CLASSIFICAÇÃO DE IMAGENS COM DISTINTOS PARÂMETROS UTILIZANDO O GOOGLE EARTH ENGINE E ALGORITMO DE APRENDIZADO DE MÁQUINA

Samuel Gameiro [1], Jairo Matos da Rocha [1], Igor Rodrigues dos Santos [1], Manuel Eduardo Ferreira [1], Karen Borges-Almeida [2], Rosane Garcia Collevatti [2]

[1]Universidade Federal de Goiás – Laboratório de Processamento de Imagens e Geoprocessamento, Campus Samambaia, Almeida Palmeiras, s/n - Chácaras Califórnia, samuelgameiro@discente.ufg.br; devjairomr@gmail.com; santosigor@discente.ufg.br; manuel@ufg.br

[2]Universidade Federal de Goiás - Laboratório de Genética & Biodiversidade, Campus Samambaia, Almeida Palmeiras, s/n - Chácaras Califórnia, karen.cbiologicas@gmail.com; rosanegc68@icloud.com

RESUMO

O mapemento do uso e cobertura do solo para um dado território é de extrema importância para diversas análises socioambientais e econômicas. O uso da plataforma Google Earth Engine (GEE) e algoritmos de aprendizado de máquina ganham cada vez mais destaque nesse processo. Nesse artigo objetivamos avaliar, por meio da plataforma GEE e o algoritmo *Random Forest*, três modelos de classificação, utilizando diferentes bandas espectrais de imagens Planet, junto com índices de vegetação derivados desse conjunto de dados. As imagens referem-se ao ano de 2022, cobrindo uma grande área da região Centro-Sul do estado de Goiás. O modelo utilizando a maior quantidade de parâmetros (63) foi o que obteve mapas mais detalhados e com a maior acurácia, quando comparado aos modelos com 12 e 21 parâmetros, demonstrando a importância de se utilizar parâmetros variados para o processo de classificação e não apenas as tradicionais bandas RGB deste sensor.

Palavras-chave — *Random Forest*, índice de vegetação, acurácia, Planet, variáveis.

ABSTRACT

The land use and land cover mapping for a given territory is crucial for several socio-environmental and economic analyses. The use of the Google Earth Engine (GEE) platform and machine learning algorithms are increasingly highlighted in these processes. In this article, we aim to evaluate three classification models using different spectral bands from Planet images, operating the GEE platform and the Random Forest algorithm, in addition to vegetation indices derived from this dataset. The images refer to the year 2022, covering a large area in the Center-South region of Goiás state. The model using the largest number of parameters (63) was the one that obtained more detailed maps and the highest accuracy when compared to models with 12 and 21 parameters, demonstrating the importance of using different parameters for the classification process and not only the traditional RGB bands of this sensor.

Keywords *— Random Forest, vegetation index, accuracy, Planet, variables.*

1. INTRODUÇÃO

O uso da plataforma Google Earth Engine (GEE) para a classificação de imagens e produção de mapas sobre o uso e a cobertura do solo vem aumentando a cada dia, e se tornando uma das principais plataformas para realizar tais estudos [1], como observado, por exemplo, em [2] e [3].

Uma das vantagens dessa plataforma é poder utilizar algoritmos robustos e avançados para a classificação de um grande volume de imagens. Os algoritmos de aprendizado de máquina se destacam nesse quesito, entre estes o Random Forest (RF) [4, 5] e o Classification and Regression Trees (CART) [2, 3].

Além do uso de algoritmos robustos, distintos parâmetros espectrais para a classificação também podem influenciar na qualidade dos resultados, como evidenciado por [6] e [7], que utilizaram diferentes índices espectrais para classificação de área queimadas e na identificação de alterações em áreas da caatinga, respectivamente.

Neste contexto, objetiva-se com o presente trabalho avaliar três modelos de classificação do uso e cobertura do solo utilizando o algoritmo Random Forest e a plataforma do GEE, a partir de bandas espectrais e índices de vegetação advindos de imagens do satélite Planet.

2. MATERIAL E MÉTODOS

A área de estudo refere-se um quadrante de 100x100 km localizado na região Centro-Sul do estado de Goiás, envolvendo a Região Metropolitana de Goiânia. Esta área teve origem em um Projeto Ecológico de Longa Duração (PELD) iniciado no município de Silvânia-GO, ampliada recentemente para extrapolar modelos ecológicos que avaliam o impacto da fragmentação e conversão da paisagem natural sobre a fauna e flora do bioma Cerrado.

As classificações foram feitas no ambiente do Google Earth Engine, com a utilização do algoritmo Random Forest e imagens do mosaico Planet de julho de 2022. As classificações utilizaram 20 árvores de decisão, divididas em

6 classes de uso e ocupação do solo. A coleta de amostras para treino do modelo de classificação foi feita visualmente, sendo utilizadas 200 amostras de cada classe temática. No processo de teste e validação, foram utilizadas 25 amostras de cada classe, coletadas diretamente em campo.

O processo de classificação utiliza o valor dos pixels para identificar amostras de espectro semelhante, normalmente restritas às bandas espectrais dos satélites. Aqui foram utilizadas 3 combinações diferentes de bandas, sendo elas denominadas de BI, MMM e BIMMM (Figura 1).

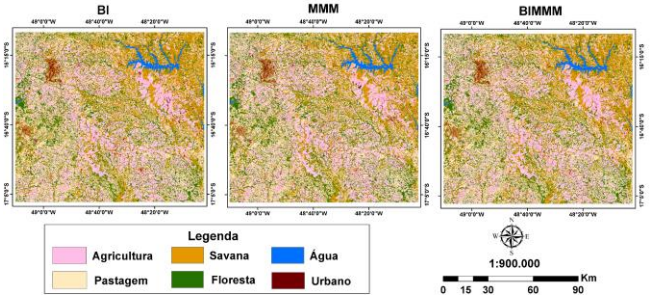

Figura 1. Classificação de uso e ocupação da área total com os distintos parâmetros

A classificação BI consistiu na utilização de bandas R/G/B/NIR, presentes nas imagens Planet, além de 17 índices de vegetação, sendo eles NDVI, SAVI, VARI, NDWI, NDVIB, ENDVI, EXG, TGI, GLI, GNDVI, GRVI, MNLI, MSR, RDVI, TVDI, OSAVI e EVI, totalizando 21 parâmetros para a classificação.

A classificação MMM foi realizada utilizando os valores mínimos, máximos e as médias das bandas R/G/B/NIR, totalizando 12 parâmetros.

Já a BIMMM é baseada nos dois anteriores, utilizando os valores mínimos, médios e máximos de todas as bandas e todos os índices de vegetação, perfazendo um total de 63 parâmetros.

Para análise dos resultados, foram quantificadas as percentagens de classificação de cada classe temática, além de calculados o índice kappa (K), acurácia global (AG), acurácia do consumidor (AC) e do produtor (AP) e a importância de cada variável no processo de classificação.

3. RESULTADOS E DISCUSSÃO

As classes obtiveram percentagens semelhantes na sua grande maioria, com pequenas discrepâncias entres os três modelos de classificação, conforme evidenciado na tabela 1.

	BI	MMM	BIMMM
AGRICULTURA	26,46%	26,88%	26,64%
PASTAGEM	25,26%	25,77%	25,35%
SAVANA	25,43%	24,16%	25,31%
FLORESTA	18,89%	19,11%	18,94%
ÁGUA	1,59%	1,63%	1,60%
URBANO	2,34%	2,42%	2,12%

Tabela 1. Percentagem das classes temáticas em cada um dos modelos de classificação.

A classe com maior diferença em percentagem de área foi a savana, apresentando no MMM diferenças maiores do que 1% em relação às outras classificações, o que equivale a 56 km². Todas as outras classes obtiveram diferenças em torno de 0,5%, o que ainda é significativo, pois representam áreas com pouco mais de 25 km² (~2500 ha).

Na etapa de análise visual (figura 2), foram selecionadas 5 áreas com discrepâncias, representadas pelos círculos vermelhos nas imagens. As principais diferenças estão representadas pelas classes de urbano, agricultura e savana.

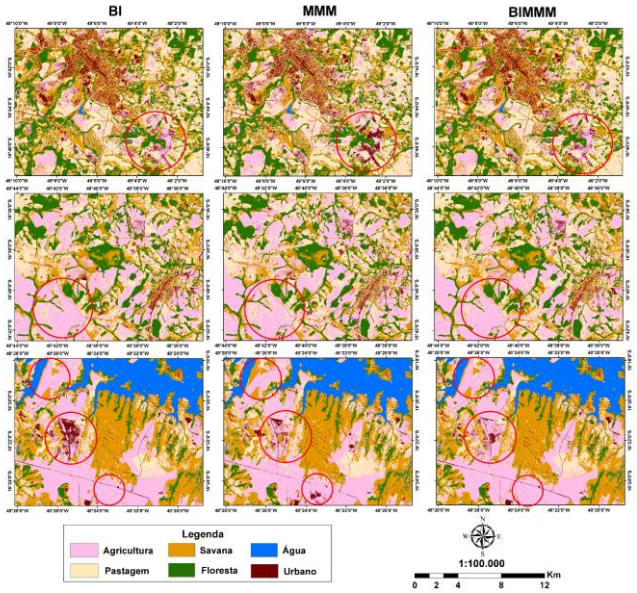

Figura 2. Áreas discrepantes (círculos vermelhos) nos mapas.

Na primeira comparação (Figura 2), localizada próximo à cidade de Anápolis, é exibida na classificação MMM uma região classificada como urbana, enquanto nas outras classificações essas áreas são de agricultura. Isso demonstra uma confusão do modelo MMM com possíveis áreas de solo exposto dentro de áreas agricultáveis.

A segunda comparação (Figura 2) detalha a região de Silvânia, junto à Floresta Nacional do município. A região é rodeada, principalmente, de áreas agricultáveis. Porém, algumas dessas áreas foram classificadas como áreas de pastagem e savana, principalmente no modelo MMM. Isso salienta a confusão desse modelo para classificar áreas agricultáveis.

A última comparação (Figura 2) três áreas são destacadas. Na área do topo da figura, evidenciam-se diferenças nas classificações com o MMM em relação à

classe de agricultura, classificada erroneamente como savana. Na região central da imagem (ainda na 3ª linha da Figura 2), o modelo BI classificou uma grande área como classe urbana, enquanto os outros dois modelos evidenciam uma melhor área para tal classe; já na parte mais baixa da imagem, o modelo MMM novamente classificou a área de agricultura como urbano.

Essas comparações visuais são de grande relevância e indicam, principalmente, uma diferença do modelo MMM com os outros dois, ressaltando que o mesmo possui outras discrepâncias, principalmente em áreas de agricultura.

3.1. Acurácias

O erro de omissão (tabela 2), ou Acurácia do Produtor (AP), ocorre quando o algoritmo deixa de mapear um *pixel* corretamente, baseado nas classes e amostras de *pixels* selecionadas pelo analista. Enquanto o erro de comissão (tabela 3), ou Acurácia do Consumidor (AC), ocorre quando o algoritmo atribui uma determinada classe a um *pixel*, sabendo que ele pertence a outra classe [8].

	BI	MMM	BIMMM
AGRICULTURA	0,976	0,974	0,98
PASTAGEM	0,965	0,963	0,974
SAVANA	0,976	0,967	0,967
FLORESTA	0,985	0,987	0,985
ÁGUA	0,994	1	0,994
URBANO	0,979	0,974	0,979

Tabela 2. Acurácia do produtor.

O modelo BIMMM obteve valores de AP maiores para as classes de agricultura e pastagem, sendo classes que, juntas, ocupam mais de 50% de toda a área de estudo. Seus valores para as classes de savana e urbano se igualam ao modelo BI, sendo os maiores entre os três modelos confeccionados.

O modelo MMM se mostrou mais adequado para áreas de floresta e recursos hídricos, obtendo valores maiores do que os outros dois modelos. Entretanto, se mostrou com menor potencial de discriminação para as outras classes temáticas.

O modelo BI demonstrou ser o mais estável ou intermediário entre os três modelos avaliados. Isto é, ficou sempre na média dos modelos comparados.

	BI	MMM	BIMMM
AGRICULTURA	0,972	0,971	0,971
PASTAGEM	0,972	0,960	0,972
SAVANA	0,962	0,967	0,974
FLORESTA	0,989	0,989	0,981
ÁGUA	1	1	1
URBANO	0,994	0,989	0,994

Tabela 3. Acurácia do consumidor.

Na AC, o modelo BIMMM foi o melhor apenas na classe de savana, quando comparado aos demais modelos; entretanto, obteve valores iguais a BI em urbano e pastagem.

Na classe de agricultura, que predomina na região, os valores dos três modelos ficaram parecidos, com apenas o modelo BI obtendo 0,001 de acurácia a mais do que os outros modelos. A classe de água foi correta e obteve valor máximo (1) para todos os modelos.

O modelo MMM não foi superior em nenhuma das análises de AC. Reforçando a constatação de não ser o modelo mais adequado para essa classificação.

As últimas análises foram realizadas com os índices K (Kappa), AG e determinação da importância de cada variável, apresentadas na tabela 4.

	BI	MMM	BIMMM
Kappa	0,9722	0,9694	0,9732
Acurácia Global	0,9776	0,9752	0,9783
Importância da variável	NDVIB 55	Rmean 100	EXGmax 18,66

Tabela 4. Índice Kappa, acurácia global e importância da variável.

Nos valores de K e AG, o modelo BIMMM mostrou certa superioridade em comparação com os outros dois modelos, com valores de 0,9732 e 0,9783, respectivamente, sendo considerado o melhor modelo.

O modelo BI foi o intermediário, com valores de K e AG iguais a 0,9722 e 0,9776, respectivamente, enquanto o modelo MMM obteve os menores valores de K e AG, com 0,9694 e 0,9752, respectivamente. Esses valores de acurácia demonstram a inferioridade do modelo MMM na

classificação das imagens, quando comparada aos modelos BI e BIMMM.

Analisando as variáveis utilizadas, o modelo BI, o qual utilizou 17 parâmetros, obteve o índice de vegetação NDVIB como o mais influente e importante no processo de classificação, com valor de 55. Para o modelo MMM, que utilizou 12 parâmetros, a banda Rmean foi a mais influente, com valor de 100. Esse resultado indica a razão do modelo MMM obter valores mais acurados na classe de floresta, visto que a banda R é amplamente utilizada para classificações que envolvem vegetação. No modelo BIMMM, o qual utilizou 63 parâmetros, o índice EXGmax foi o mais relevante, com valor de 18,66. Essa análise é de grande importância, pois parâmetros que obtiverem pouca ou nenhuma importância na classificação podem ser excluídos do processo, agilizando-se a etapa de processamento e análise.

4. CONCLUSÕES

O uso da plataforma GEE e de algoritmos de aprendizado de máquina são extremamente eficientes no processo de classificação do uso e cobertura do solo, com a possibilidade de utilizar os mais diversos parâmetros como variáveis de entrada para a classificação, tornando os modelos gerados mais robustos e precisos.

Dentre os três modelos testados, o modelo BIMMM foi o que obteve os melhores resultados, tanto visuais quanto de acurácia, obtendo valores mais altos nos índices K e AG, de 0,9732 e 0,9783, respectivamente. Além disso, apresentou valores mais altos na análise de AC e AP, demonstrando que a utilização de várias bandas e vários índices de vegetação podem trazer melhores mapeamentos de uso e cobertura do solo.

O modelo MMM, apesar de apresentar resultados com alta acurácia, demonstrou ser o modelo com os valores mais baixos quando comparado aos outros dois, não sendo indicado a sua metodologia, a qual utiliza apenas os valores de mínimas, médias e máximas das bandas.

Estudos com novos índices, ou com a exclusão de índices que obtiveram baixa importância na análise, ainda são necessários, para que haja o aperfeiçoamento das técnicas de classificação, tornando o resultado cada vez mais próximo da realidade encontrada em campo.

5. REFERÊNCIAS

[1] L. Yang, J. Driscol, S. Sarigai, Q. Wu, H. Chen, C. D. Lippitt. Google Earth Engine and Artificial Intelligence (AI): a Comprehensive Review. *Remote Sensing*, v. 14, 3253, 2022.

[2] C.O.F. Silva. Classificação supervisionada de área irrigada utilizando índices espectrais de imagens landsat-8 com google earth engine, *Irriga*, v. 25, n.1, pp. 160-169, 2020.

[3] W. S. Carvalho, F. J. C. M. Filho, T. L. dos Santos. Uso e cobertura do solo utilizando a Plataforma Google Earth Engine (GEE): Estudo de caso em uma Unidade de Conservação, *Brazilian Journal of Development*, v. 7, n. 2, pp. 15280-15300, 2021.

[4] L. E. S. Currihuinca, J. M. Chaves, W. J. S. F. Rocha, J. S. B. Lobão, P. M. Falcão, Identificação das Dunas do Atacama (Norte do Chile) a partir da avaliação de três algoritmos no Google Earth Engine. *Revista Brasileira de Geografia Física*, v. 14, n. 6, pp. 3294-3315, 2021.

[5] S. S. Lima, J. L. P. Cordeiro, L. P. Teixeira, R. P. Maia, M. V. C. da Silva, M. F. Moro. Caracterização geográfica e dinâmica de uso da terra da Ibiapaba e seu entorno, Domínio Fitogeográfico da Caatinga. *Revista Brasileira de Geografia Física*, v. 15, n. 5, pp. 2500-2524, 2022.

[6] J. A. S. Júnior, A. P. Pachêco. Avaliação de índices espectrais e Classificação Normal Bayes usando imagens OLI e TIRS para o mapeamento de áreas queimadas no Cerrado. *Revista Brasileira de Meio Ambiente*, v. 10, n. 3, pp. 132-147, 2022.

[7] U. J. S. Junior, R. M. Gonçalves, L. M. M. de Oliveira, J. A. S. Júnior. Sensibilidade Espectral dos Índices de Vegetação: GNDVI, NDVI e EVI na Mata Ciliar do Reservatório de Serrinha II - PE, Brasil. Revista Brasileira de Cartografia, v. 73, n. 1, 2021.

[8] R. G. Pontius Jr & M. Millones. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, v. 32(15), pp. 4407-4429, 2011.