CLASSIFICAÇÃO DIGITAL PARA IDENTIFICAÇÃO E MAPEAMENTO DE VEGETAÇÃO SECUNDÁRIA E PASTAGENS NO CERRADO

Adriane Calaboni[1], Lídia Sanches Bertolo[1], Júlio Cesar Dalla Mora Esquerdo[2], João Francisco Gonçalves Antunes,[2] Alexandre Camargo Coutinho[2]

[1]Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH., Av. José Rocha Bonfim, 214, Jardim Santa Genebra, Praça Capital, Ed. Frankfurt, Sala 227, CEP 13080-650, Campinas, SP, Brasil, {adriane.calaboni, lidia.bertolo}@giz.de; [2]Embrapa Agricultura Digital, Av. André Tosello, 209, CEP 13083-886, Campinas, SP, Brasil, {julio.esquerdo, joao.antunes, alex.coutinho}@embrapa.br.

RESUMO

A identificação e mapeamento de áreas de vegetação secundária e de pastagens, no Bioma Cerrado, é um grande desafio para as instituições que trabalham no monitoramento da cobertura e uso da terra no Brasil. Com o objetivo de promover avanços metodológicos para a automatização dos mapeamentos, foram executados processamentos de dados satelitários, por meio do pacote sits (*Satellite Image Time Series Analysis for Earth Observation Data Cubes*), para construir um modelo acurado. Amostras iniciais foram coletadas por interpretação visual e tratadas pelo método *Self-organizing Maps* (SOM). O modelo criado usando o algoritmo *Random Forest* obteve acurácia global de 96% e o mapeamento apresentou exatidão global de 89%. O classificador foi capaz de diferenciar e delimitar áreas de pastagem, áreas com predomínio de arbustos, com predomínio de indivíduos arbóreos e com estabelecimento de dossel, apesar do limite entre essas classes ser caracterizado por transições graduais da vegetação, não caracterizando bordas discretas.

Palavras-chave *— BDC, sits, aprendizado de máquina, SOM, Random Forest.*

ABSTRACT

The identification and mapping of secondary vegetation and pastureland in the Cerrado Biome (Brazilian savanna) is a great challenge to the institutions which monitor land use and land cover in Brazil. To promote methodological advances on automatic map classification of Cerrado Biome, we performed satellite image processing using Satellite Image Time Series Analysis for Earth Observation Data Cubes (sits) to build an accurate model. Reference samples were collected by visual interpretation and filtered by Self-organizing Maps (SOM) method. The model was build using the Random Forest algorithm and obtained 96% of accuracy while the map overall accuracy was 89%. As a result, the model identified and classified pasturelands and areas with the dominance shrubs or areas with dominance of trees and canopy well develop, although the limits between those classes are gradual transitions.

Key words *— BDC, sits, machine learning, SOM, Random Forest.*

1. INTRODUÇÃO

O Cerrado é o segundo maior bioma brasileiro em termos de extensão territorial, representando, ao mesmo tempo, uma das regiões de maior biodiversidade no mundo [1] e estratégica para o desenvolvimento do agronegócio brasileiro [2]. Por isso, o Brasil enfrenta hoje a difícil tarefa de aliar a conservação do Cerrado e ao mesmo tempo investir na expansão da agropecuária nacional. Uma informação estratégica, essencial, na busca por soluções de baixo impacto ambiental e de relevância estratégica para intensificação e expansão do agronegócio nacional é o mapeamento e discriminação das áreas de pastagens e de vegetação secundária, como subsídio para o entendimento da dinâmica de desmatamento e regeneração do bioma. No entanto, a classificação automática das áreas de vegetação secundária arbórea, arbustiva e a tipificação das pastagens no Cerrado é um grande desafio, já que constituem um gradiente natural, cujos limites entre essas classes temáticas não são bem definidos. Parte do enfrentamento deste desafio está relacionada à obtenção de conjuntos de amostras com alta qualidade para a parametrização de modelos de classificação baseados em aprendizado de máquina. Estas amostras precisam ter alto poder discriminatório entre as classes para proporcionar uma boa performance do classificador.

O *Satellite Image Time Series Analysis for Earth Observation Data Cubes* (*sits*) é um pacote, desenvolvido em linguagem R, com funções para avaliação e melhoria da qualidade de amostras, parametrização de modelos e classificação da cobertura e uso da terra por meio de séries temporais de imagens de satélite [3]. A utilização das séries temporais de imagens de satélite permite que as classes de vegetação sejam discriminadas com base no comportamento espectro-temporal expresso por suas fenologias, aumentando a capacidade de diferenciação de alvos quando comparado aos métodos tradicionais, baseados em uma ou poucas imagens de períodos específicos [3]. Com base nisso, por meio do *sits*, buscamos construir um classificador capaz de

identificar e mapear áreas de vegetação secundária arbórea (VS arbórea), vegetação secundária arbustiva (VS arbustiva) e pastagem (PA) no Bioma Cerrado.

2. MATERIAL E MÉTODOS

Foram coletadas, por meio de interpretação visual, 270 amostras em pixels das classes VS arbórea, VS arbustiva e PA, totalizando 810 amostras distribuídas em uma grade 3x3 de cenas Landsat-8/OLI no estado de Mato Grosso do Sul, onde uma região na cena central foi posteriormente classificada (tile BDC 084105). Posteriormente, foram coletadas 101 amostras em pixels representativos de corpos d'água, para promover a melhoria do potencial discriminatório do modelo para as classes de interesse. Contudo, essa classe não foi avaliada por não fazer parte do objetivo do estudo (Figura 1).

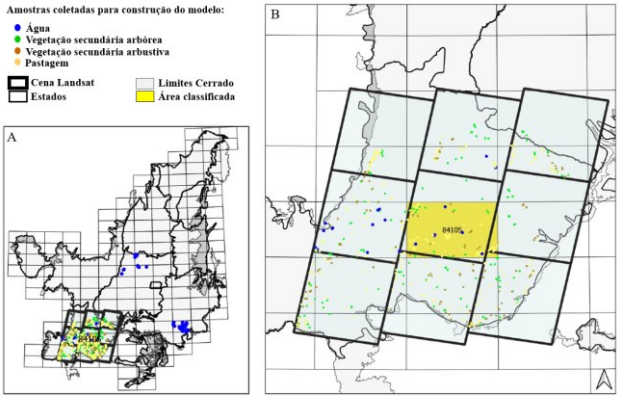

Figura 1. Distribuição das amostras coletadas manualmente em uma grade 3x3 de cenas Landsat no estado de Mato Grosso do Sul (A) e a área classificada, em amarelo (tile BDC/Sentinel-2), localizada na cena central da grade (B).

Utilizando como referência imagens Sentinel-2/MSI, Planet Scope e do Google Earth, as amostras foram coletadas observando-se o contexto da área e o pixel da imagem Sentinel-2 (composição falsa cor; bandas B8, B4 e B3) para determinação da classe. Foram denominadas como PA áreas com predomínio de gramíneas naturais ou exóticas. Áreas com predomínio de arbustos e poucos indivíduos arbóreos foram denominadas como VS arbustiva e como VS arbórea foram consideradas aquelas áreas com predomínio de indivíduos arbóreos e com a formação de dossel.

A avaliação da qualidade das amostras foi feita pelo método SOM (*Self-Organizing Map*) [4] utilizando a função *sits_som_map* do pacote *sits*, que utiliza redes neurais não supervisionadas para reconhecimento de padrões em séries temporais de dados espaciais extraídos de cubos de dados e inferência bayesiana para filtragem das amostras [3]. Por meio da função *sits_get_data* do pacote *sits,* as séries temporais de cada amostra foram extraídas de cubos de dados Sentinel-2 disponibilizados pelo *Brazil Data Cube* (BDC) [5]. Os cubos foram constituídos de composições de 16 dias obtidas ao longo do ano agrícola 19/20 (28/07/2019 a 28/08/2020), contendo as bandas espectrais B02, B03, B04, B05, B06, B07, B8A, B11, B12, os índices de vegetação EVI e NDVI e, também, o produto de mascaramento de nuvens CLOUD. Para a classificação foi aplicado o algoritmo *Random Forest* e um filtro Bayesiano pós-classificação, ambos usando parâmetros padrões do pacote *sits* sendo 200 árvores de decisão e janela 5x5, respectivamente. A escolha do algoritmo baseou-se em um balanço entre performance do classificador e tempo de processamento que são condições importantes para mapeamentos de áreas muito extensas. O mapa gerado foi validado por meio da coleta e classificação de 50 pontos aleatórios por classe (N=150).

3. RESULTADOS E DISCUSSÃO

As 810 amostras foram processadas pelo SOM e 155 (38 de VS arbórea, 74 de VS arbustiva e 43 de PA) foram consideradas *noisy samples* pelo método para serem analisadas posteriormente. Destas 155 amostras, 20 de VS arbórea e 44 de PA foram removidas manualmente por serem amostras coletadas em áreas de transição entre classes, restando, portanto, 655 amostras. O modelo gerado com base nessas amostras (N=655) obteve acurácia global igual à 0,957 e IC95% = 0,016. No entanto, o mapa criado pelo modelo apresentou áreas de inclusão de VS arbustiva em PA, geralmente áreas úmidas, e inclusão de VS arbórea em pixels representativos de corpos d'água. Desse modo, foram adicionadas ao modelo amostras coletadas manualmente em áreas de inclusão de VS arbustiva em PA (N=77) e amostras coletadas em pixels representativos de rios, lagos e reservatórios (N=101), embora a classe corpo d'água não tenha sido avaliada, como dito anteriormente. O novo conjunto de amostras (N=924) foi novamente aplicado ao método SOM (Figura 2), e nessa nova rodada nenhuma amostra foi removida.

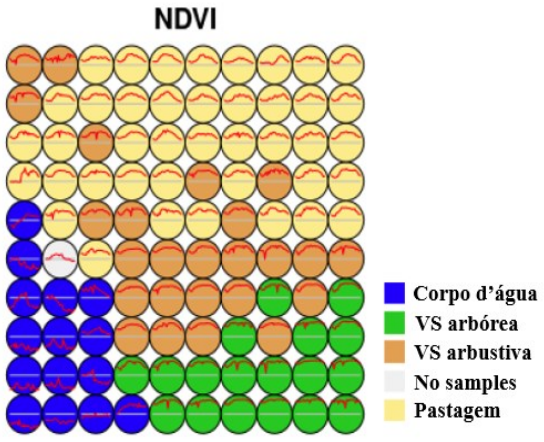

Figura 2. Distribuição de neurônios gerada pelo método SOM (*Self-Organizing Maps)* para o conjunto de amostras (N=924) utilizado para a construção do modelo.

Após análise do SOM (Figura 2), um novo modelo foi gerado e obteve acurácia de 0,96 (IC95% = 0,14). O mapa criado por meio deste modelo apresentou exatidão global de 89% (Tabela 1).

		Classes do mapa					
		VS arbustiva	VS arbórea	PA	Total Geral	Acurácia do usuário (%)	Erros de omissão (%)
Verdade	VS arbustiva	39	6		45	87	13
	VS arbórea	6	44		50	88	12
	PA	5		50	55	91	9
	Total Geral	50	50	50	150		
	Acurácia do produtor (%)	78	88	100			
	Erros de comissão (%)	22	12	0			
	Acurácia global (%)	89					

Tabela 1. Matriz de transição do mapa de vegetação secundária e pastagem do tile BDC 084105, Mato Grosso do Sul, Brasil.

A acurácia do mapa foi superior àquelas obtidas por outro estudo que usou aprendizado de máquina e o algoritmo *Random forest* para construção do modelo (N=21000) para classificar pastagem, formação savânica e formação florestal no Cerrado inteiro para 33 anos (acurácias entre 67% e 74%) [6]. É importante observar, contudo, que comparativamente o modelo criado pelo presente estudo foi construído com menos amostras (N=924) e testado em uma área menor, um tile apenas. O classificador foi capaz de diferenciar áreas com predomínio de indivíduos arbóreos e com dossel formado (VS arbórea) daquelas áreas com predomínio de arbustos (VS arbustiva). Além disso, o classificador também identificou satisfatoriamente áreas de PA.

Os percentuais de acerto da classificação em relação à referência terrestre (i.e., acurácia do usuário) para VS arbustiva, VS arbórea e PA foram de 87%, 88% e 91%, respectivamente (Tabela 1). Novamente, acertos de classificação superiores ao mesmo estudo citado anteriormente [6], cujas acurácias para as classes formação savânica, formação florestal e pastagem foram de 73%, 84% e 73%, respectivamente.

A confusão entre as classes ocorreu devido ao gradiente de similaridade entre elas. O erro de comissão da classe VS arbustiva foi de 22% (Tabela 1), a qual acabou incluída em áreas de VS arbórea e PA. Estas inclusões ocorreram, geralmente, em áreas de VS arbórea sem dossel bem definido, mas com predomínio de indivíduos arbóreos (Figura 3).

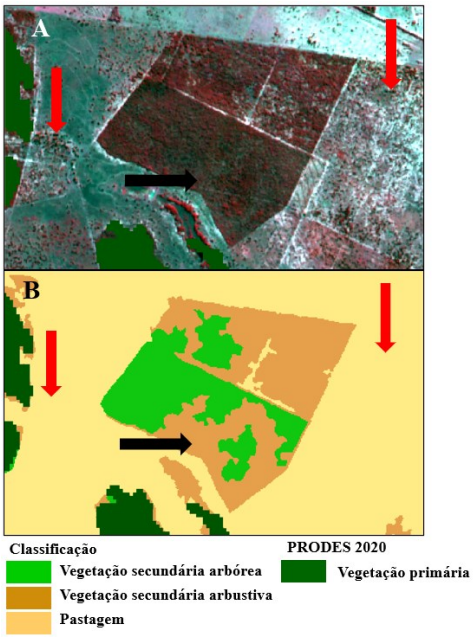

Figura 3. Imagem Sentinel-2 (A) e mapeamento (B) de área localizada no município de Ribas do Rio Pardo, Mato Grosso do Sul, Brasil. As setas indicam áreas em que houve inclusão de vegetação secundária arbustiva em vegetação secundária arbórea (pretas) e áreas de vegetação secundária arbustiva classificadas como pastagem (vermelhas).

Já as inclusões de VS arbustiva em PA foram observadas em áreas com predomínio de gramíneas mais altas e densas, e poucos arbustos (Figura 4).

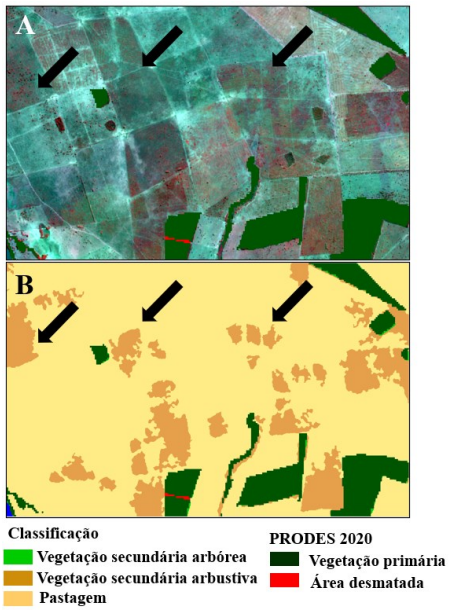

Figura 4. Imagem Sentinel-2 (A) e o mapeamento (B) de área localizada no município de Ribas do Rio Pardo, Mato Grosso do Sul, Brasil. Setas na cor preta indica a área onde houve inclusão de vegetação secundária arbustiva em áreas de pastagem.

A inclusão de VS arbórea em VS arbustiva (12% de erro de comissão) ocorreu em áreas em que há predomínio de arbustos e presença de indivíduos arbóreos, porém não há formação de dossel (Figura 5).

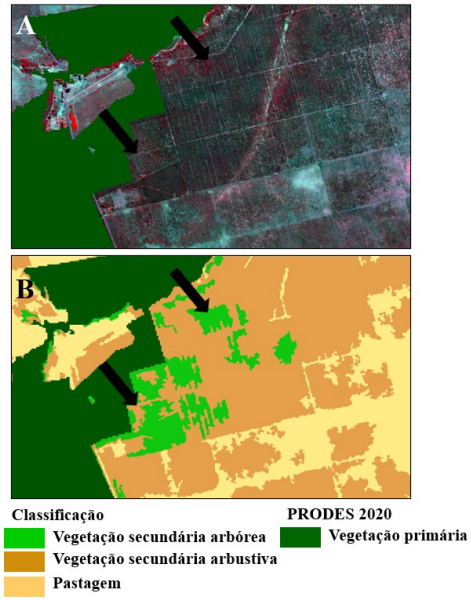

Figura 5. Imagem Sentinel-2 (A) e o mapeamento (B) de área localizada no município de Ribas do Rio Pardo, Mato Grosso do Sul, Brasil. Setas na cor preta indicam a área onde houve inclusão de vegetação secundária arbórea em áreas de vegetação secundária arbustiva.

Áreas com predomínio de arbustos, cujo estrato de gramíneas estava aparente, foram classificadas como PA (Figura 3). No entanto, o erro de comissão de PA em relação à VS arbustiva foi igual a zero (Tabela 1), indicando que este erro está associado à coleta das amostras usadas na construção do modelo, a partir da referência terrestre. Logo, amostras coletadas em pixels de áreas com estas características devem ser incluídas no modelo sob o rótulo de VS arbustiva para que este seja capaz de classificá-las corretamente.

4. CONCLUSÃO

O modelo foi capaz de classificar VS arbórea, VS arbustiva e PA, embora áreas de transição entre elas tenham sido fonte de confusão. Essas classes temáticas são naturalmente difíceis de serem identificadas visualmente em imagens de satélite, o que dificulta a coleta e a rotulação adequada de amostras. Contudo, o uso do método SOM aplicado a séries temporais de imagens Sentinel-2 permitiu identificar e coletar amostras com maior poder discriminatório. O mapa gerado pelo modelo de classificação mostrou-se acurado. Porém, são necessários novos testes que incluam nova coleta de amostras, principalmente em áreas de confusão, bem como a classificação de outras regiões do Cerrado para avaliar o desempenho do classificador em áreas com diferentes fitofisionomias.

5. AGRADECIMENTOS

Os autores agradecem ao projeto *Brazil Data Cube* (BDC - http://brazildatacube.org/) desenvolvido pelo Instituto Nacional de Pesquisas Espaciais (INPE) e aos desenvolvedores do pacote sits (*Satellite Image Time Series Analysis for Earth Observation Data Cubes*), por todo o apoio e suporte recebidos durante o desenvolvimento deste trabalho.

6. REFERÊNCIAS

[1] R. A. Mittermeier, W. R. Turner, F. W. Larsen, T. M. Brooks, and C. Gascon. Global biodiversity conservation: the critical role of hotspots. In: F.E. Zachos, J.C. Habel. *Biodiversity hotspots: Distribution and Protection of Conservation Priority Areas*. Springer: Berlin, 2011, p. 3-22

[2] C. C. Mueller, G. B. Martha Jr. Agropecuária e o desenvolvimento socioeconômico recente no Cerrado. In: F. G. Faleiro, A. L. F. Neto (Ed.). *Savanas: desafios e estratégias para o equilíbrio entre sociedade, agronegócio e recursos naturais.* Planaltina, DF: Embrapa Cerrados, pg.104-169, 2008.

[3] R. Simões, G. Câmara, G. Queiroz, F. Souza, P. R. Andrade, L. Santos, A. Carvalho, and K. Ferreira. Satellite Image Time Series Analysis for Big Earth Observation Data. *Remote Sens. 13*, 2428, 2021.

[4] L. A. Santos, K. Ferreira, M. Picoli, G. Câmara, R. Zurita-Milla and E.W. Augustijn. Identifying Spatiotemporal Patterns in Land Use and Cover Samples from Satellite Image Time Series. *Remote Sens.* 13, 974, 2021.

[5] K. R. Ferreira, G. R. Queiroz, L. Vinhas et al. Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products. *Remote Sens.* 12, 4033, 2020.

[6] A. Alencar, J. Z. Shimbo, F. Lenti et al. Mapping three decades of changes in the Brazilian savanna native vegetation using Landsat data processed in the Google Earth Engine Plataform. *Remote Sens.* 12, 924, 2020.