CLASSIFICAÇÃO SUPERVISIONADA DE IMAGENS SENTINEL 2 COM MACHINE LEARNING NO GOOGLE EARTH ENGINE

Clovis Cechim Junior [1], Hideo Araki[2], Rodrigo Macedo[3], Heloise Milena Dambrat[4]

[1,4] Sistema de Tecnologia e Monitoramento Ambiental do Paraná (SIMEPAR),
clovis.cechim@simepar.br[1]; heloise.dambrat@simepar.br[2]
[2,3] Universidade Federal do Paraná (UFPR),
haraki@ufpr.br[2]; rodrigo.macedo@ufpr.br[3]

RESUMO

O objetivo deste trabalho foi avaliar o uso do Mapeamento de uso do solo realizado pelo MapBiomas, como conjunto amostral de treinamento para classificação supervisionada de Imagens Sentinel 2. Para a classificação foram utilizados os classificadores *Classification and Regression Trees* (CART), *Random Forest* (RF) e *Gradient Tree Boost* (GTB), implementados no *Google Earth Engine* (GEE). A região de estudo usada na classificação foi a região fitogeográfica Savana, localizada no estado do Paraná. A classificação por RF obteve a maior acurácia espacial com um índice Kappa de 0,94 e exatidão global de 96 %.

Palavras-chave **—** sensoriamento remoto, uso do solo, *Random Forest*.

Abstract

The objective of this work was to evaluate the use of the Land Use Mapping carried out by MapBiomas, as a training sample set for supervised classification of Sentinel 2 Images. for the classification, the classifiers used were Classification and Regression Trees (CART), Random Forest (RF) and Gradient Tree Boost (GTB), implemented in Google Earth Engine (GEE). The study region used in the classification was the Savana phytogeographic region, located in the state of Paraná. The RF classification obtained the highest spatial accuracy with a kappa index of 0.94 and global accuracy of 96%.

Keywords **—** *remote sensing, land use, Random Forest.*

1. INTRODUÇÃO

O Aprendizado de Máquina, ou *Machine Learning* (ML), é um campo de estudo da ciência da computação que fornece a capacidade de aprender ou prever dados usando métodos computacionais [3]. O aprendizado de máquina tem sido usado extensivamente em diversas aplicações, como biologia, visão computacional, economia e sensoriamento remoto [8]; [9]; [10]. Com os avanços recentes em *hardware*, técnicas, habilidades de otimização e coleta de dados, o aprendizado profundo, *Deep Learning* (DL), que está enraizado na teoria da rede neural artificial, tornou-se recentemente uma importante área de foco na comunidade de aprendizado de máquina (ML) devido ao seu potencial para um melhor aprendizado com representações de dados usando várias camadas [7].

De acordo com Hird *et al.,* (2017) [6], os avanços modernos na computação em nuvem e nos algoritmos de uso de máquina estão mudando a maneira como os dados de observação da Terra são usados para monitoramento ambiental, principalmente na era dos fluxos de dados de satélite de acesso aberto e gratuitos. Na década de 2000, surgiram soluções para os problemas sobre ajuste e alta demanda computacional típicos das redes neurais, permitindo o desenvolvimento de redes neurais muito maiores e mais profundas. Com isso, DL se tornou um subcampo de rápido crescimento do aprendizado de máquina. Nas geociências, a aquisição de grandes volumes de dados de sensoriamento remoto está se acelerando devido à proliferação de técnicas e fontes de sensoriamento [3].

2. MATERIAL E MÉTODOS

A área de estudo corresponde à região fitogeográfica Savana, localizada no estado do Paraná, localizada entre as coordenadas 25°00'00"S e 50°00'00"O (Figura 1).

A metodologia pode ser dividida em 4 etapas, sendo:

(I) Seleção de imagens Sentinel 2;
(II) Geração de amostras de treinamento;
(III) Classificação supervisionada usando *Machine Learning* no GEE;
(IV) Validação de desempenho (acurácia espacial).

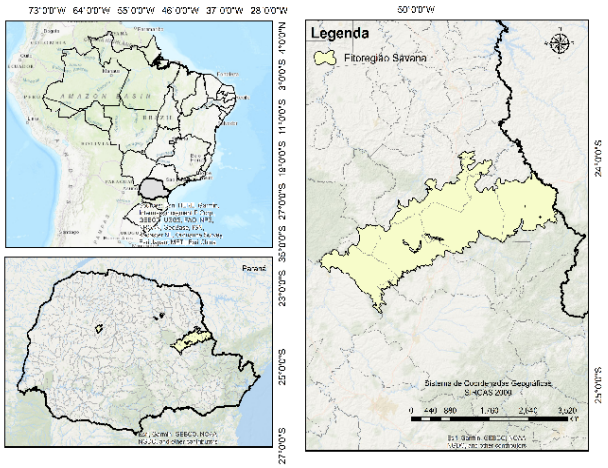

Figura 1. Região Fitogeográfica Savana, estado do Paraná.

A seleção de imagens Sentinel 2 (European Union/ESA/Copernicus, 2021), se deu a partir de períodos de entressafra agrícola, período compreendido entre agosto e setembro de 2016, sendo gerado uma composição de mediana, e selecionando imagens com menor incidência de nuvens, procedimento feito no GEE (Figura 2).

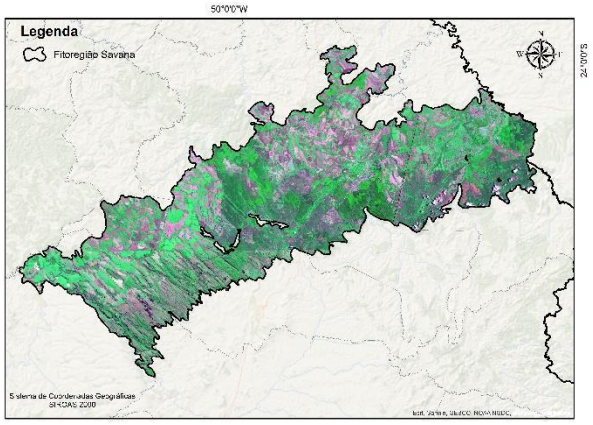

Figura 2. Imagem Sentinel 2, ano de 2016, RGB-382.

Imagens selecionadas em períodos de entressafra facilitam o procedimento de classificação, pois áreas de solo exposto, ou cultivos com baixo vigor vegetativo, permitem uma melhor descriminação com outros alvos, como pastagem, floresta e reflorestamento.

Foi aplicado também uma máscara de nuvens, utilizando a banda máscara de nuvem do Sentinel 2 (QA60) de 60m de resolução espacial, para filtrar eventuais nuvens ou algum tipo de vapor de água, da composição mediana usada para classificação.

Foi gerado uma composição incluindo as bandas do visível (b2, b3, b4) e infravermelho (b8), com 10m de resolução espacial, e as bandas de borda do vermelho (b5, b6, b7) e a banda do infravermelho de ondas curtas (SWIR) (b8A) com resolução espacial de 20m, totalizando 8 bandas espectrais.

A segunda etapa consistiu na geração de amostras de treinamento, sendo utilizado o mapeamento produzido pelo MapBiomas para o ano de 2016, na resolução espacial de 30m. Deste mapeamento do MapBiomas foram gerados 11.000 pixels amostrais, para as seguintes classes de uso do solo: Formação Florestal; Floresta Plantada; Formação Campestre; Pastagem; Infraestrutura Urbana; Outra Área não vegetada; Rio, Lago e Oceano; Lavoura Perene; Soja; Outras Lavouras Temporárias.

A terceira etapa consistiu nos testes de classificação por ML, utilizando 3 classificadores distintos, *Classification and Regression Trees* (CART) (Breiman et al., (1984) [1]), *Random Forest* (RF) e *Gradient Tree Boost* (GTB). Para esta análise foram utilizados o mesmo número amostral.

Para os classificadores RF e GBT, foram utilizados como parâmetro de entrada 70 árvores de decisão a serem criadas. Para o classificador CART não foi definido nenhum limite para o parâmetro de entrada.

A quarta etapa consistiu na análise de acurácia espacial para a validação (acurácia), das classificações geradas foi usados o índice de exatidão global (EG), o índice de concordância Kappa (IK) e os erros de inclusão (EI) e Omissão (EO) [4]; [5].

Para a obtenção das métricas estatísticas de acurácia foi utilizado uma amostragem aleatória estratificada não proporcional com 100 pontos amostrais (Pixels) sorteados em toda área de estudo. Estes pontos amostrais utilizados na validação foram gerados de forma independente do conjunto amostral usado no treinamento dos classificadores. A imagem Sentinel 2 foi usada como referência para determinar as amostras para validação.

3. RESULTADOS E DISCUSSÃO

Para a classificação por RF obteve-se um IK de 0,94 com uma EG de 96.00% (Figura 3).

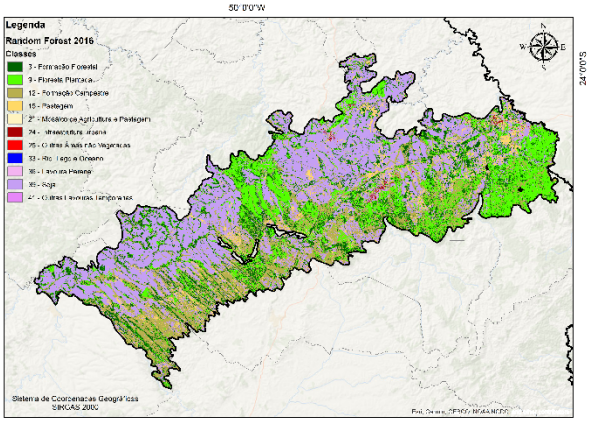

Figura 3. Classificação *Random Forest*.

Pode-se observar que em todos os mapeamentos a classe de água, utilizando os pixels amostrais do MapBiomas, foi omitida do mapeamento para alguns cursos de água com até 10 metros de largura, isto pode ser explicado em virtude da resolução espacial utilizada pelo MapBiomas que provém de imagens Landsat com 30 m de resolução espacial.

A classificação por GTB obteve um IK de 0,86 com um EG de 90,82% (Figura 4).

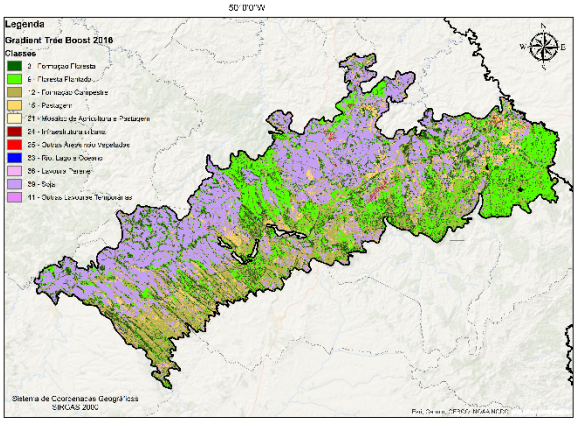

Figura 4. Classificação *Gradient Tree Boost.*

Para a classificação CART, obteve-se um IK de 0,73 com uma EG de 81,00% (Figura 5). Nesta classificação foi observado que houve uma maior confusão entre as classes de uso do solo.

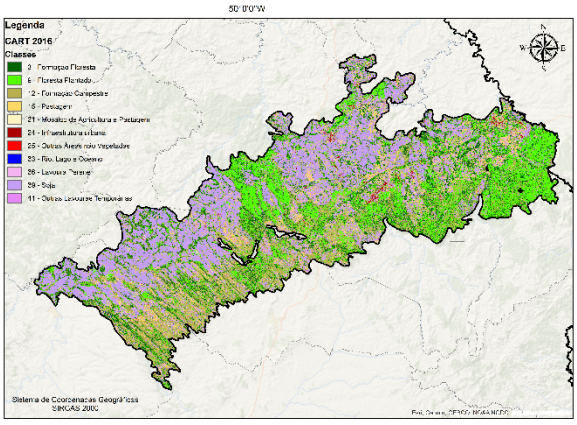

Figura 5. Classificação CART.

Foi observado que as maiores confusões foram entre as classes de Formação Florestal e Floresta Plantada, e entre áreas de Pastagens e Agricultura.

A tabela 1, apresenta os resultados das métricas estatísticas de avaliação de acurácia espacial.

	RF	GTB	CART
IK	0,94	0,86	0,73
EG (%)	96,00	90,82	81,00

Tabela 1. Acurácia espacial dos classificadores.

Os classificadores RF e GTB apresentaram maior semelhança entre as classes com uma melhor definição e suavização entre as classes dos mapeamentos (Figura 6-A e 6-B).

Pode-se observar que o classificador CART (Figura 6-C) demonstrou uma maior confusão espectral entre as classes de uso e cobertura do solo, o que pode ser constatado pela representação mais pixelada.

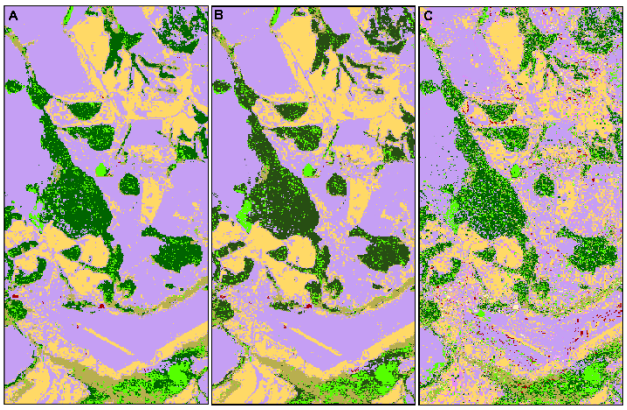

Figura 6. Comparação entre os classificadores RF (A), GT (B) e CART (C).

4. CONCLUSÕES

O mapeamento do MapBiomas na resolução espacial de 30 metros provenientes das imagens Landsat pode ser utilizado para gerar amostras de treinamento e validação.

Este conjunto amostral pode ser utilizado em ML para classificações de imagens Sentinel 2.

Todavia alguns alvos com menor representatividade nas região Fitogeográfica de Savana foram omitidos, como a classe de água, para alguns cursos d'água com largura inferior a 30m.

Dentre os classificadores utilizados, o classificador RF obteve o melhor resultado de acurácia na classificação de uso e cobertura do solo. Não foi avaliado o uso de nenhum tipo de filtro espacial, somente o algoritmo de classificação e sua performance utilizando um mesmo conjunto amostral.

5. REFERÊNCIAS

[1] Breiman, L. Friedman, J. Olshen, R. Stone, C. Classification and Regression Trees," Chapman and Hall, 1984.

[2] Chatziantoniou, A .; Psomiadis, E .; Petropoulos, G.P. Co-Orbital Sentinel 1 e 2 para mapeamento LULC com ênfase em áreas úmidas em um ambiente mediterrâneo baseado em aprendizado de máquina. Remote Sens. 2017, 9, 1259. https://doi.org/10.3390/rs9121259

[3] Chi, J.; Kim, H.-c. Prediction of Arctic Sea Ice Concentration Using a Fully Data Driven Deep Neural Network. Remote Sens. 2017, 9, 1305. https://doi.org/10.3390/rs9121305

[4] Congalton, R. G.; Green, K. Assessing the accuracy of remotely sensed data: principles and practices. Boca Raton: CRC Press, 1999. 160 p.

[5] Congalton, R. G. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, v.37, p.35-46, 1991.

[6] Hird, J.N.; DeLancey, E.R.; McDermid, G.J.; Kariyeva, J. Google Earth Engine, Open-Access Satellite Data, and Machine Learning in Support of Large-Area Probabilistic Wetland Mapping. *Remote Sens.* **2017**, *9*, 1315. https://doi.org/10.3390/rs9121315

[7] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015, 28;521(7553):436-44. doi: 10.1038/nature14539

[8] Mountrakis, G.; IM, J.; Ogole, C. Support vector machines in remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 2011, 66, 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

[9] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

[10] Praticò, S.; Solano, F.; Di Fazio, S.; Modica, G.. (2021). Machine Learning Classification of Mediterranean Forest Habitats in Google Earth Engine Based on Seasonal Sentinel-2 Time-Series and Input Image Composition Optimisation. Remote Sensing. 13. 586. https://doi.org/10.3390/rs13040586

[11] Sousa, C.; Fatoyinbo, L.; Neigh, C.; Boucka, F.; Angoue, V, Larsen, T. Cloud-computingand machine learning in support of country-levelland cover and ecosystem extent mapping in Liberia and Gabon. 2020. PLoS ONE 15 (1). https://doi.org/10.1371/journal.pone.0227438