	Perda de Treino	Perda de Validação
1	0.9702	8.4698
5	0.4971	0.5901
8	0.4574	0.6800
10	0.4316	0.9603
13 (melhor)	**0.3866**	**0.5117**
20	0.2992	0.6323
23	0.2834	1.1726

Figure 1: Exemplos das imagens RGB de entrada, das máscaras binárias de referência e das predições geradas pelo modelo.



O treinamento foi interrompido na época 23, após 10 épocas consecutivas sem melhora na perda de validação, conforme critério de early stopping. A melhor performance foi obtida na época 13, cujos pesos foram restaurados automaticamente para avaliação final.

3.2 Avaliação no Conjunto de Teste

O limiar de binarização foi definido com base no conjunto de validação, sendo obtido o valor de 0,8 como aquele que maximizou o coeficiente de *Dice*.

Aplicando-se esse limiar ao conjunto de teste, mantido completamente isolado durante o treinamento e ajuste de hiperparâmetros, o modelo alcançou coeficiente médio de *Dice* igual a 0,76, com desvio padrão de 0,29.

Esse resultado indica boa concordância média entre as máscaras preditas e as máscaras de referência, embora com variabilidade entre diferentes amostras, refletindo a heterogeneidade espacial e espectral das regiões analisadas.

3.3 Visualização dos Resultados

A Figura 1 apresenta amostras do conjunto de teste, incluindo a imagem RGB original, a máscara de referência e a predição binarizada gerada pela rede.

4 Considerações Finais

Os resultados mostram que a *U-Net 3D* conseguiu aprender padrões espaciais e temporais relevantes para identificar áreas de soja. A curva de perda apresentou redução consistente até o acionamento do *early stopping*, com melhor desempenho na época 13. As visualizações evidenciam que as predições seguem de forma coerente as regiões de soja, mesmo em áreas heterogêneas.

Limitações persistem, como dificuldades em regiões com baixa reflectância ou transições sutis. Métodos baseados em séries temporais mais longas [6] podem contribuir para superar essas limitações.

Como trabalhos futuros, recomenda-se incluir múltiplas safras para aumentar a variabilidade espacial e temporal, ampliando a robustez do modelo, como sugere [7].

Referências

[1] Alfredo R. Huete, Hongxing Liu, Kamel Batchily, and Willem van Leeuwen. A comparison of vegetation indices over a global set of tm images for eos-modis. *Remote Sens. Environ.*, 59(3):440–451, 1999.

[2] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.*, 57(7):4229–4245, 2019.

[3] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 4489–4497, 2015.

[4] Noel Gorelick, Matt Hancher, Mike Dixon, Sergey Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.*, 202:18–27, 2017.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[6] Anderson Ruhoff, Gustavo Araújo, Diogo Antunes, et al. Soybean mapping using multi-temporal remote sensing data and deep learning approaches. *Remote Sens.*, 14(9):2051, 2022.

[7] Jeferson dos Santos, Daniel da Silva, Michelle Picoli, et al. Mapping soybean croplands in brazil using satellite remote sensing and deep learning. *ISPRS J. Photogramm. Remote Sens.*, 168:65–81, 2020.