Previsão de Produtividade de Soja Utilizando Adaptação de Domínio

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

Edivan José Possamai
Instituto de Desenvolvimento Rural do Paraná – IDRPR
edivanjp@idr.pr.gov.br

Anderson de Toledo
Instituto de Desenvolvimento Rural do Paraná – IDRPR
anderson@idr.pr.gov.br

Abstract

Soybean yield prediction is essential for optimizing agricultural management and supporting decision-making in the agribusiness sector. The lack of yield data at the farm level hinders the application of machine learning models for accurate estimation. This work proposes the use of domain adaptation techniques as a strategy to overcome this limitation, transferring knowledge from a model trained at the municipal level, where yield labels are publicly available, to farm-level predictions. The methodology employs time series of vegetation indices and climate variables derived from satellite data, processed by a deep learning model to capture patterns relevant to soybean yield. For evaluation, the results are compared with yields from previous harvests of farms with known data, allowing assessment of the estimation accuracy. This approach is expected to contribute to more efficient precision agriculture practices and support strategic decisions in agricultural planning.

Keywords

machine learning, time series analysis, agricultural productivity, precision agriculture

1 Introdução

A previsão de produtividade agrícola é essencial para o planejamento estratégico do setor agropecuário. Trabalhos recentes mostram o potencial do uso de imagens de satélite combinadas com arquiteturas de *deep learning* para previsão de produtividade em escalas agregadas [1, 2]. Modelos capazes de capturar padrões espaciais e temporais têm obtido bons resultados em nível municipal, permitindo a extração de informações relevantes a partir de séries temporais multisensoriais.

Métodos de *domain adaptation* tornam possível transferir conhecimento de domínios com rótulos (por exemplo, municípios) para domínios sem rótulos (propriedades), permitindo estimativas localizadas sem a necessidade de grandes bancos de dados rotulados a nível de propriedade [3, 4]. Este trabalho explora uma abordagem adversarial para adaptação de domínio, com o objetivo de prever a produtividade de soja em propriedades rurais a partir de dados municipais rotulados e dados de propriedades sem rótulos.

2 Fundamentação Teórica

2.1 Previsão em Escalas Agregadas

Em estudos de previsão agrícola, diferentes níveis de agregação espacial podem ser considerados. Escalas agregadas referem-se a unidades territoriais amplas, como municípios ou regiões administrativas, nas quais os dados de produtividade são disponibilizados de forma consolidada por órgãos oficiais. Nesses casos, os valores representam médias espaciais da produção em grandes áreas cultivadas.

Modelos baseados em *deep learning* têm sido amplamente empregados para previsão de produtividade agrícola em nível regional ou municipal, utilizando séries temporais de imagens de satélite e dados climáticos [1, 2, 5]. Esses trabalhos demonstram que arquiteturas profundas são capazes de capturar padrões espaço-temporais complexos, especialmente quando aplicadas a dados multissensoriais.

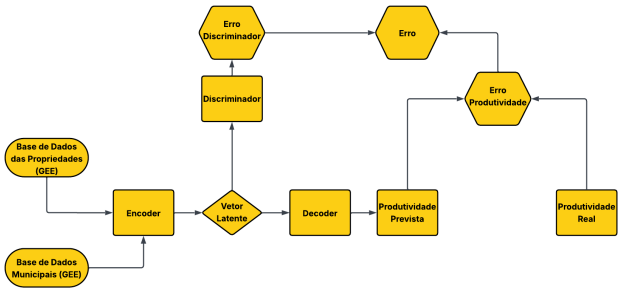
Embora modelos treinados em escalas agregadas apresentem boa capacidade de generalização regional, sua aplicação direta em nível de propriedade rural pode resultar em perda de precisão devido à heterogeneidade espacial, diferenças de manejo e variações microclimáticas. Essa discrepância caracteriza um cenário clássico de mudança de domínio, no qual as distribuições estatísticas entre conjuntos de treinamento e aplicação diferem significativamente [4, 6].

2.2 Adaptação de Domínio

Adaptação de domínio é uma subárea de aprendizado de máquina inserida no contexto de *transfer learning*, cujo objetivo é transferir conhecimento de um domínio fonte, no qual há dados rotulados disponíveis, para um domínio alvo com pouca ou nenhuma rotulação [4]. Formalmente, considera-se que os domínios possuem distribuições de probabilidade distintas, isto é, $P_s(X, Y) \neq P_t(X, Y)$.

Abordagens clássicas de adaptação buscam reduzir a divergência entre domínios por meio do alinhamento de distribuições estatísticas, como no método *Deep CORAL* [7], ou por estratégias de generalização de domínio [8]. Métodos mais recentes utilizam treinamento adversarial para promover invariância de domínio no espaço latente [3, 6]. Nesse paradigma, um discriminador tenta distinguir a origem dos dados (fonte ou alvo), enquanto o extrator

Figure 1: Arquitetura geral do modelo proposto



Fonte: Autoria própria (2025).

de características é treinado para tornar essa distinção impossível, promovendo alinhamento representacional.

Neste trabalho, o domínio fonte corresponde aos dados municipais, para os quais há produtividade oficialmente reportada, enquanto o domínio alvo refere-se às propriedades rurais individuais. A abordagem adversarial utilizada busca alinhar as distribuições latentes extraídas pelo *encoder* por meio de um discriminador, minimizando a divergência entre os domínios no espaço de representação.

3 Solução Proposta

O processo incluiu extração dos dados de satélite e climáticos via *Google Earth Engine (GEE)*, pré-processamento (limpeza, agregação mensal, balanceamento e normalização), treinamento do modelo com *encoder–latente–decoder* e discriminador adversarial e avaliação em nível municipal e em propriedades fornecidas pelo IDR-Paraná.

As principais fontes de dados foram *MapBiomas*[1], *Sentinel-2*[2] e *ERA5-Land*[3], cobrindo as safras de 2019 a 2023. As *features* utilizadas foram *NDVI*, *NDWI*, temperatura máxima, temperatura mínima, precipitação e radiação solar, agregadas mensalmente ao longo do ciclo da safra, produzindo entradas no formato $(n, 5, 6)$.

Na etapa de seleção e balanceamento, municípios com área plantada entre 1000 e 20000 ha e propriedades entre 1 e 50 ha foram mantidos. *Outliers* fora do intervalo 600–4300 kg/ha foram removidos. Aplicou-se balanceamento por faixas e avaliou-se a similaridade estatística entre domínios utilizando a distância de *Wasserstein* por *feature* e por mês, orientando o alinhamento prévio entre distribuições.

A arquitetura do modelo combina *encoder* convolucional 1D, vetor latente compacto, discriminador adversarial (*MLP*) e *decoder* de regressão quantílica (Q_{10}, Q_{50}, Q_{90}). A Figure 1 apresenta uma visão geral dessa arquitetura proposta.

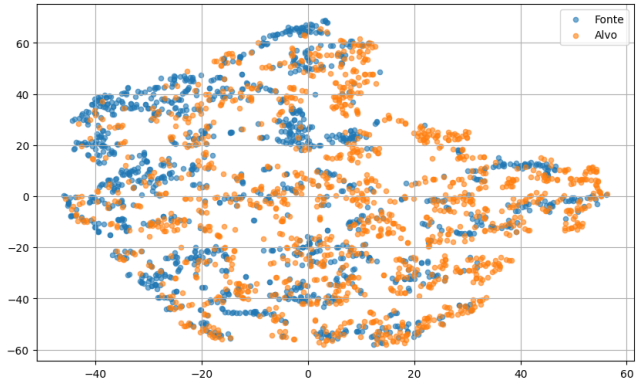
A função de perda total combina *MAE*, perda quantílica e termo adversarial:

$$\mathcal{L}_{total} = \lambda_r \mathcal{L}_{reg} + \lambda_d \mathcal{L}_{adv}.$$

Gradientes penalizados garantiram estabilidade.

[1] https://mapbiomas.org/
[2] https://sentinel.esa.int/web/sentinel/missions/sentinel-2
[3] https://cds.climate.copernicus.eu/

Figure 2: Resultado da sobreposição entre os espaços latentes



Fonte: Autoria própria (2025).

A validação incluiu avaliação municipal, análise de sobreposição de espaços latentes via *t-SNE* e teste em propriedades reais do IDR-Paraná.

4 Resultados e Discussão

Após as etapas de pré-processamento e seleção das amostras, o conjunto de dados final foi composto por aproximadamente 1400 amostras em nível municipal e 1200 amostras referentes a propriedades rurais. Essa diferença reflete as restrições impostas pela disponibilidade de dados de campo validados, especialmente no contexto de propriedades individuais.

A análise de similaridade estatística entre os domínios de treinamento (municipal) e validação (propriedades) foi conduzida por meio da distância de *Wasserstein*. Os resultados indicaram menor divergência para os índices espectrais, como *NDVI* e *EVI*, enquanto variáveis climáticas apresentaram maior variabilidade entre os domínios. Esse comportamento evidencia que informações espectrais tendem a ser mais consistentes espacialmente, enquanto variáveis climáticas são fortemente influenciadas por condições locais, dificultando o processo de adaptação de domínio.

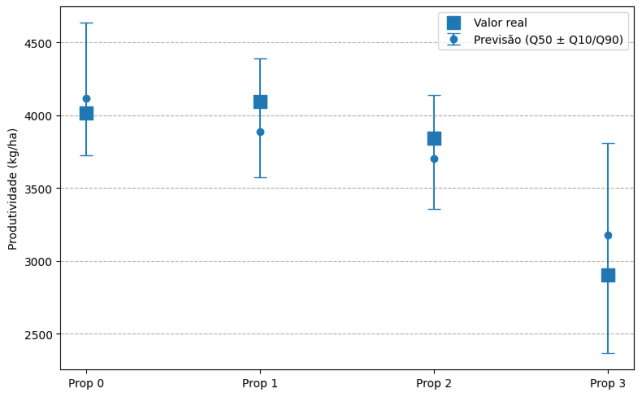
Em nível municipal, o modelo apresentou um erro absoluto médio (*MAE*) de aproximadamente 401,66 kg/ha para o quantil mediano (Q_{50}), valor considerado competitivo frente a trabalhos relacionados na literatura. A projeção *t-SNE* apresentada na Figure 2 evidencia uma forte sobreposição entre os espaços latentes dos domínios de origem e alvo, indicando que o mecanismo adversarial foi eficaz em alinhar as representações internas do modelo, reduzindo o deslocamento de domínio.

No nível de propriedades, foram avaliadas quatro áreas experimentais monitoradas pelo IDR-Paraná. As previsões geradas pelo modelo para essas propriedades estão apresentadas na Table 1. Observa-se que o modelo manteve coerência entre os valores previstos e os valores reais, resultando em um *MAE* aproximado de 206,1 kg/ha para o quantil Q_{50}. Esse desempenho superior em relação ao nível municipal sugere que, apesar do conjunto reduzido de amostras, o modelo foi capaz de capturar padrões produtivos locais de forma consistente.

Table 1: Previsões do modelo em comparação com os valores reais (kg/ha)

Amostra	Q_{10}	Q_{50}	Q_{90}	Real
0	3724,75	4115,32	4633,32	4016,40
1	3573,84	3887,53	4391,18	4090,80
2	3353,07	3703,73	4138,89	3843,00
3	2366,15	3174,11	3806,72	2901,00

Figure 3: Resultado da previsão a nível de propriedade



Fonte: Autoria própria (2025).

A Figure 3 ilustra os intervalos quantílicos estimados pelo modelo e sua relação com os valores observados nas propriedades analisadas. Nota-se que os valores reais encontram-se dentro do intervalo definido pelos quantis Q_{10} e Q_{90}, indicando boa calibração da incerteza preditiva. Essa característica é particularmente relevante em aplicações agrícolas, nas quais a variabilidade inerente ao sistema produtivo deve ser considerada no processo de tomada de decisão.

De forma geral, os resultados demonstram que o modelo apresenta bom desempenho em escala municipal e mantém coerência ao ser transferido para o nível de propriedades, mesmo diante de um cenário de adaptação de domínio. Entre as principais limitações do estudo, destacam-se o tamanho reduzido do conjunto de validação em propriedades, as divergências climáticas entre os domínios analisados e a presença de lacunas temporais nas séries do Sentinel-2, causadas principalmente por cobertura de nuvens. Esses fatores indicam que a incorporação de séries temporais mais longas, bem como a ampliação do número de propriedades avaliadas, pode contribuir para aumentar a robustez e a generalização do modelo em trabalhos futuros.

5 Considerações Finais

Este trabalho apresentou uma metodologia de previsão de produtividade de soja baseada em *deep learning* e *domain adaptation* adversarial. A adaptação de domínio se mostrou promissora para transferir conhecimento do nível municipal para propriedades.

Modelos quantílicos facilitaram a interpretação da incerteza, e o pré-processamento baseado em *Wasserstein* foi crucial.

Como trabalho futuro, pretende-se ampliar a validação em nível de propriedade com mais amostras, visando robustez estatística e maior confiabilidade.

Referências

[1] A. Kamilaris and F. X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.
[2] M. Fathi, R. Shah-Hosseini, and A. Moghimi. 3d-resnet-bilstm model: A deep learning model for county-level soybean yield prediction with time-series sentinel-1, sentinel-2 imagery, and daymet data. *Remote Sensing*, 15(23):5551, 2023.
[3] Y. Ganin and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016.
[4] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
[5] T. He, M. Li, and D. Jin. Deep learning-based time series prediction for precision field crop protection. *Frontiers in Plant Science*, 16:1575796, 2025.
[6] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Computing Surveys*, 53(2):1–34, 2020.
[7] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 443–450, 2016.
[8] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2551–2559, 2015.