SELEÇÃO DE AMOSTRAS E PARAMETRIZAÇÃO DE MODELO PARA CLASSIFICAÇÃO DE ÁREAS AGRÍCOLAS NO CERRADO USANDO CUBO DE DADOS SENTINEL-2

Lídia Sanches Bertolo[1], Adriane Calaboni[1], João Francisco Gonçalves Antunes [2], Júlio Cesar Dalla Mora Esquerdo[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, 13080-650 Campinas/SP – Brasil, {lidia.bertolo, adriane.calaboni}@giz.de;
[2]Embrapa Agricultura Digital, Av. André Tosello, 209, Caixa Postal:6041, CEP 13083-886 – Campinas, SP, Brasil, {joao.antunes, julio.esquerdo, alex.coutinho}@embrapa.br

RESUMO

Muitos foram os avanços tecnológicos e metodológicos no mapeamento da cobertura e uso da terra, incluindo análise de séries temporais e classificações baseados em algoritmos de aprendizado de máquina. Este trabalho pretendeu testar, usando soluções inovadoras, como BDC/sits, o desenvolvimento e a sistematização de um método capaz de selecionar amostras e classificar a cobertura e uso da terra do Cerrado. O método deveria, ainda, ampliar a automatização do mapeamento, minimizando a edição vetorial pós-classificação com o cubo de imagens Sentinel-2 para as classes temáticas de agricultura: temporárias de 1 ciclo ou mais, semiperene, perene e silvicultura. A comparação entre o modelo desenvolvido e o produto TerraClass 2020 apresentou 75% de concordância. A validação com pontos distribuídos aleatoriamente entre as classes e inspeção visual mostrou 89% de acurácia global. Conclui-se que o método se mostrou eficaz, gerando um modelo para classificação de áreas agrícolas no Cerrado com um desempenho bastante satisfatório.

Palavras-chave — série temporal, aprendizado de máquina, *random forest*, cultura agrícola, BDC, sits.

ABSTRACT

Land cover and use mapping are in constant technological and methodological advances, including time series analysis and classifications based on machine learning algorithms. This work aimed to test, using innovative solutions such as BDC/sits, the systematization and development of a method capable of selecting samples and classifying the Cerrado's land cover and use. The method should also increase mapping automation, minimizing post-classification vector editing with the Sentibel-2 image cube, for agriculture thematic classes: temporary with 1 cycle or more, semi-perennial, perennial and forestry. The comparison between the developed model and the TerraClass 2020 product showed 75% agreement. Validation with points randomly distributed between classes and visual inspection showed 89% overall accuracy. It is concluded that the method proved to be effective, generating a model for classifying agricultural areas in the Cerrado with a very satisfactory performance.

Key words — *time series, machine learning, random forest, agriculture, BDC, sits.*

1. INTRODUÇÃO

Aproximadamente de 25% do território brasileiro é ocupado pelo bioma Cerrado. O Prodes Cerrado 2020 (Monitoramento do Desmatamento no Cerrado Brasileiro por Satélite) [1] estimou que cerca de 51% de seus 2 milhões de km² já foram antropizados. O TerraClass Cerrado tem o objetivo de mapear a cobertura e uso da terra dessas áreas detectadas pelo Prodes para identificar e qualificar as principais atividades antrópicas desenvolvidas no bioma. Dados do TerraClass Cerrado 2020, ainda em processo de validação, mostram que as terras agrícolas cultivadas somam cerca de 325 mil km² (31% do total da área desflorestada até esta data).

Os avanços metodológicos na classificação automática da cobertura e uso da terra apontam, cada vez mais, na direção de métodos baseados em aprendizado de máquinas. Além disso, o uso de séries temporais para detecção de alvos na superfície terrestre também é uma tendência e vem se mostrando cada vez mais promissor, em substituição à classificação convencional, baseada em informações obtidas a partir de uma imagem de data específica [2,3]. Nesse sentido, o sits (*Satellite Image Time Series Analysis for Earth Observation Data Cubes*), pacote desenvolvido em R [4], e o cubo de dados disponibilizado pelo BDC (*Brazil Data Cube*) [5], surgem no cenário das tecnologias nacionais, como ferramentas complementares e estratégicas para o desenvolvimento de novas abordagens, em apoio a instituições vinculadas ao Governo Federal, responsáveis pelo monitoramento da dinâmica da cobertura e uso da terra no Brasil. Santos *et al*., [6], destacaram que a obtenção de bons resultados na classificação está diretamente relacionada à quantidade e qualidade de amostras de treinamento. Com base nisso, a ferramenta *Self-Organizing Maps* (SOM), acoplada ao pacote computacional sits tem apoiado de forma muito eficiente a avaliação e seleção dos conjuntos de amostras processados pelo sistema.

A busca por um método capaz de identificar, delimitar, quantificar e espacializar a dinâmica da cobertura e uso da terra, com boa acurácia, de forma rápida, detalhada e sistemática foi a principal motivação do TerraClass Cerrado.

Nesse contexto, este trabalho tem por objetivo buscar soluções tecnológicas e metodológicas para ampliar a automatização do processo de mapeamento da cobertura e uso da terra no Cerrado, visando principalmente as classes temáticas de agricultura, sendo temporárias de 1 ciclo ou mais, semiperene, perene e silvicultura.

2. MATERIAL E MÉTODOS

Para testar o desempenho do novo protocolo de mapeamento, foi selecionado um recorte territorial localizado na região sudeste do Cerrado.

Após a finalização do mapeamento da cobertura e uso da terra do Cerrado 2020, realizado a partir da metodologia convencional aplicada nos produtos do projeto TerraClass, foram extraídos 33.544 pontos, localizados no centroide dos polígonos, com informações sobre suas respectivas coordenadas geográficas e a classe temática atribuída, para a realização dos testes da nova metodologia. O trabalho foi desenvolvido seguindo o fluxograma metodológico apresentado na Figura 1.

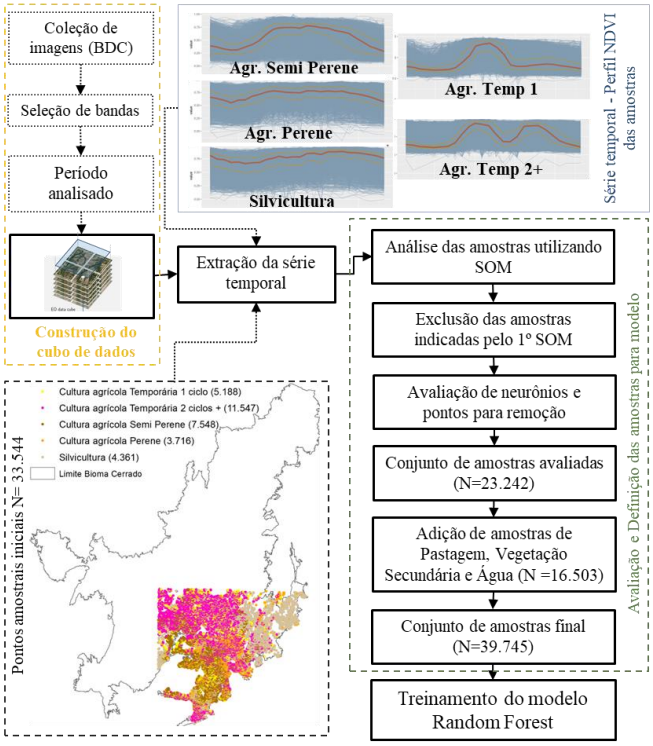

Figura 1. Fluxograma metodológico

Para o processamento digital, foi utilizado o cubo de dados Sentinel-2/MSI, disponibilizado pelo *Brazil Data Cube* (BDC), composto por imagens de composições máximas de 16 dias e, as análises foram executadas através do pacote computacional sits.

Após a realização de alguns testes preliminares, foi definida a construção de um cubo composto por imagens do período de 28/07/2019 a 28/08/2020, e das 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. Extraída a série temporal dos pontos iniciais, o conjunto de dados foi submetido à uma primeira iteração no SOM, procedimento que apresenta uma figura composta por uma matriz bidimensional, cujos neurônios resultantes definem os agrupamentos amostrais, os arranjos e relações de proximidade e distância das classes e, também, os valores atípicos das amostras (Figura 2). Além disso, o algoritmo também indica quais amostras devem ser mantidas, verificadas ou removidas. Após análise preliminar foi acatada a indicação para remoção de 1.295 amostras de agricultura perene, 1.726 de semiperene, 1.850 de agricultura temporária de 1 ciclo, 2.131 de mais de 1 ciclo e 896 de silvicultura. Após esses ajustes no conjunto amostral, o conjunto resultante de dados foi submetido a uma segunda iteração no SOM. Mas, desta vez, os neurônios resultantes foram analisados de acordo com o conjunto de pontos agrupados em cada um deles, bem como dos neurônios do seu entorno. Assim, um conjunto final de 23.242 amostras das classes de interesse foi consolidado, ao qual acrescentou-se 16.503 amostras de corpos d'água, vegetação secundária e pastagem para promover a melhoria do potencial discriminatório do modelo. Como as classes de maior interesse são as agrícolas, esse conjunto complementar de amostras não foi avaliado pelo SOM. A distribuição final dos neurônios e a quantidade de amostras utilizadas para a construção do modelo estão ilustradas na Figura 2.

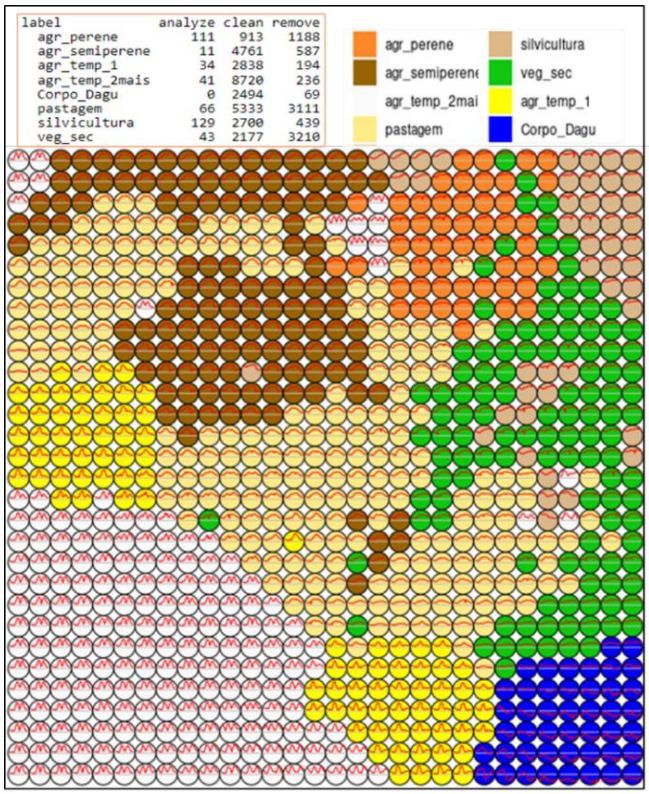

Figura 2. Distribuição espacial dos neurônios das amostras finais utilizadas para modelo

Uma área piloto localizada no oeste de Minas Gerais, abrangendo todas as classes agrícolas de interesse para o mapeamento, foi selecionada para aplicação do modelo. Para a classificação foi utilizado o algoritmo de aprendizado de máquina *Random Forest* e aplicado um filtro Bayesiano pós -classificação, ambos adotando os parâmetros indicados pelo sits, sendo 200 árvores de decisão e janela 5x5, respectivamente. Uma vez classificada, a imagem final foi então submetida a mais uma filtragem para remoção de pequenos polígonos e falhas por meio do *plugin* SCP para QGIS (*Semi-Automatic Classification Plugin*) usando como parâmetros *size threshold* 300 e *pixel connection* 4.

Por fim, foi realizada uma comparação do dado do TerraClass 2020, baseado em metodologia convencional, utilizando séries temporais do sensor MODIS e interpretação visual com base em imagens Landsat-8/OLI, e o produto final do método desenvolvido neste trabalho. Essa comparação visou comparar, por meio da matriz de confusão e distribuição espacial, as concordâncias e discordâncias resultantes da aplicação de ambos os métodos. Adicionalmente, foram sorteados e avaliados 50 pontos, para cada classe de interesse, avaliados por meio de interpretação visual com base em imagens do satélite Sentinel-2, imagens de alta resolução espacial disponíveis no Google Earth e de séries temporais de índices vegetativos MODIS disponibilizadas pelo Sistema de Análise Temporal da Vegetação (SATVeg) [8].

3. RESULTADOS E DISCUSSÃO

A Figura 3 apresenta a matriz de confusão entre o mapa TerraClass 2020 e o modelo baseado em Sentinel-2 (S2) e a Figura 4 mostra alguns exemplos comparando os produtos.

		TC 2020								
	88104	Agr. Perene	Agr. Semiperene	Agr. Temp. 1C	Agr. Temp. 2+	Past.	Silv.	VS	Total	Exatidão Usuário
Modelo	Agr. Perene	74.538	105	74	2.307	2.261	263	554	80.102	93%
	Agr. Semiperene	98	79.178	1.762	3.743	2.324	99	115	87.319	91%
	Agr. Temp. 1C	452	2.864	32.231	13.662	7.858	31	31	57.128	56%
	Agr. Temp. 2+	4.448	4.853	12.570	331.082	19.320	431	382	373.086	89%
	Pastagem	27.743	13.731	15.872	49.184	411.768	3.431	6.008	527.737	78%
	Silvicultura	2.565	50	60	448	1.535	35.384	771	40.814	87%
	VS	11.690	4.923	2.707	20.185	78.410	6.307	25.595	149.818	17%
	Total	121.534	105.704	65.276	420.612	523.476	45.947	33.456	1.316.005	
	Exatidão Produtor	61%	75%	49%	79%	79%	77%	77%	Conc. Global	75%

Figura 3. Matriz de confusão entre TC2020 e Modelo Sentinel

Apesar de métodos distintos terem sido utilizados para obtenção dos mapas, é possível notar uma concordância global de 75% entre os mapeamentos. O resultado da classificação do modelo S2 apresenta um detalhamento maior dos polígonos (Figura 3), o que era esperado em função deste ser baseado em imagens Sentinel-2 (resolução espacial de 10m), enquanto o TerraClass 2020 foi mapeado usando como base imagens Landsat-8/OLI (resolução espacial de 30m). A exatidão do usuário foi alta para classes que, historicamente, vêm se mostrando difíceis de serem mapeadas de forma automática, como a agricultura perene (93%), semiperene (91%) e silvicultura (87%). Para a agricultura temporária de 2 ou mais ciclos, a exatidão do usuário também foi alta (89%), indicando que o classificador apresentou baixo percentual de erro de comissão dessas classes. Já a agricultura temporária de 1 ciclo apresentou uma menor exatidão (56%) e, quando investigada a razão, notou-se que essas áreas representam pequenos talhões em meio à agricultura de 2 ciclos ou eram áreas de reforma de cana-de-açúcar que foram subestimadas pelo TC Cerrado.

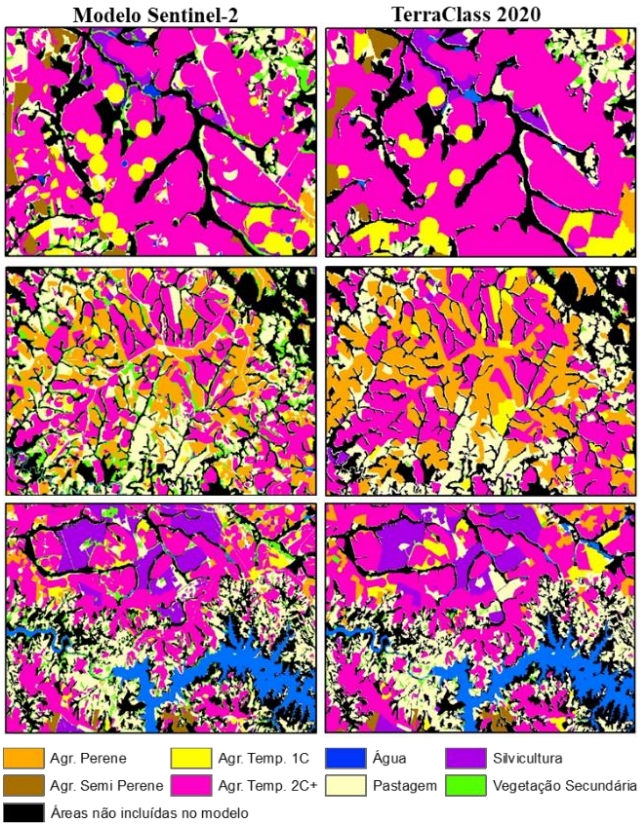

Figura 4. Exemplos comparando o mapeamento gerado pelo modelo Sentinel-2 com o produto TC Cerrado 2020 (Landsat 8)

Com relação à exatidão do produtor, os valores foram um pouco mais baixos para as culturas perene, semiperene e silvicultura 61%, 75% e 77%, respectivamente, expressando uma quantidade maior de erros de omissão, assim como para agricultura temporária de 2 ou mais ciclos (79%). Esses valores obtidos podem ser explicados pelo melhor detalhamento inerente do uso de uma imagem de maior resolução espacial. Quando verificados alguns pontos que mostravam essas diferenças, notou-se que áreas de plantio mais recente de silvicultura e cultura perene também não foram incluídas na classificação pelo modelo S2. A cultura temporária de 1 ciclo apresentou a menor condordância dentre as classes mapeadas, ficando próxima a 50%. Áreas de cultivo de algodão de 1 ciclo não foram classificadas pelo modelo S2, indicando a necessidade de inclusão de pontos que representem essa cultura, uma vez que o perfil temporal do NDVI para essa classe pode ser um pouco mais longo (130

a 220 dias) do que o perfil de culturas de 1 ciclo, como soja e milho (100 a 160 dias).

Visando confrontar esses resultados, foi realizada uma validação dos mapas com a distribuição de 250 pontos de controle nas classes de interesse, com 242 pontos finais utilizados para a validação, sendo que 8 deles caíram em borda de área urbana ou corpo d'água e por isso foram removidos. A Figura 5 apresenta as matrizes de confusão, pelas quais observa-se a discordância de alguns pontos no mapa TC2020 com pastagem e vegetação secundária, novamente relacionadas à diferença de resolução espacial entre os produtos, onde muitos dos pontos foram identificados como áreas de bordas de polígonos, os quais se apresentam mais bem definidos no modelo Sentinel-2. Já a confusão de cultura perene com pastagem no modelo S2 pode ser explicada por áreas em estágios iniciais do plantio, onde o perfil temporal do NDVI se assemelha às pastagens. Apesar disso, o modelo S2 apresentou percentuais de acurácia do usuário e do produtor acima de 91% para as classes de interesse, com exceção da agricultura temporária de 1 ciclo (73%) e perene (80%). Por fim, o fato do método desenvolvido utilizar imagens de melhor resolução espacial resultou em uma acurácia global do modelo S2 superior à do mapa apresentado pelo TC Cerrado (89% frente a 83%).

	Pontos amostrados								
TC 2020	Perene	Semi Per.	Temp. 1C	Temp. 2+	Past	Silv	VS	Total Geral	Ac. Usuário
Agric. Perene	44			2	2		2	50	88%
Agric. Semi Perene		37	2	1	5		3	48	77%
Agric. Temp. 1C			37	4	6		1	48	77%
Agric. Temp. 2+			1	42	3		2	48	88%
Pastagem								0	
Silvicultura				1		41	6	48	85%
VS								0	
Total Geral	44	37	40	50	16	41	14	242	
Ac. Produtor	100%	100%	93%	84%		100%			
								Ac. Global	83%
Mod RF_SE	Perene	Semi Per.	Temp. 1C	Temp. 2+	Past	Silv	VS	Total Geral	Ac. Usuário
Agric. Perene	35							35	100%
Agric. Semi Perene		35	1					36	97%
Agric. Temp. 1C		1	29					30	97%
Agric. Temp. 2+			5	48				53	91%
Pastagem	7	1	5	1	16	1		31	52%
Silvicultura	1					38		39	97%
VS	1			1		2	14	18	78%
Total Geral	44	37	40	50	16	41	14	242	
Ac. Produtor	80%	95%	73%	96%	100%	93%	100%		
								Ac. Global	89%

Figura 5. Matriz de confusão – validação dos mapeamentos TC e modelo RF_SE

4. CONCLUSÕES

A estratégia metodológica apresentada neste trabalho para seleção de amostras se mostrou eficaz e gerou um modelo para classificação de áreas agrícolas no Cerrado com desempenho bastante satisfatório. Os avanços metodológicos aqui apresentados não seriam possíveis sem a utilização da infraestrutura de cubos Sentinel-2 disponibilizada pelo BDC/INPE e do uso do pacote computacional sits, adotado para a avaliação das amostras e classificação das imagens utilizando séries temporais. O modelo Sentinel-2 possibilitou a geração de uma classificação com pouca necessidade de edição vetorial, condicionando uma produção de mapas mais eficiente. Futuramente, pontos adicionais podem ser incluídos no modelo para melhorar a acurácia de algumas classes.

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] Instituto Nacional de Pesquisas Espaciais. Coordenação Geral de Observação da Terra. PRODES – Incremento anual de área desmatada no Cerrado Brasileiro. Disponível em: http://www.obt.inpe.br/cerrado.

[2] Chaves, M. E. D.; Picoli, M. C. A.; Sanches, I. D. Recent Applications of Landsat 8/OLI and Sentinel-2/MSI for Land Use and Land Cover Mapping: A Systematic Review. *Remote Sensing*. 2020; 12(18):3062. https://doi.org/10.3390/rs12183062

[4] Simões, R.; Camara, G.; Queiroz, G; Souza, F.; Andrade, P.; Santos, L.; Carvalho, A.; Ferreira, K. Satellite Image Time Series Analysis for Big Earth Observation Data. *Remote Sensing*, 13: 2428, 2021. doi:10.3390/rs13132428.

[5] Ferreira, K. R.; Queiroz, G. R.; Vinhas, L.; Marujo, R. F. B.; Simoes, R. E. O.; Picoli, M. C. A.; Camara, G.; Cartaxo, R.; Gomes, V. C. F.; Santos, L. A.; Sanchez, A. H.; Arcanjo, J. S.; Fronza, J. G.; Noronha, C. A.; Costa, R. W.; Zaglia, M. C.; Zioti, F.; Korting, T. S.; Soares, A. R.; Chaves, M. E. D.; Fonseca, L. M. G. Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products. *Remote Sens*ing 2020, 12, 4033. https://doi.org/10.3390/rs12244033

[6] Santos, L.; Ferreira, K, Camara, G.; Picoli, M.; Simoes, R. Quality control and class noise reduction of satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:75-88, 2021. doi:10.1016/j.isprsjprs.2021.04.014.

[7] Congedo, L. Semi-Automatic Classification Plugin: A Python tool for the download and processing of remote sensing images in QGIS. *Journal of Open Source Software*, 6(64), 3172, 2021. https://doi.org/10.21105/joss.03172

[8] Embrapa (Empresa Brasileira de Pesquisa Agropecuária). SATVeg. Disponível em: <https://www.satveg.cnptia.embrapa.br/>. Acesso em: 01 set. 2022.