Análise de dados estatísticos dos satélites Sentinel 1 e 2 para classificação de cobertura e uso do solo na região Amazônica.

Luís Nascimento[1], Rogério Batista[2] e Victor Barros[3]
Laboratório Associado de Computação e Matemática Aplicada (LABAC)
Instituto Nacional de Pesquisas Espaciais(INPE)
São José dos Campos, Brasil
{luis.esnascimento, rogbatista, vicssb}@gmail.com

***Resumo*—As informações de uso e cobertura da Terra são de elevada importância para uma sociedade bem estrutura, uma vez que auxiliam órgãos de gestão ambiental e desenvolvimento agropecuário a identificar regiões de desmatamento e avanço de urbanização, bem como áreas naturais que devem ser protegidas. O projeto MapBiomas, utiliza-se de imagens de satélite e da tecnologia da informação para processar e classificar a cobertura da Terra, informando e disponibilizando as transformações no território brasileiro anualmente. A versão mais recente disponibilizada pelo projeto, denominada coleção 6, utiliza-se de dados estatísticos em diversas frequências de ondas, captadas pelos instrumentos do satélite LANDSAT, e produz uma classificação com acurácia aproximada de 97% para a região de estudo.**
Este trabalho visa a análise exploratória de atributos estatísticos dos satélites Sentinel-1 e Sentinel-2 afim de analisar a possibilidade de classificação do uso e cobertura da Terra, em uma região especifica, através de uma Rede Neural Artificial (RNA), utilizando dados da coleção 6 do MapBiomas como parâmetro de referência o que possibilitaria o monitoramento das mudanças em um período de tempo mais curto.

***Palavras-chave*—Análise Exploratória, Sentinel-1, Sentinel-2, MapBiomas**

I. Introdução

Desde 2015 o projeto MapBiomas, uma rede colaborativa formada por ONGs, universidades e empresas de tecnologias, busca soluções computacionais para a classificação do uso e cobertura da Terra com o objetivo de anualmente divulgar as informações publicamente. O projeto cria classes por biomas, sendo estes: Global, Amazônia, Mata Atlântica, Cerrado, Caatinga, Pantanal e Pampa. Especificamente para o bioma Amazônia, o projeto é capaz de aferir sua cobertura em um nível mais refinado com 96.6%, o que é comparável a uma medida direta e possibilita o uso da informação como medida de validação. Este trabalho analisou dados de dois canais do radar de abertura sintética (SAR) do satélite Sentinel-1, que sofre pouca interferência de nuvens por se tratar de um sensor micro-ondas passivo. Também foram analisados dados de 13 bandas ópticas do satélite Sentinel-2 com resolução espacial entre 10 e 60 metros. O pré-processamento dos dados foi realizado pela Dr. Tahisa Kuck do Sistema de Proteção da Amazônia (SIPAM), nele foram identificados polígonos referentes as classes, e de cada polígono foram extraídas informações estatísticas como a média dos pixels, mediana, desvio padrão e outros, totalizando 12 variáveis estatísticas para cada canal, ou 180 variáveis estatísticas ao todo. Cada amostra foi identificada com sua respectiva classe de acordo com dados do projeto MapBiomas e cedido para este estudo em formato *Comma-separeted values* (CSV). A área de estudo está localizada no estado do Mato Grosso, dentro do bioma Amazônia e foram analisados dois períodos, o primeiro entre os dias 28 e 31 de maio de 2020 e o segundo entre os dias 07 e 08 de outubro de 2020. A partir dos canais disponibilizados foram produzidas outras 3 variáveis muito utilizadas na classificação de uso do solo, *Normalized Ratio Procedure between Bands* (NRPB), *Normalized Difference Vegetation Index* (NDVI) e *Normalized Difference Moisture Index* (NDMI) e analisado a correlação das demais variáveis com estas. Também foram analisados parâmetros como *outliers* e duplicidade de informação. O resultado mostrou-se satisfatório para algumas variáveis candidatas a parâmetros de *input* em metodologias de Inteligência Artificial (IA) como as RNAs.

II. Dados

Segundo [3] o SAR abordo do satélite Sentinel-1 é um instrumento eficaz para monitoramento do uso do solo devido as suas resoluções espaciais e temporais, que podem ser observadas na TabelaI. O SAR a bordo do Sentinel-1 pode operar em dois modos, sendo eles o modo de polarização única (HH ou VV) ou de dupla polarização (HH+VV ou VV+VH). Neste estudo foi utilizado dados em polarização única coletados nos dias 28 de maio de 2020 e 07 de outubro de 2020. Além do Sentinel-1 também foram utilizados dados de sensores ópticos *Multi-Spectral Imager* (MSI) do satélite Sentinel-2, que podem ser observados na TabelaII. As datas referentes aos dois satélites não são equivalentes, mas próximas com não mais de 3 dias, o que para o problema em questão de uso do solo é pouco provável que tenha ocorrido uma mudança significativa, sendo as datas observadas pelo Sentinel-2 os dias 31 de maio de 2020 e 08 de outubro de 2020.

Para cada canal de cada um dos satélites foram disponibilizados 12 variáveis estatísticas conforme a TabelaIII, totali-

Tabela I
SAR/SENTINEL-1

Resolução	$20 * 5m^2$
Swath	$20 * 20km^2$ a cada 100 km
Polarização	simples (VH/VV)

Fonte: space.oscar.wmo.int

Tabela II
MSI/SENTINEL-2

Canal	Resolução (m)	Comprimento de onda (nm)
B1	60	443
B2	10	490
B3	10	560
B4	10	665
B5	20	705
B6	20	740
B7	20	783
B8	10	842
B8A	20	865
B9	60	945
B10	60	1375
B11	20	1610
B12	20	2190

Fonte: space.oscar.wmo.int

zando 180 valores observados para cada polígono identificado na região de estudo.

Tabela III
VARIÁVEIS ESTATÍSTICAS

Variável	Descrição
Count	Número de pixel no polígono
Sum	Somatório dos valores de pixel
Mean	Média dos valores de pixel
Median	Mediana dos valores de pixel
StDev	Desvio padrão dos valores de pixel
Min	Mínimo entre os valores de pixel
Max	Máximo entre os valores de pixel
Range	Valor (Max-Min) dos valores de pixel
Minority	Valor do pixel menos representativo
Majority	Valor do pixel mais representativo
Variety	Número de valores de pixel distintos
Variance	Variância dos valores de pixel

Fonte: mapbiomas.org

Além das informações descritas acima cada um dos polígonos foi classificado com o algoritmo do projeto MapBiomas, que na região de estudo limitou-se a 9 classes, conforme a Tabela IV.

Todas essas variáveis foram cedidas pela Dr. Tahisa Kuck da Divisão de Sensoriamento Remoto do Sistema de Proteção da Amazônia (SIPAM), onde a partir destes foram produzidas 4 outras variáveis, utilizando-se os valores de média dos demais parâmetros, relacionadas ao uso do solo e cobertura da Terra que são o NRPB(1) e o NDVI(2) do Sentinel-1 conforme [1]:

$$NRPB = \frac{VH - VV}{VH + VV} \tag{1}$$

$$NDVI = 2.57 - 0.05 * VH + 0.17 * VV + 3.42 * NRPB \tag{2}$$

Tabela IV
CLASSES IDENTIFICADAS NA REGIÃO DE ESTUDO

ID	Classe	Grupo
3	Formação Florestal	Floresta
4	Formação Savânica	Floresta
11	Campo Alagado e Área Pantanosa	Formação Natural não Florestal
12	Formação Campestre	Formação Natural não Florestal
15	Pastagem	Agropecuaria
24	Área Urbanizada	Área não Vegetada
33	Rio, Lago e Oceano	Corpo D'água
39	Soja	Agropecuaria
41	Outras Lavouras Temporárias	Agropecuaria

Fonte: mapbiomas.org

E os índices NDVI(3) e NDMI(4) do Sentinel-2 conforme [2]:

$$NDVI = \frac{B8 - B4}{B8 + B4} \tag{3}$$

$$NDMI = \frac{B8A - B11}{B8A + B11} \tag{4}$$

Com isso o Conjunto de dados disponível para análise passou a ter 184 variáveis associados a 1 classe totalizando um total de 180.128 amostras somando os 2 períodos disponíveis.

III. ANÁLISE EXPLORATÓRIA DOS DADOS

A. Junção dos conjuntos de dados de entrada

Os dados do Sentinel-1 e do Sentinel-2 foram disponibilizados em arquivos separados, assim como, os dados do primeiro e do segundo período, totalizando 4 arquivos CSV com 90.064 amostras cada, porém cada amostra em cada arquivo foi identificada com um identificador numérico único (fid), o que possibilitou a união das amostras dos diferentes satélites de um mesmo período. Após a junção dos conjuntos de dados dos diferentes satélites, foi realizada também a junção das amostras dos dois períodos resultando em um arquivo único com 180.128 amostras.
Conforme a 1, não houve alteração na classificação dos poligonos entre os perídos.

B. Exclusão de Gaps

Após a junção dos arquivos foi verificada a existência de valores nulos no conjunto de dados, o que ocorreu em 42 linhas que foram removidas do montante, além desses foi identificado também 108 amostras classificadas com a classe "0"que não está prevista na legenda disponibilizada na documentação do projeto MapBiomas e que também foram removidas.

C. Análise do balanceamento dos dados

Conforme observado na Figura2 as amostras classificadas com a classe 15 (Pastagem) correspondem a 84.940, cerca de 47% dos dados, também a classe 41 (Outras Lavouras Temporárias) correspondem a 50.904 amostras, cerca de 28% e ambas pertencem ao mesmo grupo (Agropecuária), ou seja, somados as duas correspondem a 75% do conjunto de dados

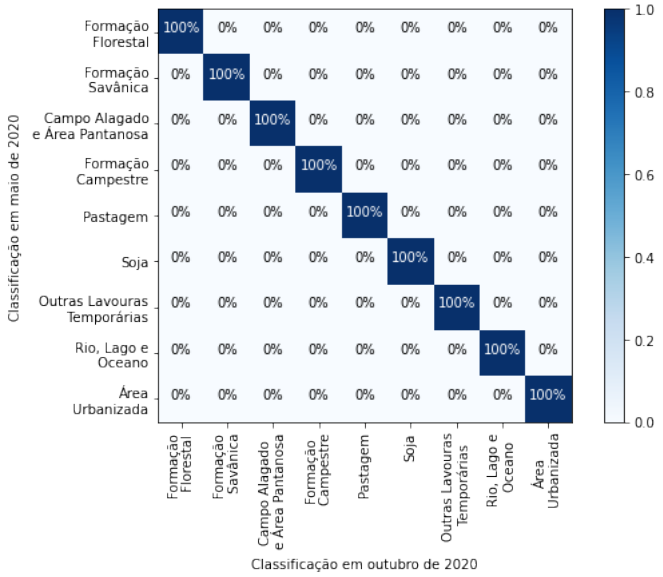

Figura 1. Mudança de classificação entre o período de maio a outubro de 2020

enquanto a classe 24 (Área Urbanizada) corresponde a apenas 50 amostras ou aproximadamente 0.03% do total o que nos diz que o conjunto de dados é desbalanceado, e se nenhum tratamento for realizado provavelmente uma técnica de IA como uma RNA produza classes tendenciosas as classes 15 e 41 e mesmo assim apresente uma boa acurácia, o que não é o objetivo, porém antes de balancear o conjunto de dados outras questões como a análise de correlação e a existência de outliers serão avaliadas.

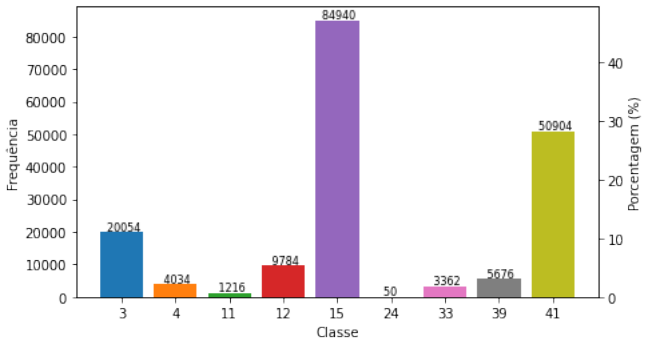

Figura 2. Distribuição dos dados no conjunto de dados

D. Análise de correlação

O número excessivo de variáveis prejudica uma análise de correlação entre cada uma delas, por isso as variáveis estatísticas dos satélites Sentinel-1 e Sentinel-2 sem outliers foram correlacionadas com os índices S1_NRPB, S1_NDVI, S2_NDVI e S2_NDMI, observando as variáveis com correlação superior a 50%. Esta informação associada aos gráficos de histograma e boxplot foram fundamentais para a análise e escolha das variáveis candidatas a input em um classificador. Nesta etapa nenhuma variável apresentou correlação com os índices NRPB e NDVI do Sentinel-1, porém várias delas apresentaram correlação com os índices NDVI e NDMI do Sentinel-2.

E. Exclusão das variáveis

As variáveis que não apresentaram correlação com os índices de vegetação e não se mostraram promissoras nos gráficos de histograma foram excluídas do conjunto de dados. Dentre as que sobraram, Min, Minority e Majority, apresentaram grande parte das informações equivalentes, sendo assim as variáveis Minority e Majority também foram excluídas, bem como Count, Variety e Range restando um total de 50 variáveis candidatas a input em um processo de classificação. Conforme a tabela V

F. Outliers

Quando há disponível um grande número de amostras é comum a exclusão dos outliers, porém nesse caso, como existe um desbalanceamento dos dados, uma exclusão sem a devida análise pode reduzir ainda mais ou eliminar definitivamente classes com poucas amostras, assim neste estudo as classes 4, 11, 24 e 33 foram mantidas sem alterações e as demais tiveram os outlies excluídos.

G. Balanceamento dos dados

Após as etapas acima terem sido concluídas, conforme Figura3 é possível perceber que a quantidade de informação no conjunto de dados reduziu significativamente, contudo o conjunto de dados permanece desbalanceado e para minimizar a diferença, como não é viável a obtenção de mais amostras foi aplicado uma técnica de undersampling, que consiste em retirar amostras aleatoriamente das classes com um maior número, afim de melhorar o balanceamento das informações.

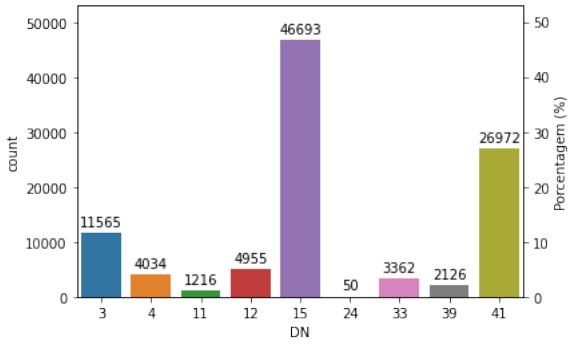

Figura 3. Distribuição dos dados no conjunto de dados após a remoção das outliers

IV. Conclusão

Considerando a distribuição do conjunto de dados após as correções, uma diferença elevada entre as classes ainda permanece, o que indica a necessidade de obtenção de mais amostras das classes em menor quantidade, esse desbalanceamento pode prejudicar, por exemplo a separação das classes

Tabela V
SELEÇÃO DE VARIÁVEIS

Variável estatística	Instrumento	Parâmetro
Média	S1	VV
	S1	VH
	S1	NRPB
	S1	NDVI
	S2	B1
	S2	B2
	S2	B3
	S2	B4
	S2	B5
	S2	B8
	S2	B8A
	S2	B10
	S2	B11
	S2	B12
	S2	NDVI
	S2	NDMI
Mediana	S1	VV
	S1	VH
	S2	B1
	S2	B2
	S2	B3
	S2	B4
	S2	B5
	S2	B8
	S2	B8A
	S2	B10
	S2	B11
	S2	B12
Desvio Padrão	S2	B2
	S2	B3
	S2	B4
Mínimo	S2	B1
	S2	B2
	S2	B3
	S2	B4
	S2	B5
	S2	B8
	S2	B11
	S2	B12
Máximo	S2	B1
	S2	B2
	S2	B3
	S2	B4
	S2	B5
	S2	B8A
	S2	B11
	S2	B12
Variância	S2	B2
	S2	B3
	S2	B4

Fonte: Autores

39 (soja) e 41 (Outras Lavouras Temporárias) já que elas pertencem ao mesmo grupo (Agropecuária) e possivelmente tenham uma resposta espectral semelhante, contudo a classe 24 (Área Urbanizada) que tem o menor número de amostras pertence a um grupo (Área não Vegetada) diferente de todos os demais, assim como sua resposta espectral, o que pode vir a ser uma característica determinante para que a RNA aprenda a classifica-la mesmo com poucas amostras. Nossa conclusão é que os dados disponibilizados apresentam correlação com o objeto de estudo e podem ser utilizados para classificação, porém a baixa quantidade de determinadas amostras pode vir a prejudicar a acurácia do classificador.

REFERÊNCIAS

[1] Roberto Filgueiras, Everardo Chartuni Mantovani, Daniel Althoff, Elpídio Inácio Fernandes Filho, and Fernando França da Cunha. Crop ndvi monitoring based on sentinel 1. *Remote Sensing*, 11(12), 2019.

[2] Josef Lastovicka, Pavel Svec, Daniel Paluba, Natalia Kobliuk, Jan Svoboda, Radovan Hladky, and Premysl Stych. Sentinel-2 data in an evaluation of the impact of the disturbances on forest vegetation. *Remote Sensing*, 12(12), 2020.

[3] Dipankar Mandal, Vineet Kumar, Debanshu Ratha, Subhadip Dey, Avik Bhattacharya, Juan M. Lopez-Sanchez, Heather McNairn, and Yalamanchili S. Rao. Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 sar data. *Remote Sensing of Environment*, 247:111954, 2020.