USO DE SÉRIES TEMPORAIS E TÉCNICAS DE MACHINE LEARNING PARA CLASSIFICAR ÁREAS AGRÍCOLAS EM PETROLINA-PE

Pedro V. da Silva Brito[1], *Herica F. de Sousa Carvalho*[2], *Michel E. D. Chaves*[3], *Rafael D. Coelho dos Santos*[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE), Av. dos Astronautas, 1758 - São José dos Campos-SP, Brasil, (pedro.brito; rafael.santos)@inpe.br

[2]Centro de Tecnologia da Informação - Renato Archer (CTI) - Campinas-SP, Brasil (hfscarvalho@cti.gov.br)

[3]Universidade Estadual Paulista - Júlio de Mesquita Filho (UNESP) - Câmpus de Tupã (michel.dantas@unesp.br)

RESUMO

O objetivo deste estudo foi avaliar o uso de bandas de imagens de satélite para mapear áreas agrícolas em Petrolina-PE, utilizando séries temporais e algoritmos de machine learning. Foram comparados os modelos Random Forest (RF) e Temporal Convolutional Neural Network (TempCNN) na identificação de classes de uso e cobertura da terra com ênfase em cultivos agrícolas de importância econômica para a região. Imagens Sentinel-2 foram organizadas em cubos de dados, e amostras coletadas para treinar os classificadores e validar os mapas. Ambos os modelos distinguiram bem as classes, com o RF destacando-se em algumas classes específicas, e o TempCNN alcançando maior acurácia global, mostrando-se robusto para mapeamento em larga escala de áreas agrícolas.

Palavras-chave – Machine Learning, Cubos de dados, Cultivos agrícolas.

ABSTRACT

The objective of this study was to evaluate the use of satellite image bands to map agricultural areas in Petrolina-PE, utilizing time series and machine learning algorithms. The Random Forest (RF) and Temporal Convolutional Neural Network (TempCNN) models were compared in identifying land use and cover classes, with an emphasis on agricultural crops that are economically important to the region. Sentinel-2 images were organized into data cubes, and samples were collected to train the classifiers and validate the maps. Both models performed well in distinguishing the classes, with RF excelling in some specific classes, while TempCNN achieved higher overall accuracy, demonstrating robustness for large-scale mapping of agricultural areas.

Key words – *Machine Learning, Data cubes, Agricultural crops.*

1. INTRODUÇÃO

As mudanças no uso e cobertura da terra, como a conversão de áreas naturais em cultivos agrícolas, são um dos fatores que impulsionam as mudanças climáticas [1]. No Brasil, a fruticultura ocupa parte das terras agrícolas, tornando o país o terceiro maior produtor mundial de frutas, considerando a produção para consumo direto e processamento [2]. O desenvolvimento da agricultura irrigada no Semiárido, especificamente no Vale do São Francisco, nas cidades de Juazeiro - BA e Petrolina - PE, vêm promovendo importantes transformações socioeconômicas devido às atividades do agronegócio [3].

Com o desenvolvimento agrícola nesses municípios, torna-se essencial a criação de mapas de Uso e Cobertura da Terra (em inglês, *Land Use and Land Cover*, LULC) atualizados e precisos para ajudar na criação de políticas públicas e incentivos à gestão sustentável. Esses mapas são fundamentais para estudos como estimativas de rendimento e gestão de recursos hídricos. Imagens de satélite organizadas como cubos de dados de observação da Terra (em inglês, *Earth Observation*, EO) permitem o acompanhamento temporal de áreas específicas, facilitando a extração de séries temporais (em inglês, *Satellite Image Time Series*, SITS) e o uso de técnicas de aprendizado de máquina. No Brasil, a iniciativa é do Brazil Data Cube (BDC) que oferece suporte ao mapeamento de LULC por meio de análises de SITS e processamento de dados de observação da Terra [4].

Uma das implementações de software disponibilizadas e gerenciadas pelo BDC para processamento de cubos de dados é o pacote `SITS`, desenvolvido na linguagem R de código aberto [5]. Este pacote oferece um conjunto de funções para análise e classificação de SITS a partir de métodos de aprendizado de máquina e profundo (em inglês, *Machine Learning e Deep Learning*, ML e DL), utilizando as abordagens *Time-First* e *Space-Later* [5, 6]. Entre os modelos disponíveis neste pacote estão o *Random Forest* (RF) [7] e *Temporal Convolutional Neural Network* (TempCNN) [8]. Ambos conhecidos por serem métodos robustos na classificação de LULC.

O uso de modelos de ML e DL para classificação de imagens de satélite possibilita a extração de características complexas e padrões de grandes conjuntos de dados espaciais. Usando estas tecnologias, este estudo tem como objetivo avaliar o uso de bandas de imagens de satélite para mapear áreas de cultivos agrícolas no município de Petrolina - PE, Brasil, utilizando séries temporais e algoritmos de ML.

2. MATERIAL E MÉTODOS

2.1. Localização e caracterização da área de estudo

O estudo foi realizado no município de Petrolina (7°37'35" 9°43'56" latitude sul e 41°12'11" 39°7'27" latitude oeste) (Figura 1). Este município é reconhecido como o principal polo de fruticultura irrigada do Semiárido brasileiro pelo fácil acesso as água do rio São Francisco. O clima é classificado como BSw'h'- Semiárido quente e seco, de

acordo com a classificação de Köppen-Geiger [9]. Devido à sua localização geográfica e à influência da Zona de Convergência Intertropical (ZCIT), o padrão de precipitação é irregular durante o ano, com maior volume ocorrendo de fevereiro a maio [10]. O município abriga extensas áreas agrícolas irrigadas voltadas para a produção de frutas, como manga (43%), uva (31%), coco (9%), goiaba (5%), banana e acerola (ambos 4%). O sucesso dessas culturas é atribuído à irrigação, às condições edafoclimáticas e ao uso de tecnologias agrícolas avançadas [11].

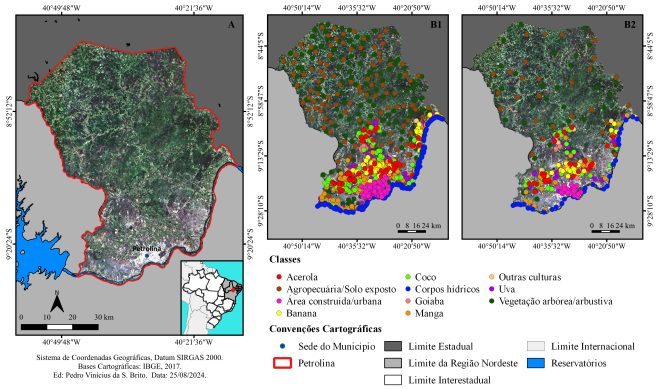

Figura 1: (A) Localização geográfica do município, (B1) Amostras de treinamento e (B2) Amostras de validação.

2.2. Coleta e processamento dos dados

2.2.1. Amostras de uso e cobertura

Coletou-se amostras de LULC para treinamento e validação por meio de interpretação visual de imagens Sentinel-2/MSI e séries temporais de Índice de Vegetação por Diferença Normalizada (em inglês *Normalized Difference Vegetation Index*, NDVI), utilizando a plataforma *TerraCollect*, desenvolvida no âmbito do projeto BDC [4]. Para treinar os classificadores foram coletadas 1320 amostras (120 por classe) com longitude, latitude, data inicial, data final e classe (Figura 1 B1). As classes foram Acerola, Agropecuária/Solo exposto, Área construída/urbana, Banana, Coco, Corpos hídricos, Goiaba, Manga, Outras culturas, Uva e Vegetação arbórea/arbustiva. Para validar os mapas classificados coletou-se 330 amostras (30 por classe) (Figura 1 B2).

2.2.2. Dados de satélite

Os dados de sensoriamento remoto incluem cubos de dados anuais de bandas geradas pelo satélite Sentinel-2/MSI, coleção BDC S2 SEN2COR_10_16D_STK-1. Esses dados possuem imagens com resolução espacial de 10 metros projetadas e recortadas na grade BDC considerando função de composição temporal de 16 dias, usando a abordagem de empilhamento de imagens com menos nuvem de um período temporal. Os dados foram baixados gratuitamente do servidor BDC [4] abrangendo o período de 01/10/2019 a 30/09/2020. No total, 8 bandas foram utilizadas, dentre elas: B02 (blue), B03 (green), B04 (red), B08 (nir), B11 (swir16), B12 (Swir22), B8A (Nir08), e SCL (cloud), contendo 23 imagens por ano.

2.2.3. Avaliação de amostras de uso e cobertura

Foram extraídas as séries temporais das amostras a partir dos cubos de dados. As amostras foram filtradas pelo método proposto por Santos et al. [12], que usa mapas auto-organizáveis (em inglês, *Self-Organizing Maps*, SOM) [13]. Nesta etapa, foram gerados testes para cada um dos experimentos variando o tamanho da grade de neurônios para o conjunto amostral. Como base na escolha dos testes, inicialmente adotou-se os parâmetros indicados por Santos et al. [12], que foram $5\frac{\sqrt{N}}{2}$ neurônios, em que, N é o número de amostras, taxa de aprendizado decrescente, de $0,50$ a $0,01$, e distância euclidiana. Para maior amplitude entre as confusões das amostras nos agrupamentos SOM, os tamanhos de grade analisadas foram 9×9 à 18×18 e 200 interações.

A validação cruzada *k-fold*, foi aplicada para selecionar o melhor modelo RF diferenciando-se pelo número de árvores. Para o modelos TempCNN foi aplicado o *model tuning*. Após a seleção dos modelos RF e TempCNN mais precisos, aplicou-se o *F1-score* para avaliar a precisão das classes de uso e cobertura da Terra. Para validar o modelo TempCNN foi utilizado 20% das amostras coletadas para treinamento, não usadas para treinar o modelo.

2.2.4. Classificação e avaliação dos mapas

Os modelos RF e TempCNN selecionados foram treinados, as imagens classificadas, e por fim, refinadas por suavização *Bayesiana*. A precisão dos mapas classificados foi obtida utilizando a técnica de área ponderada, de acordo com as melhores práticas de Olofsson et al. [14]. Qualitativamente os mapas foram avaliados e comparados por interpretação visual a conhecimento sobre a região. As etapas de classificação e avaliação foram todas realizadas a partir do pacote `SITS`.

3. RESULTADOS E DISCUSSÃO

3.1. Padrões espectro-temporais

Os padrões espectro-temporais de cada banda das classes de LULC são apresentados na Figura 2. É possível identificar padrões sazonais específicos para cada classe, evidenciando a variação temporal dos valores das bandas em função do tipo de uso do solo. Será dada ênfase às classes de culturas (Acerola, Banana, Coco, Goiaba, Manga, Outras culturas e Uva), dado o papel econômico e agrícola significativo dessas culturas no município de Petrolina-PE. As demais classes, serão incluídas para oferecer uma visão geral do uso e cobertura da terra.

Nas classes referente as culturas as bandas B08 e B8A apresentaram os maiores valores espectrais, indicando a presença de vegetação saudável e vigorosa ao longo do ano. Observa-se uma redução nos valores espectrais entre os meses de abril e julho, seguido por aumento entre julho e outubro, sendo essa variação mais evidente nas classes Acerola, Goiaba, Manga, Uva e Outras culturas. Esse comportamento pode estar associado a prática de manejo cultural associado a cada uma delas, como por exemplo, poda e colheita.

A classe Banana apresentou valores espectrais consistentes ao longo do ano, especialmente nas bandas B08 e B8A,

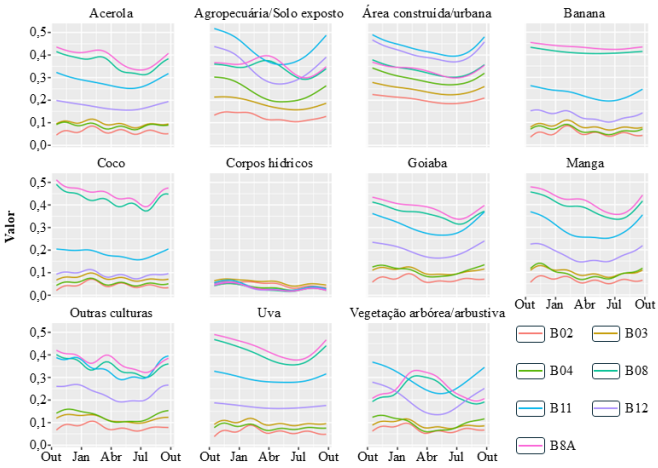

Figura 2: Padrões espectro-temporais das bandas via GAM.

assim como a classe Coco, porém, com maior variação de valores nas bandas. A classe de Outras culturas exibiu maior variabilidade espectral, refletindo a diversidade de culturas e diferentes estágios fenológicos.

3.2. Avaliação das amostras

A Figura 3A apresenta o agrupamento obtido via SOM. Observa-se a presença de 39 neurônios *outliers*. Esses neurônios correspondem a casos em que a classe predominante difere das classes da vizinhança [12]. A presença de neurônios *outliers* não afirma erro de rotulagem de amostras, mas pode representar amostras com diferentes padrões de classes de LULC no tempo ou espaço, ou inseparáveis utilizando SITS [12].

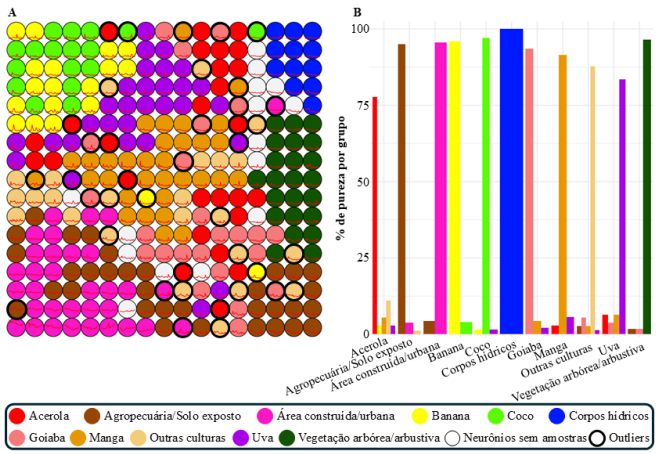

Figura 3: Agrupamento das séries temporais e porcentagem de confusão entre os grupos do SOM com grade 17x17.

O agrupamento das amostras resultou em 15 neurônios sem amostras (Figura 3A). No SOM, cada neurônio é rotulado com a classe predominante das amostras a ele associadas. Em alguns casos, nenhum neurônio específico possui amostras associadas, sendo assim, rotulados como "neurônios sem amostras" [15]. Isso indica a ausência de exemplos suficientemente representativos de certas classes nos dados para serem atribuídos a esses neurônios específicos.

A principal razão para os melhores agrupamentos nas classes Corpos hídricos e Vegetação arbórea/arbustiva pode ser atribuída ao comportamento característico de suas assinaturas espectrais, facilitando ao modelo distingui-las. Resultados semelhantes foram encontrados por Brito et al. [16] ao utilizar apenas séries temporais de NDVI. No entanto, observa-se maiores confusões entre os neurônios das classes Acerola, Agropecuária/Solo exposto, Banana, Goiaba e Outras culturas.

Na Figura 3B, são exibidos os resultados do percentual de pureza por grupo após a remoção das amostras consideradas ruidosas por meio da aplicação do SOM. Foram identificados agrupamentos com grau de pureza igual ou superior a $77,77\%$ (Acerola). A classe Corpos hídricos alcançou o maior percentual, com 100% de pureza.

3.3. Ajuste dos modelos

Após a avaliação do percentual de confusão, os resultados de acurácia de acordo com o número de árvores obtidos por meio da validação *k-fold* ($k = 5$) no conjunto de treinamento utilizando o modelo RF foram: $88,3\%$; $88,4\%$; $87,9\%$; $90,3\%$ e $88,4\%$ nos respectivos números de árvores 100, 500, 1000, 1500 e 2000, enquanto utilizando o modelo TempCNN foi $89,2\%$. Ambos os modelos apresentaram precisão igual ou superior a $87,9\%$ (RF com 1000 árvores). A maior precisão também foi obtida com o modelo RF ao utilizar 1500 árvores ($90,3\%$).

Após a avaliação utilizando *k-fold*, são apresentados os percentuais de acurácia para cada classe com base na métrica *F1-score* para os modelos RF e TempCNN, respectivamente: Acerola ($82,2\%$, 85%); Agropecuária/Solo exposto (100%, $97,8\%$); Área construída/urbana ($93,2\%$, 91%); Banana ($97,8\%$, $95,4\%$); Coco ($85,7\%$, 100%); Corpos hídricos ($96,9\%$, 69%); Goiaba ($78,3\%$, $77,3\%$); Manga ($71,9\%$, $77,5\%$); Outras culturas (84%, 83%); Uva ($86,8\%$, $76,9\%$) e Vegetação arbórea/arbustiva ($91,2\%$, 92%). Observa-se que, os menores percentuais de precisão ocorreram nas classes Corpos hídricos e Manga, para os modelos TempCNN e RF, com 69% e $71,9\%$, respectivamente. Por outro lado, Agropecuária/Solo exposto e Coco apresentaram 100% de precisão nos modelos RF e TempCNN, respectivamente.

3.4. Classificação dos mapas de LULC

A seguir são apresentados os mapas classificados de LULC (Figura 4). De modo geral, as classes Corpos hídricos e Vegetação arbórea/arbustiva foram melhor identificadas com ambos os modelos. Resultados semelhantes foram encontrados na mesma área de estudo ao utilizar o modelo RF e NDVI [16].

As áreas de Uva e Manga foram melhor identificadas com o modelo RF (regiões a e b, respectivamente), o mesmo foi observado para a classe Coco (região c). A classe Banana (região d) foi reconhecida em ambos os modelos, porém, apresentou bastante confusão, em especial, com a classe Coco. Na classificação usando o modelo TempCNN, observou-se um aumento de áreas de Uva classificadas erroneamente como Acerola, Goiaba e Outras culturas. Em ambos os mapas, são evidentes diferenças pontuais nas delimitações das classes de LULC, bem como em alvos específicos como culturas agrícolas, corpos de água

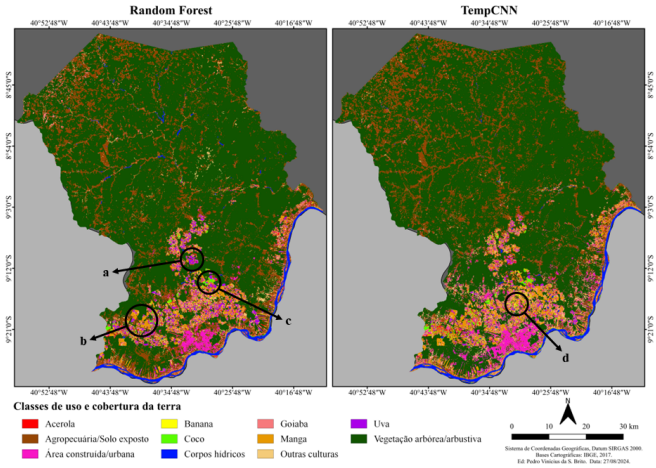

Figura 4: Classificações de uso e cobertura da Terra para Petrolina-PE, utilizando os modelos RF e TempCNN.

e vegetação. Resultados equivalentes foram observados por [8].

3.5. Avaliação dos mapas

Usando a técnica de área ponderada, observa-se que a classe Outras culturas apresentou 0% nas métricas de acurácia do produtor e usuário (Tabela 1). Isso pode ter ocorrido devido à dificuldade dos modelos em identificar essa classe, possivelmente em função do seu tamanho reduzido. Com exceção da classe anterior, Banana foi a que obteve menor percentual (17% e 18% para o RF e TempCNN, respectivamente, na acurácia do produtor). O modelo TempCNN obteve maior valor de acurácia global, com diferença de 1% do RF. Esse resultado corrobora com os encontrados por [8].

Tabela 1: Valores de acurácia das imagens classificadas.

	AP(%)		AU(%)		AG(%)	
Classes	**RF**	**TC**	**RF**	**TC**	**RF**	**TC**
Acerola	21	25	65	79	93	94
Agropecuária/Solo exposto	100	98	85	100		
Área construída/urbana	100	100	94	91		
Banana	17	18	62	61		
Coco	92	93	74	69		
Corpos hídricos	100	100	100	100		
Goiaba	42	42	73	58		
Manga	62	88	85	64		
Outras culturas	0	0	0	0		
Uva	54	51	54	76		
Vegetação arbórea/arbustiva	100	100	100	100		

RF = *Random forest*, TC = TempCNN, AP = Acurácia do produtor; AU = Acurácia do usuário, AG = Acurácia global.

4. CONCLUSÕES

Ambos os modelos foram capazes de identificar as classes de interesse, com o modelo RF apresentando maior valor de acurácia na identificação de classes específicas, como Uva e Manga. Por outro lado, o TempCNN obteve uma acurácia global maior, mostrando-se uma alternativa robusta para mapeamento em larga escala de áreas agrícolas.

5. AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao projeto Brazil Data Cube (BDC).

6. REFERÊNCIAS

[1] C. Liao, K. Nolte, D. G. Brown, J. Lay, and A. Agrawal. The carbon cost of agricultural production in the global land rush. *Global Environmental Change*, 80:102679, 2023.

[2] B. B. Maggi, E. R. Novacki, L. R. Barcelos, R. da C. Cavalcanti Junior, and F. J. de P. C. Fonseca. Plano nacional de desenvolvimento da fruticultura, 2018.

[3] R. R. D. Ramos and J. C. F. de Melo Júnior. Mapping of the current land use in part of the irrigated perimeter nilo coelho, petrolina-pe, brazil. *Comunicata Scientiae*, 10:89–97, 2019.

[4] K. Ferreira and et al. Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12:4033, 2020.

[5] R. Simoes, , and et al. Satellite image time series analysis for big earth observation data. *Remote Sensing*, 13:2428, 2021.

[6] G. Camara and et al. Big earth observation data analytics: matching requirements to system architectures. In *Proceedings...*, pages 1–6, San Francisco, 2016. ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON ANALYTICS FOR BIG GEOSPATIAL DATA.

[7] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[8] C. Pelletier, G. I. Webb, and F. Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11:523, 2019.

[9] C. A. Alvares and et al. Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, 22:711–728, 2013.

[10] J. A. Marengo, R. R. Torres, and L. M. Alves. Drought in northeast brazil—past, present, and future. *Theoretical and Applied Climatology*, 129:1189–1200, 2017.

[11] DINC. Distrito de irrigação senador nilo coelho, 2022. Disponível em: <https://www.dinc.org.br/>. Acesso em: 25/08/2024.

[12] L. Santos and et al. Quality control and class noise reduction of satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:75–88, 2021.

[13] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78:1464–1480, 1990.

[14] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148:42–57, 2014.

[15] Lorena Santos and et al. Self-organizing maps in earth observation data cubes analysis. In *International Workshop, WSOM+ 2019, Barcelona*, pages 70–79, 2020.

[16] P. Brito and et al. Uso de séries temporais para classificações de uso e cobertura da terra em Petrolina, Pernambuco. In *Anais...*, pages 2530–2533, Florianópolis, 2023. Simpósio Brasileiro de Sensoriamento Remoto, SBSR.