ANÁLISE DE DESEMPENHO DO CLASSIFICADOR *RANDOM FOREST* NA DETECÇÃO DE CLASSES DE USO E COBERTURA DA TERRA EM ÁREAS DE NÃO FLORESTA EM SALVATERRA/PA.

Nicole Rodrigues de Magalhães[1], *Alessandra Rodrigues Gomes*[1], *Denison Lima Corrêa*[2]

[1]INPE - Instituto Nacional de Pesquisas Espaciais, Av. dos Astronautas 1754, São José dos Campos, 12227-010 Brasil, nicole.rodrigues.magalhaes@gmail.com; alessandra.gomes@inpe.br.
[2]Universidade Estadual do Pará – UEPA, Rodovia, PA-125, s/n, 68625-000 - Paragominas - PA, Brasil, denison.correa@uepa.br.

RESUMO

O presente trabalho teve como objetivo avaliar o desempenho do classificador *Random Forest* utilizando imagens do Sentinel-2B na detecção de classes de uso e cobertura da terra em áreas de Não floresta, baseadas nas classes do Projeto TerraClass, em uma região de Salvaterra, no Pará nos anos de 2020 e 2021. A metodologia foi executada em sete etapas: seleção de imagens, geração de mosaicos, coleta de pontos amostrais, seleção de bandas e classificação, pós-classificação e avaliação dos modelos. Foram utilizados os índices Kappa, acurácia global e matriz de confusão para avaliação de desempenho. O trabalho apresentou bons resultados na classificação, segundo avaliação dos modelos. Sugeriu-se que haja a validação dos dados em *in loco*, no intuito de melhorar a precisão na detecção de classes de uso e cobertura do solo e aprimoramento dos procedimentos metodológicos.

Palavras-chave **—** Random Forest, Google Earth Engine, TerraClass, Não-floresta, Salvaterra.

ABSTRACT

The present work aimed to evaluate the performance of the Random Forest classifier using Sentinel-2B images in the detection of land use and land cover classes in no forest areas, based on the classes of the TerraClass Project, for a region of Salvaterra, Pará in the years 2020 and 2021. The methodology was performed in seven steps: image selection, mosaic generation, collection of sampling points, band selection and classification, post-classification and evaluation of the models. Kappa indices, global accuracy and confusion matrix were used for performance evaluation. The work presented good results in the classification, according to the evaluation of the models. It was suggested that the data be validated in loco, in order to improve the accuracy in the detection of land use and land cover classes and to improve the methodological procedures.

Key words **—** *Random Forest, Google Earth Engine, TerraClass, No forest, Salvaterra.*

1. INTRODUÇÃO

A Amazônia, enquanto maior ecossistema florestal tropical contínuo, conta com uma extensão de aproximadamente 5,4 milhões de km² sendo que cerca de 62% está em território brasileiro, além de ser responsável por 15% da fotossíntese total que ocorre no planeta e detém 25% de todas as espécies florestais terrestres do mundo [1]. Ainda que a composição do bioma Amazônia seja de grandes extensões florestais, o ecossistema é complexo e possui alta diversidade florística [2]

Dentro desta complexidade, encontram-se as vegetações de fitofisionomias não florestais que correspondem a uma área de 290 mil km², ou seja, sua área equivale a 7% dentro do bioma Amazônia [3]. Estas áreas, conhecidas como não-florestas (NF), são compostas principalmente por campiranas e savanas que podem ser oriundas de incidência natural ou como resultado de ações antrópicas [4].

Com a pressão intensa do desflorestamento nas vegetações de fitofisionomias florestais e não florestais, o PRODES (Projeto de Monitoramento do Desmatamento na Amazônia Legal) realiza desde 1988 o monitoramento da perda de cobertura em áreas florestais, fornecendo informações acerca das taxas anuais de desmatamento. Apesar da importância deste projeto para monitoramento das áreas florestais, ele não abrange áreas de formação não florestais.

O Projeto TerraClass [3], desenvolvido numa parceria entre INPE e Embrapa, tem como objetivo mapear uso e cobertura da terra em áreas desflorestadas identificadas pelo Projeto PRODES e procura, a cada ano, melhorar a metodologia para sistematizar o mapeamento, a geração de mapas e estatisticas de uso e cobertura da terra e a divulgação e disponibilização das informações e dados gerados à toda a sociedade [3].

Entendendo a importância das áreas de NF e a necessidade de entender os processos de ocupação nelas, este trabalho tem como objetivo avaliar o desempenho do classificador *Random Forest* utilizando imagens do Sentinel-2B na detecção de classes de uso e cobertura da terra em áreas de NF, baseadas nas classes do Projeto TerraClass, para uma região de Salvaterra, Estado do Pará, nos anos de 2020 e 2021.

2. MATERIAL E MÉTODOS

2.1 Área de estudo

O município de Salvaterra está localizado ao leste do Arquipélago do Marajó, sua sede está localizada nas coordenadas 00º45'12" S, 48º31'00" W e possui 918,563 km² de extensão territorial.

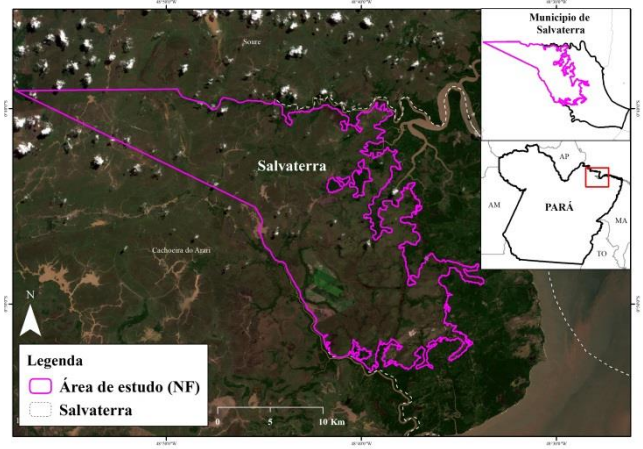


Figura 1. Área de estudo e município de Salvaterra, Pará.

De acordo com os dados do TerraClass do ano de 2014, o município de Salvaterra possui 433.3 km² de NF identificadas mas não mapeadas, ocupando 41,35% do território. Na área de NF encontram-se as savanas amazônicas, lavrados e campiranas [4]. Além de ser a classe que ocupa a maior área no município, sua escolha justifica-se devido à peculiaridade da fitofisionomia encontrada na região, como as manchas de savana ou savana parque que são caracterizadas pela presença de árvores baixas com espaçamento, intercaladas por arbustos e vegetação herbácea, que podem ter sido resultado de um ecossistema anterior ou que podem ter sido modificadas por atividades pecuárias e agrícolas [5].

2.2 Processamento de dados

O satélite Sentinel-2B possui sensores multiespectrais de alta resolução espacial e temporal com a finalidade de realizar o monitoramento da vegetação, do solo e áreas costeiras [6]. Com o aumento da necessidade de técnicas de classificação mais sofisticadas, os algoritmos de predição de classes se tornam útil e realizam a fusão de imagens com melhor resolução espacial e temporal [7]. Entre os algoritmos de classificação mais utilizados está o *Random Forest* um algoritmo baseado em aprendizado de máquina (*machine learning)* que realiza as predições e classificações através de um conjunto de árvores de decisão independentes [8].

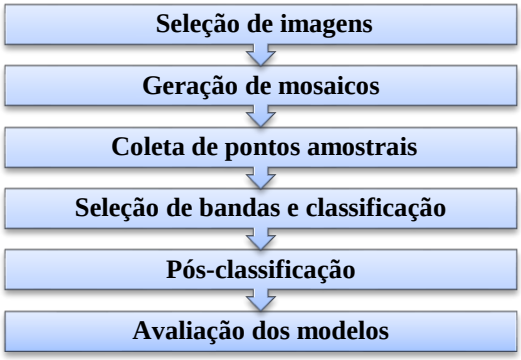


Figura 2. Fluxograma das etapas de execução da metodologia

Os dados foram processados dentro da plataforma *Google Earth Engine* (GEE) utilizando o catálogo de imagens Sentinel-2B, com resolução espacial de 10 metros, de junho e julho de 2020 e 2021 com percentual máximo de 30% de nuvens. Nas imagens foram aplicados algoritmos de remoção de nuvens (*Cloud detector*) de acordo com a probabilidade de ocorrência destas. A escolha dos meses de junho e julho se justifica devido a menor incidência de cobertura de nuvens, pois é período de menor índice pluviométrico e menor atividade convectiva [9].

Após a geração dos mosaicos anuais foram coletadas amostras de treinamento e validação para as seguintes classes de uso e cobertura: agricultura, hidrografia, mosaico de ocupações, campo/savana e floresta. Para classificação dos mosaicos, foi executado o classificador *Random Forest* e visando o melhor desempenho computacional e qualidade da classificação foi determinado o uso de 50 árvores.

Para os resultados de classificação foram obtidas amostras de teste e validação, conforme Tabela 2.

Classe	Nº de amostras			
	Treinamento		Validação	
	2020	2021	2020	2021
1	38.282	38.214	12.969	16.881
2	2.852	2.852	4.063	4.063
3	6.038	6.168	869	869
4	121.218	121.972	83.558	73.589
5	5.615	5.626	6.597	6.597

Tabela 2. Nº de amostras para cada classe.

Na etapa de coleta de amostras os dados foram divididos em dois conjuntos: um pra treinamento e outro para validação. Esta divisão foi realizada por meio de um gerador de números pseudoaleatórios (PRNG - *Pseudo Random Number Generator*) também conhecido como gerador de bits aleatórios determinísticos (DRBG - *Deterministic Random Bit Generator*), onde é gerada uma sequência de números não inteiramente aleatórios a partir de uso de um valor inicial ou também chamado de “semente”

(*seed*) [10]. A repetição do mesmo valor inicial permite a reprodução da sequência e, portanto, aprimora a padronização das amostras. A definição das amostras ocorreu através de inspeção visual com base em aspectos texturais e tonais, além da utilização de mapas do TerraClass e MapBiomas como referência.

No mapeamento das classes de agricultura, hidrografia, mosaico de ocupações, campo/savana e floresta são geradas variáveis de entrada baseadas nas composições de imagens, conforme Tabela 1.

Variável	Estatística	Acrônimo
BLUE	Mediana	blue
GREEN	Mediana	green
RED	Mediana	red
RED EDGE 1	Mediana	red_edge_1
NIR	Mediana	nir
SWIR 1	Mediana	swir1
SWIR 2	Mediana	swir1

Tabela 1. Variáveis das características de entrada

3. RESULTADOS E DISCUSSÕES

Após a classificação foram aplicados dois filtros espaciais: 1) Filtro Majoritário - substitui o pixel alvo baseado na moda dos três pixels em seu entorno e 2) Filtro de Área – que identificou os pixels conectados da classe agricultura que totalizavam no mínimo cinco hectares e seu valor foi substituído pela classe majoritária.

Para análise da área de estudo foi realizada uma adaptação da legenda do Projeto TerraClass, incluindo a classe campo/savana pois esta não pode ser diferenciada (Tabela 3):

Classe	Área (km²)	
	2020	2021
Agricultura (1)	14.35	16.31
Hidrografia (2)	6.75	4.98
Mosaico de Ocupações (3)	4.71	10.91
Campo/savana (4)	371.5	364.65
Floresta (5)	29.47	30.54

Tabela 2. Área das classes detectadas

É possível observar que a classe de campo/savana teve sua área reduzida nos anos de 2020 e 2021, o que pode justificar é que houve conversão de campo/savana para agricultura (visualiza-se no crescimento da área na tabela 2). Nas amostras de validação, o *RF* entendeu que a área ainda era campo natural e identificou como boa parte de campo/savana:

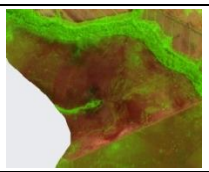
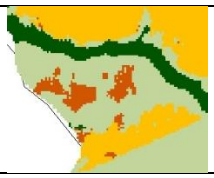
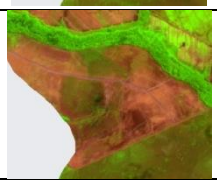
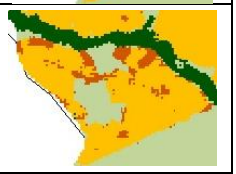
Ano	Falsa-cor	Classificação final
2020		
2021		

Tabela 3. Detalhe de conversão de campo/savana para Agricultura.

Para melhor compreensão da conversão temporal ocorrida da classe de campo/savana para agricultura, as amostras das feições foram coletadas em composição de falsa-cor: R (B11 – SWIR 1), G (B08 – NIR infravermelho próximo) e B (B04 – Red), conforme Tabela 3. A composição de falsa-cor também auxiliou a identificar uma confusão ocorrida na classificação da área, onde pixels classificados como mosaico de ocupações no ano de 2020 foram confundidos com a classe agricultura no mesmo ano.

O aumento de área da classe de mosaico de ocupações em 2021 também pode ser explicado pela confusão do classificador que não compreendeu como uma fase de transição e que a área de agricultura ainda estava em fase inicial, ou seja, a modalidade de uso da terra empregada na área e sua escala de trabalho adotada, fez com que não fosse possível a identificação dos diferentes componentes da paisagem.

Com o objetivo de avaliar o desempenho do *Random Forest* e a confiabilidade da classificação, haja vista que não houve validação dos dados em campo, o trabalho utilizou o índice *Kappa*, acurácia global e matriz de confusão:

Para fins de comprovação, na avaliação da classificação (tabela 4) observa-se que no ano de 2020 o índice *Kappa* de confiabilidade é abaixo de 0,90, mas que ainda sim a concordância da classificação é quase perfeita.

2020					
Classe	1	2	3	4	5
1	**38142**	0	38	7	0
2	0	**2852**	0	0	0
3	0	0	**5948**	1	0
4	140	0	52	**121210**	5
5	0	0	0	0	**5610**
Kappa	0.83				
Ac. Global	0.93				
2021					
Classe	1	2	3	4	5

1	**38171**	0	8	7	0
2	0	**2852**	0	0	0
3	2	0	**6132**	0	0
4	41	0	28	**121962**	29
5	0	0	0	0	**5597**
Kappa	0.91				
Ac. Global	0.96				

Tabela 4. Avaliação do modelo

A conversão de campo/savana para agricultura pode ser explicada através da matriz de confusão no ano de 2020, a classe 4 foi a única que obteve pixels da classe 1. Na classificação, as classes 1, 3 e 5 foram interpretadas pelo *RF* como campo/savana. Essa distribuição de pixels entre as classes pode ser explicada pela similaridade espectral.

No ano de 2021, é possível observar que o nº de pixels da classe 1 aumentou, e o número de pixels associados às outras classes foram reduzidos. Essa alteração pode ser confirmada pelo aumento do indice *Kappa* e acurácia global. A única classe que não teve distribuição de seus pixels as outras classes foi a classe 2, que corresponde a Hidrografia.

5. CONCLUSÕES

A avaliação do desempenho do *Random Forest* para detectar classes de uso e cobertura da terra, no município de Salvaterra de junho e julho dos anos de 2020 e 2021 foi considerada muito satisfatória. O uso de amostragem estratificada fez com que houvesse a divisão de classes razoavelmente homogêneas, sendo assim, com o aumento na quantidade de amostras, maiores as chances dos pixels parecerem entre si, e, portanto, maior a precisão da classificação.

Ainda que o trabalho tenha apresentado bons resultados na classificação realizada através do algoritmo *Random Forest* na plataforma *Google Earth Engine*, sugere-se que haja a validação dos dados em *in loco*, no intuito de melhorar a precisão na detecção de classes de uso e cobertura do solo e aprimoramento dos procedimentos metodológicos.

AGRADECIMENTOS

Os autores agradecem o apoio do Programa de Capacitação Institucional, uma parceria INPE e CNPq, através do Processo Nº 444327/2018-5 e Processo individual Nº 300349/2022-0.

8. REFERÊNCIAS

[1] MALHI, Y. *et al*. *Climate change, deforestation, and the fate of the Amazon*. Science, v. 319, n. 5860, p. 169–172, 2008.

[2] ROSSETTI, D. F.; TOLEDO, P. M. *Biodiversity from a historical geology perspective: A case study from Marajo Island, lower Amazon*. Geobiology, v. 4, n. 3, p. 215–223, 2006.

[3] COUTINHO, A. C. *et al. Uso e cobertura da terra nas áreas desflorestadas da Amazônia Legal: TerraClass 2008*. Brasília, DF: Embrapa; São José dos Campos: INPE, 2013.

[4] IBGE. *Manual Técnico da Vegetação Brasileira*. Rio de Janeiro, IBGE, 2012.

[5] FURTADO A. M. M. *et al. Distribuição espacial das manchas de savana parque no município de Salvaterra, Ilha de Marajó, Pará*. IV Simpósio Nacional de Geomorfologia. Goiânia, 2006.

[6] EUROPEAN SPACE AGENCY. Sentinel-2. *Copernicus*, 2021b. Disponível em: <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>. Acesso em: 29 ago. 2022

[7] GAO, F. *et al*. *On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance*. Ieee Transactions on Geoscience and Remote Sensing, v.44, p. 2207 -2218, 2006.

[8] BREIMAN, L. *Random Forests*. Machine Learning, Amsterdam, v. 45, p. 5–32, 2001.

[9] FISCH, G. *et al. Uma revisão geral sobre o clima da Amazônia*. Acta Amazônica, v. 28, n. 2, p. 121-126, 1998

[10] ROSSETTI, R. *et al*. *Técnicas de Concepção de Algoritmos: Algoritmos Aleatórios*. FEUP-MIEIC-CAL, Porto, 2010/2011.