DETECTANDO PADRÕES ATÍPICOS DE AGRICULTURA EM APLICAÇÕES DE CRÉDITO RURAL BASEADO EM SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITES

Baggio Luiz de Castro e Silva[1], *Karine Reis Ferreira*[1], *Gilberto Ribeiro de Queiroz*[1], *Marcos Adami*[1], *Thales Körting*[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE),
Av. Dos Astronautas 1758, 12227-010 – São José dos Campos, SP – Brazil
{baggio.silva, karine.ferreira, gilberto.queiroz, marcos.adami, thales.korting}@inpe.br

RESUMO

Este artigo apresenta uma metodologia para a detecção de padrões atípicos em aplicações de crédito rural agrícola, baseada em séries temporais de imagens de satélites. A metodologia utiliza os métodos de agrupamento Mapas Auto-Organizáveis (*Self-Organizing Maps - SOM*) e Hierárquico, empregando as distâncias Euclidiana e DTW (*Dynamic Time Warping*). A metodologia proposta foi aplicada a um conjunto de glebas de soja adquiridas do Sistema de Operações do Crédito Rural e do Proagro (Sicor), obtendo resultados promissores apresentados nesse trabalho.

Palavras-chave – Sensoriamento remoto, séries temporais de imagens, aprendizado não supervisionado, padrões atípicos, agricultura.

ABSTRACT

This paper presents a methodology for detecting atypical patterns in agricultural rural credit applications, based on time series from satellite images. The methodology employs clustering methods, specifically Self-Organizing Maps (SOM) and Hierarchical clustering, using Euclidean and Dynamic Time Warping (DTW) distances. The proposed methodoly was applied to soybean fields obtained from the Rural Credit Operations System and Proagro (Sicor), obtaing promising results presented in this work.

Key words – *Remote sensing, image time series, unsupervised learning, atypical patterns, agriculture.*

1. INTRODUÇÃO

O setor de crédito rural e seguro agrícola no Brasil é essencial para garantir a segurança alimentar e o desenvolvimento econômico e social do país, fornecendo suporte financeiro e proteção para que agricultores de diversos perfis — desde pequenos produtores familiares até grandes empresas — possam investir em práticas agrícolas modernas e sustentáveis [1]. Programas como o *Programa de Garantia da Atividade Agropecuária* (Proagro), gerido pelo Banco Central do Brasil, foram desenvolvidos para proteger os produtores rurais contra perdas decorrentes de eventos climáticos adversos, pragas e doenças, assegurando a continuidade da atividade agropecuária e a estabilidade financeira [2].

De acordo com o Manual de Crédito Rural (MCR), os registros das operações de crédito rural concedidas pelas instituições financeiras autorizadas, bem como os enquadramentos de empreendimentos no Proagro, são realizados no *Sistema de Operações do Crédito Rural e do Proagro* (Sicor) [1]. Para assegurar que apenas perdas legítimas sejam cobertas, o Proagro possui mecanismos de comprovação e auditoria, incluindo a Comunicação de Perdas (COP) e o Relatório de Comprovação de Perdas (RCP), gerados por técnicos independentes. Esses documentos são essenciais para validar as informações fornecidas e garantir que os recursos sejam destinados a quem realmente necessita [1]. Entretanto, práticas fraudulentas, como declarações incorretas sobre áreas cultivadas e simulação de perdas, comprometem a sustentabilidade do programa e a alocação equitativa dos recursos públicos.

Nesse contexto, a integração de tecnologias, como a análise de séries temporais de imagens e sensoriamento remoto, surge como uma abordagem promissora para identificar fraudes de maneira mais precisa e eficaz. A análise de séries temporais de imagens permite monitorar o comportamento das atividades agrícolas ao longo do tempo, possibilitando a verificação independente das informações declaradas pelos beneficiários do Proagro. O uso de índices de vegetação derivados de imagens de satélite, como o Índice de Vegetação por Diferença Normalizada (NDVI), permite distinguir áreas com diferentes práticas agrícolas, como irrigação e sequeiro, e monitorar a saúde das culturas ao longo da temporada de cultivo. Estudos indicam que a integração de dados de sensoriamento remoto e técnicas de mineração de dados pode reduzir significativamente falsos positivos e aprimorar a detecção de anomalias em programas de seguro agrícola [3].

Técnicas de agrupamento têm sido aplicadas no campo de sensoriamento remoto para identificar diferentes padrões espectro-temporais, possibilitando a identificação de padrões distintos em imagens de satélite [4]. Em particular, esses métodos permitem detectar áreas que apresentam comportamentos espectrais divergentes dos esperados para culturas agrícolas. Neste contexto, são considerados como anomalias os padrões que diferem daqueles típicos de cultivos agrícolas. A identificação inicial dessas anomalias facilita a detecção de regiões que possivelmente estão sendo utilizadas para fins distintos dos declarado, o que é crucial para a mitigação de fraudes em programas de crédito rural.

Em estudos sobre detecção de fraudes em seguros, métodos de aprendizado de máquina, têm se mostrado promissores. Por exemplo, [5] examinaram o uso de aprendizado de máquina para detectar fraudes em seguros, enquanto [6] aplicaram detecção de anomalias baseada em conjuntos de detectores para prever fraudes em seguros. Essas pesquisas

ressaltam a eficácia da integração de análise de séries temporais de imagens de sensoriamento remoto e métodos de clusterização na identificação de fraudes e na melhoria dos sistemas de monitoramento, incluindo o crédito rural. Neste sentido, o presente estudo busca expandir essas abordagens, propondo uma metodologia que combina o uso de dados de séries temporais de imagens de sensoriamento remoto com técnicas de aprendizado de máquina não supervisionados para identificar inconsistências em áreas que obtiveram financiamento para cultivo agrícola de culturas de ciclo anual, mais especificamente a cultura de soja. Assim, para o presente trabalho é definido como anomalias todos os que diferem dos padrões espectro temporais de esperados para cultivos agrícolas de cultura de ciclo anual. Com essa abordagem, espera-se aprimorar a eficácia do Proagro, garantindo que os recursos públicos destinados ao setor agrícola sejam alocados de forma justa e eficiente.

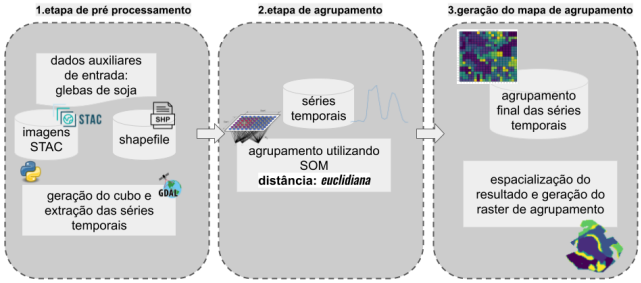

Figura 1: Metodologias de análise para glebas individuais

2. MATERIAL E MÉTODOS

Para avaliar a possibilidade de detectar anomalias em glebas de soja obtidas do SICOR por meio da clusterização de séries temporais de imagens, foram selecionadas 11 glebas de soja de diferentes estados brasileiros, incluindo Paraná e Rio Grande do Sul, para a análise. A metodologia adotada divide-se em duas abordagens principais. Para assegurar a reprodutibilidade dos resultados, o Jupyter Notebook detalhado da metodologia está disponível no seguinte link: https://www.kaggle.com/code/castrobaggio/sbsr-2025-padroes-series-temporais-imagens.

2.1. Análise Individual de Glebas

A primeira abordagem foca na identificação individualizada de anomalias, observando apenas as séries temporais de uma única gleba (Figura 1). Todas as séries temporais de uma gleba foram coletadas e submetidas à clusterização utilizando Mapas Auto-Organizáveis (*Self-Organizing Maps* - SOM) [7] com a distancia euclidiana. O SOM é um algoritmo de aprendizado não supervisionado que cria uma grade bidimensional para representar dados multidimensionais, efetivamente reduzindo a dimensionalidade. O algoritmo agrupa as séries temporais em clusters, cada um representado por um *codebook vector* (ou vetor de pesos). Assim, o conjunto de séries temporais que estão no mesmo cluster tem mais similaridades a esse padrão do que a qualquer outro. Essa abordagem permite identificar padrões não relacionados à agricultura dentro de uma gleba, como áreas florestais, edificações, corpos d'água, áreas em pousio, solo exposto dentre outros usos.

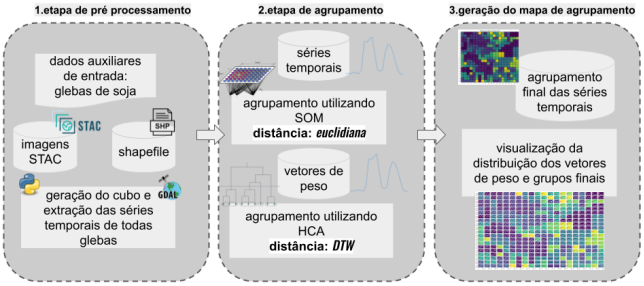

Figura 2: Metodologias de análise para múltiplas glebas

2.2. Análise para múltiplas glebas

A segunda abordagem trata da análise de grandes volumes de dados e da utilização de metodologias não supervisionadas para auxiliar na identificação de padrões a priori não agrícolas em múltiplas glebas (Figura 2). Para exemplificar, foi utilizado um conjunto de 11 glebas de soja de diferentes locais. O processo inicia-se com a clusterização utilizando Self-Organizing Maps (SOM), que agrupou os dados com base na distância euclidiana. Devido à sua capacidade de lidar com dados de alta dimensionalidade e produzir uma representação em baixa dimensionalidade adequada para visualização e interpretação, a clusterização com SOM cria uma grade bidimensional, agrupando as séries temporais para obter representatividade através dos *codebook vectors*. Entretanto, devido à complexidade e diversidade de padrões em grandes conjuntos de dados, torna-se necessário empregar técnicas de agrupamento em conjunto (*ensemble clustering*) [8] para minimizar a formação de clusters muito similares, que facilita a rápida análise por parte dos analistas de sensoriamento remoto. Para isso, a clusterização hierárquica (*Hierarchical clustering*) [9] foi aplicada aos *codebook vectors* obtidos do SOM utilizando a distância *Dynamic time warping* (DTW) [10], permitindo a seleção de um limiar de distância mínimo ou um número pré-definido de clusters finais, conforme necessário pelo analista. Esta etapa agrupa os clusters iniciais obtidos do SOM, reduzindo o número total de clusters e simplificando a análise.

2.3. Aquisição e Pré-processamento dos Dados

As séries temporais foram extraídas de cubos de dados Sentinel-2/MSI com resolução espacial de 10 metros e composição temporal de 16 dias, utilizando o método *Least Cloudy Pixel* (LCF), produzidos pelo projeto *Brazil Data Cube* [11]. *Brazil Data Cube* é uma iniciativa brasileira para produzir cubos de dados prontos para análise para todo o território brasileiro. Os cubos de dados foram preparados para uso direto nos modelos, assegurando padronização temporal e espacial das imagens, o que facilitou a extração das séries temporais para análise.

Na análise individual de glebas, utilizou-se exclusivamente o Índice de Vegetação por Diferença Normalizada (NDVI). A clusterização foi aplicada a cada gleba com base nos valores de NDVI coletados entre 9 de agosto de 2021 e 9 de agosto de 2022, totalizando 24 imagens. Para a

clusterização de múltiplas glebas, empregaram-se os dados das bandas 'B04', 'B11' e o NDVI no mesmo período. As séries temporais foram pré-processadas para remover efeitos de nuvens, minimizando ruídos e aprimorando os perfis espectro-temporais. A banda de Classificação de Cena (SCL) identificou classes como "Sem dados", "Saturado ou defeituoso", "Áreas escuras", "Sombra de nuvem", "Probabilidade média de nuvem", "Alta probabilidade de nuvem", "Cirrus tênue"e "Neve". Valores nessas classes foram substituídos pela interpolação linear simples.

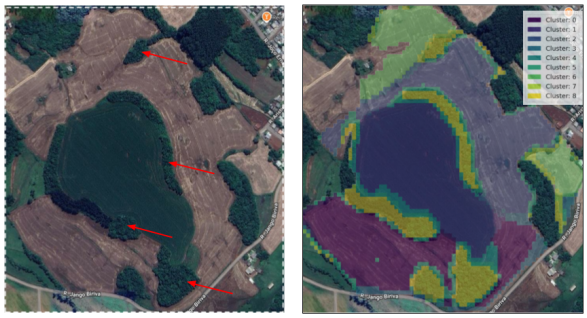

Figura 3: Figura à esquerda: Imagem de alta resolução com setas em vermelho indicado regiões com porte florestal; Figura à direta: Classificação não supervisionada da gleba indicando 9 grupos diferentes em transparência a uma imagem de alta resolução para comparação

3. RESULTADOS

Os resultados são apresentados qualitativamente, por meio de interpretação visual, mostrando a clusterização de uma gleba específica e, posteriormente, a execução para identificar padrões atípicos de agricultura em um conjunto de 11 glebas.

A Figura 3 exibe o resultado da Análise Individual de Glebas, que oferece suporte ao analista de sensoriamento remoto na interpretação dos padrões espaciais encontrados, permitindo identificar áreas atípicas em relação à agricultura e suas proporções nas glebas. Como pode ser observada nessa figura, parte da gleba era ocupada por vegetação florestal e foi identificada pelo algoritmo. Além disto, o algoritmo definiu áreas distintas dentro da mesma gleba, o que pode ser um indicativo de diferentes datas de plantio ou de cultivos distintos.

A Figura 4 representa os padrões dos vetores do codebook de cada neurônio, onde será realizada a avaliação espectro-temporal de cada padrão para auxiliar na tomada de decisão pelo analista de sensoriamento remoto. A figura permite verificar que existem diferentes padrões espectros-temporais, sendo o primeiro (azul escuro) típico de um cultivo anual e o último (amarelo) típico da área destacada como floresta na Figura 3.

A Figura 5 apresenta o resultado da análise em múltiplas glebas, onde a entrada do modelo são séries temporais de um conjunto de glebas. A figura exibe os padrões encontrados em todas as séries temporais, onde o eixo x representa o número de datas do período das séries temporais, e o eixo y corresponde à superfície de reflectância do NDVI, variando de -1 a 1. Nesta análise em múltiplas glebas, percebe-se também comportamentos similares ao observados anteriormente, com padrões variados de cultivos agrícolas, alguns mais típicos e outros menos ressaltados. Também permite verificar a ocorrência de áreas que não tiveram o cultivo agrícola.

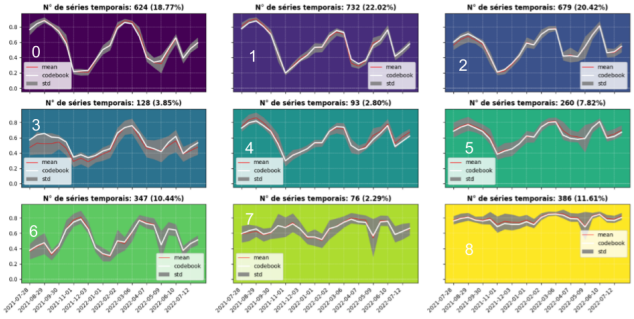

Figura 4: Grade do SOM da metodologia para análise individual de glebas contendo os 9 padrões dos vetores de peso

4. DISCUSSÕES

Avaliando a Figura 3, observa-se que a imagem à esquerda mostra as áreas da gleba com componentes florestais, enquanto a imagem à direita exibe os clusters sobrepostos a uma imagem de alta resolução, apenas para visualização. Essa imagem de alta resolução, utilizada unicamente para análise visual e não para a clusterização, permite avaliar a sobreposição espacial dos padrões. Nove clusters foram gerados, onde o cluster 8 (em amarelo) cobre majoritariamente áreas florestais, e o cluster 5 representa pixels mistos, onde parte do pixel pertence a áreas florestais e outra a áreas de soja, afetando o padrão final. Esses padrões são igualmente representados na Figura 4, onde o cluster 8 apresenta um padrão NDVI constante, típico de vegetação florestal. A análise da proporção deste padrão na gleba permite ao analista definir um limite de ocupação para padrões atípicos de agricultura. Estatísticas como desvio padrão e média das séries temporais, além do padrão do codebook, podem auxiliar na definição do tamanho ideal da grade para maximizar a separabilidade dos padrões.

Na análise em múltiplas glebas, conforme observado na Figura 5, o agrupamento utiliza uma grade 10x10 (100 neurônios), na qual o analista identifica padrões atípicos de agricultura. Cada unidade da grade possui uma cor de fundo, com um número vermelho no canto central esquerdo, representando o agrupamento hierárquico final, baseado na menor distância DTW. Esse método permite identificar rapidamente padrões atípicos de agricultura usando NDVI, um índice amplamente adotada na análise de vegetação. O grupo 34, por exemplo, situado no canto inferior esquerdo com cor verde-limão, apresenta um NDVI constante próximo de zero, indicando ausência de vegetação e, portanto, um padrão não típico de agricultura. O cluster 0, em azul-marinho, possui um NDVI elevado e constante, sugerindo uma formação florestal.

Esses clusters representam grupos de séries temporais que podem incluir múltiplas glebas, permitindo ao analista identificar aquelas com padrões específicos, demonstrando

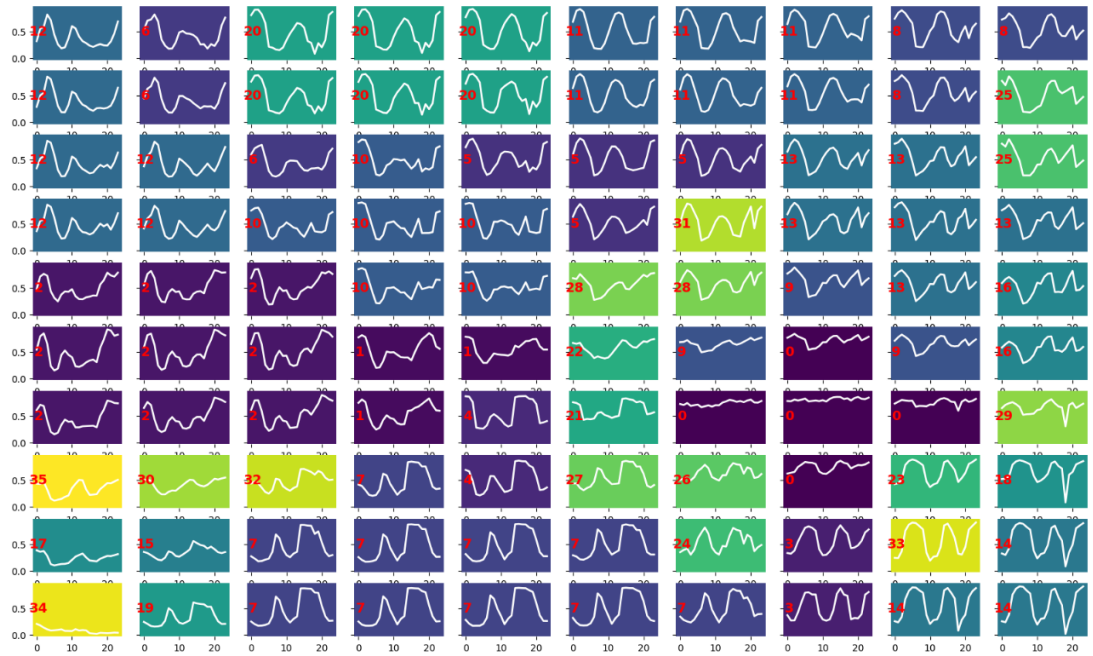

Figura 5: Grade do SOM 10x10 com as cores representado os grupos finais definidos pelo agrupamento hierárquico da metodologia para múltiplas glebas

o grande potencial deste método para análise exploratória e detecção de padrões atípicos de agricultura em grandes volumes de dados.

5. CONCLUSÕES

O presente trabalho introduziu duas técnicas de suporte para analistas de sensoriamento remoto, utilizando técnicas de agrupamento para identificação de padrões atípicos na agricultura. Essas técnicas funcionam como um recurso para a identificação de possíveis fraudes, permitindo a conferência por meio do sensoriamento remoto.

A primeira técnica identifica, em um espaço restrito, as glebas, geralmente necessitando de uma quantidade menor de grupos e tendendo a formar grupos mais homogêneos de agricultura, dependendo dos diferentes manejos da gleba. Vale ressaltar que as glebas são identificadas com uma única cultura por período para o pedido do seguro agrícola. A segunda técnica apresentou uma metodologia para identificar padrões atípicos para cultivos de culturas anuais em múltiplas glebas, mas com potencial para identificar, por exemplo, fraudes relacionadas ao plantio de diferentes tipos de culturas ou plantagens em períodos inadequados.

Como trabalho futuro, propõe-se aumentar a escala do estudo para a identificação em um grande volume de dados, visando um monitoramento rápido e eficiente.

Agradecimentos

Gostaríamos de agradecer à CAPES (Código de Financiamento 001, pela concessão de bolsa de Doutorado de Silva, B.L.C); à FAPESP (projeto 2023/09118-6) e ao CNPq (projeto 302205/2023-3 e projeto 302517/2023-5).

6. REFERÊNCIAS

[1] Banco Central do Brasil. *Manual de Crédito Rural (MCR)*, 2023. Disponível em: `https://www3.bcb.gov.br/mcr`.

[2] Ministério da Agricultura, Pecuária e Abastecimento. *Manual do Gestor ZARC (v1.1)*, 2020. Programa Nacional de Zoneamento Agrícola de Risco Climático.

[3] B Little, M Schucking, B Gartrell, B Chen, S Olson, K Ross, C Jenkerson, and R KcKellip. Remote sensing and us crop insurance program integrity: Data mining satellite and agricultural data. *WIT Transactions on Information and Communication Technologies*, 38, 2007.

[4] B. L. C. Silva, F. C. Souza, K. R. Ferreira, G. R. Queiroz, and L. A. Santos. Spatiotemporal segmentation of satellite image time series using self-organizing map. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2022:255–261, 2022.

[5] Jörn Debener, Volker Heinke, and Johannes Kriebel. Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*, 90(3):743–768, 2023.

[6] Alexander Vosseler. Unsupervised insurance fraud prediction based on anomaly detector ensembles. *Risks*, 10(7):132, 2022.

[7] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[8] Joydeep Ghosh and Abhishek Acharya. Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305–315, 2011.

[9] Brian Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Hierarchical Clustering*, chapter 4. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, USA, 5th edition, 2011.

[10] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[11] Karine R. Ferreira, Gilberto R. Queiroz, Lubia Vinhas, Rennan F. B. Marujo, Rolf E. O. Simoes, and et al. Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24), 2020.