TRATAMENTO DE DADOS DE SENSORIAMENTO REMOTO PARA A CLASSIFICAÇÃO DO USO E COBERTURA DA TERRA COM MODELOS DE INTELIGÊNCIA ARTIFICIAL

Glauber José Vaz [1,3], Victor Pedroso Curtarelli[2], João Francisco Gonçalves Antunes[1], Alexandre Camargo Coutinho[1], Júlio César Dalla Mora Esquerdo[1], Anderson Rocha[3]

[1]Embrapa Agricultura Digital, Campinas/SP, Brasil, {glauber.vaz, joao.antunes, alex.coutinho, julio.esquerdo}@embrapa.br;
[2]Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH., Campinas/SP, Brasil, victor.curtarelli@giz.de;
[3]Universidade Estadual de Campinas (UNICAMP), Recod.ai, Instituto de Computação, Campinas/SP, Brasil, anderson.rocha@unicamp.br

RESUMO

Este trabalho propõe uma metodologia para o tratamento de dados de sensoriamento remoto obtidos a partir do Brazil Data Cube, visando melhorar a qualidade das séries temporais utilizadas no treinamento de modelos de inteligência artificial para a classificação do uso e cobertura da terra. O estudo utilizou dados coletados na região amazônica da bacia do Alto Paraguai. Três cenários foram analisados: i) sem tratamento; ii) com tratamento das séries temporais; e iii) com tratamento das séries temporais e remoção das amostras não representativas. No último caso, a metodologia possibilita equilibrar o nível de qualidade exigido com a quantidade de amostras descartadas. A metodologia resultou em um ganho significativo na qualidade dos dados.

Palavras-chave — Brazil Data Cube**,** qualidade de dados, Sentinel-2, séries temporais, uso e cobertura da terra.

ABSTRACT

This work proposes a methodology for processing remote sensing data obtained from the Brazil Data Cube, aimed at improving the quality of time series used in training artificial intelligence models for land use and land cover classification. The study used data collected from the Amazon region of the Upper Paraguay Basin. Three scenarios were analyzed: i) without processing; ii) with time series processing; and iii) with time series processing and the removal of poor samples. In the latter case, the methodology makes it possible to balance the quality level with the number of removed samples. The methodology resulted in a significant improvement in data quality.

Key words — *Brazil Data Cube, data quality, Sentinel-2, time series, land use and land cover.*

1. INTRODUÇÃO

O mapeamento do uso e cobertura da terra é importante para muitas aplicações, incluindo gestão de recursos naturais e do meio ambiente, planejamento urbano e do uso da terra, conservação de biodiversidade e promoção da saúde [1, 2]. Entre os desafios para melhorar o mapeamento em larga escala via técnicas computacionais e de inteligência artificial, Zhang e Li [1] destacam a obtenção de amostras de treinamento em maior quantidade e com qualidade. No entanto, segundo os autores, há limitações em todos os métodos normalmente utilizados, como levantamentos de campo, interpretação visual e conjuntos de dados anotados existentes. Outro desafio é a inconsistência e a variabilidade dos conjuntos de dados no tempo, espaço, formatos e nas classes de interesse consideradas no mapeamento.

No Brasil, MapBiomas [3] e TerraClass [4] são importantes iniciativas que geram mapas de uso e cobertura da terra, cada uma com suas especificidades, metodologias e focos. Ambas contam com modelos de inteligência artificial para a classificação a partir de dados de sensoriamento remoto. Embora o aprendizado de máquina seja amplamente aceito na comunidade de sensoriamento remoto para aplicações como esta, observar a qualidade dos dados e a quantidade de amostras é fundamental [5].

O Brazil Data Cube (BDC) disponibiliza dados para análise a partir de grandes volumes de imagens de sensoriamento remoto para todo o território nacional, e uma das principais aplicações destes dados é o mapeamento do uso e cobertura da terra [6]. O BDC fornece diversas coleções de cubos de dados. A S2-16D-2, por exemplo, é baseada em imagens Sentinel-2/MSI de 10 m de resolução espacial em composições de 16 dias, combinação que a torna adequada para o mapeamento em larga escala. Essa coleção envolve vários produtos, incluindo faixas espectrais, índices e dados de qualidade do pixel.

Considerando-se que é muito comum ocorrer problemas com dados de sensoriamento remoto devido a, por exemplo, ruídos, cobertura de nuvens, sombras e distúrbios atmosféricos, este trabalho propõe uma metodologia de tratamento dos dados obtidos a partir do BDC, mais

especificamente da coleção S2-16D-2, para a classificação do uso e cobertura da terra. Também mostra o impacto desse tratamento na qualidade das amostras.

2. MATERIAL E MÉTODOS

Os dados considerados neste trabalho foram obtidos na região amazônica da Bacia do Alto Paraguai (BAP), que concentra 67% do passivo ambiental em reserva legal da bacia [7] e envolve pontos de diferentes características de uso e cobertura da terra. A Figura 1 exibe a localização desta área.

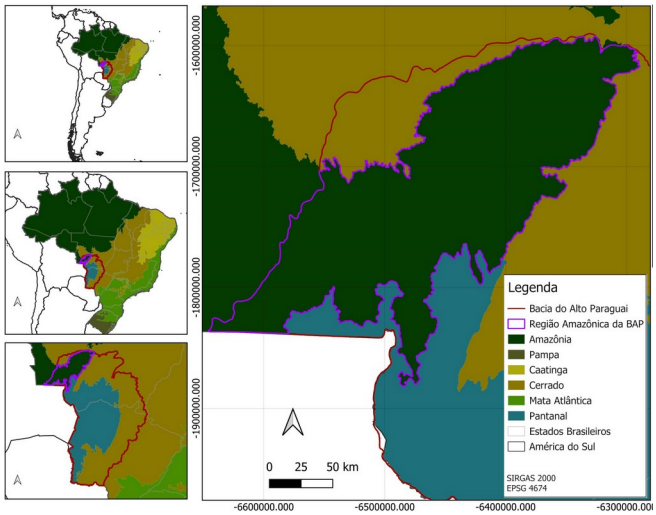

Figura 1. Localização geográfica da região estudada [8].

O conjunto de classes temáticas abordadas neste trabalho foi determinado de maneira que fosse possível compatibilizar as classes dos mapas de uso e cobertura da terra produzidos pelas iniciativas TerraClass e MapBiomas na localidade considerada. As classes são: 'Água', 'Floresta', 'Formação Natural Não Florestal', 'Silvicultura', 'Pastagem', 'Agricultura Temporária de um ciclo', 'Agricultura Temporária de mais de um ciclo', 'Agricultura Semiperene' e 'Urbano'. A determinação desse conjunto de classes e a coleta dos pontos amostrais para treinamento dos modelos foram realizadas conforme uma metodologia que procura obter pontos cuja classificação fornecida pelo TerraClass coincida com a classificação do MapBiomas [8].

A banda CLEAROB da coleção S2-16D-2 do BDC indica o número de boas observações no período, livres de nuvem e suas sombras, neve ou ocorrência de dados não disponíveis. Ela foi utilizada para ajustar as séries temporais obtidas, de maneira a anular observações ruins.

O conjunto de séries temporais pode ser analisado de duas maneiras: com as séries temporais exatamente conforme obtidas do cubo de dados, ignorando-se a banda CLEAROB, ou eliminando-se as observações que não são consideradas boas segundo a banda CLEAROB.

O BDC também provê uma banda SCL (*Scene Classification Layer*), baseada na banda de mesmo nome do Sentinel-2. Os possíveis valores desta banda indicam classificações como 'No Data', 'Vegetation', 'Not Vegetated', 'Water', entre outros. Esta banda foi usada para tratar dados que indicam a presença de água.

A banda CLEAROB foi utilizada para a validação das séries temporais. Foi considerada válida uma observação cujo valor correspondente da banda CLEAROB é maior ou igual a um. Caso contrário, a observação é anulada, atribuindo-se NaN (*not a number*), e, posteriormente, preenchida por interpolação linear. Por exemplo, se uma amostra contém a série temporal t_b=[345, 373, 499, 430, 7622, 1587, 4891, 405] para determinada banda *b*, e a série $t_{CLEAROB}$=[2, 1, 2, 1, 0, 1, 0, 1] para a banda CLEAROB, com o tratamento, a série assume os valores t_b=[345, 373, 499, 430, nan, 1587, nan, 405]. Usando interpolação para substituir os valores 'nan', o resultado é t_b=[345, 373, 499, 430, 1008.5, 1587, 996, 405].

No entanto, algumas séries temporais apresentam muitas observações inválidas, o que compromete bastante a sua qualidade e motiva seu descarte. Para isso, foram analisados dois parâmetros: a quantidade máxima permitida de observações inválidas na série temporal (N_{TOTAL}) e o tamanho da maior sequência de observações inválidas (N_{SEQ}). Para encontrar valores desses parâmetros que mantenham a qualidade das amostras e evitem o descarte excessivo de amostras, foram realizados testes com diferentes combinações de valores para todas as classes temáticas.

Portanto, as séries temporais foram comparadas em três condições: i) séries sem tratamento; ii) séries ajustadas com interpolação de valores nas observações inválidas; e iii) conjunto de séries com exclusão daquelas em que os parâmetros de qualidade não foram atendidos.

3. RESULTADOS

Para cada classe e cada banda, foram analisadas as séries temporais obtidas da coleção S2-16D-2 do BDC. A Figura 2 exibe gráficos com as séries temporais, em cinza, de todos os pontos classificados como 'Agricultura temporária de um ciclo' para as bandas 'B06' e 'NDVI'. Em destaque, aparecem as médias das séries temporais revelando padrões característicos da classe para as bandas analisadas. Cada linha da Figura 2 está associada a uma condição de tratamento das séries temporais. A primeira exibe os dados sem tratamento e a segunda exibe as séries com remoção de observações inválidas e posterior interpolação de valores. A terceira linha contém as séries com o mesmo tratamento da segunda, mas há remoção de séries temporais que não atendem aos parâmetros de qualidade relacionados à quantidade total de observações inválidas e ao maior tamanho de sequência com essas observações. Os valores utilizados para N_{TOTAL} e N_{SEQ} nesses exemplos foram, respectivamente, 5 e 2.

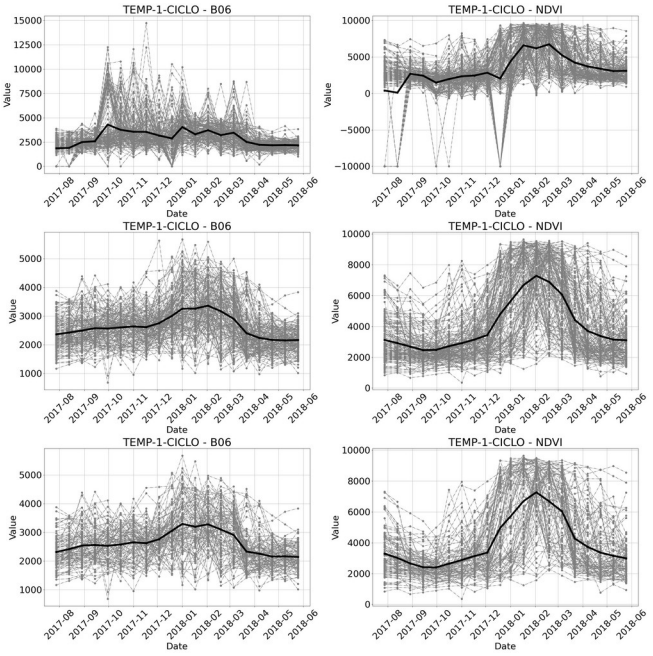

Figura 2. Perfis de séries temporais para duas bandas (colunas) e nas três condições de tratamento (linhas).

O ganho de qualidade na representação da classe nas duas últimas linhas é claro. A segunda coluna, por exemplo, exibe a banda de NDVI, muito utilizada neste tipo de aplicação. O perfil característico de uma cultura de um ciclo está muito bem representada nas duas últimas linhas, o que não ocorre no gráfico da primeira. Porém, a diferença entre a segunda linha e a terceira não é significativa. Também é possível notar a remoção de valores discrepantes que de fato são inválidos.

A Figura 3 mostra, para a mesma classe de agricultura temporária de um ciclo, em oito bandas espectrais, a diferença entre as médias das séries temporais nas três condições descritas. Em cinza, a média sem tratamento; em preto, com interpolação para observações inválidas; e, em vermelho e pontilhado, o tratamento com remoção de amostras. Esta figura torna ainda mais claro o ganho de qualidade dos dados após o tratamento proposto, dada a significativa diferença entre a curva cinza e as outras duas.

Para determinar os valores dos parâmetros N_{TOTAL} e N_{SEQ}, podem ser realizados testes com as amostras usadas. A Tabela 1 exibe a quantidade total de amostras de cada classe e a quantidade de amostras removidas considerando diferentes combinações para os valores desses parâmetros. Por exemplo, para a classe de agricultura temporária de um ciclo, são eliminadas 53 amostras com os parâmetros utilizados na geração da Figura 2 (N_{TOTAL}=5; N_{SEQ}=2).

A determinação dos valores para esses parâmetros precisa levar em consideração o compromisso entre a qualidade dos dados e a quantidade de amostras descartadas. A última coluna mostra que a adoção de parâmetros muito rígidos levam à remoção de muitas amostras, atingindo a marca de até ⅔ do total. Por outro lado, quando se aceitam sequências grandes de valores inválidos ou um total elevado de valores inválidos distribuídos na série, provavelmente, a amostra não representa adequadamente a classe correspondente.

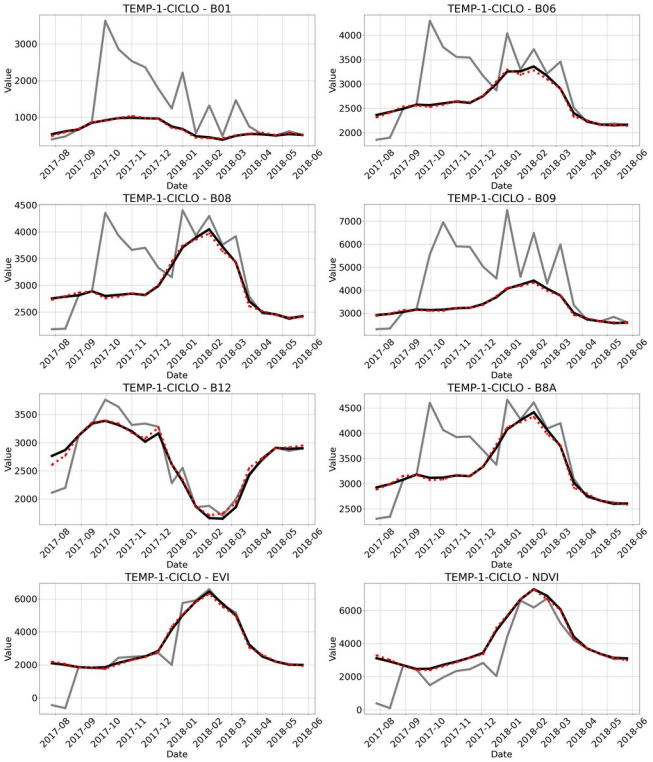

Figura 3. Comparação das médias das séries temporais nas três condições analisadas, para oito bandas.

Classe	#	(9;4)	(9;3)	(8;3)	(7;3)	(5;2)	(3;1)
Floresta	150	15	23	25	30	55	95
Temporária 1 ciclo	150	5	14	16	19	53	100
FNNF	153	21	35	40	45	61	100
Semiperene	150	4	9	12	22	53	91
Silvicultura	150	7	18	21	30	60	99
Temporária +1 ciclo	150	3	10	12	14	40	101
Urbano	123	12	17	18	28	53	85
Pastagens	150	8	17	18	22	50	87
Água	150	118	132	134	135	142	147

Tabela 1. Parâmetros de qualidade (N_{TOTAL}; N_{SEQ}) e a quantidade de amostras eliminadas por classe.

No exemplo ilustrado pelas Figura 2 e 3, os parâmetros são restritos demais, pois mais de 1/3 das amostras foram eliminadas e a representação da classe continua muito parecida. Isso pode ser verificado comparando-se os gráficos das duas últimas linhas da Figura 2 e as curvas pretas e vermelhas dos gráficos da Figura 3.

A Tabela 1 também revela que uma grande quantidade de amostras da classe ‘Água’ apresentou qualidade ruim conforme a banda CLEAROB. Então, um tratamento diferente foi realizado para amostras dessa classe.

Especificamente para a classe ‘Água’, os parâmetros N_{TOTAL} e N_{SEQ} são ignorados e a banda SCL é utilizada para determinar se a amostra é válida. Considerando-se:

- S_{WATER}: Quantidade de elementos da série classificados como ‘Water’ (código 6) na banda SCL e com valor maior do que zero na banda CLEAROB.
- S_{OTHER}: Quantidade de elementos da série classificados com valor diferente de ‘Water’ na banda SCL e com valor maior do que zero na banda CLEAROB.

Toda amostra em que a condição ($S_{WATER} > S_{OTHER}$) é satisfeita pode ser classificada como ‘Água’. Do total de 150 amostras, 53 são removidas, um valor ainda significativo, mas muito menor do que os exibidos na Tabela 1.

Adicionalmente, foi verificado que nenhuma amostra das outras classes foi classificada como ‘Água’ por este método. Portanto, uma metodologia que executa a classificação de água separadamente e antes das demais classes é uma alternativa interessante para esses dados.

4. DISCUSSÃO

Este estudo foi conduzido com dados de uma região específica, no bioma amazônico da BAP. Os dados foram coletados manualmente usando-se mapas gerados pelos projetos MapBiomas e TerraClass. Para se obter um conjunto maior de dados ou para grandes áreas, é importante traçar uma estratégia que vise a ganho de escala.

Para outras regiões do país, a validação da metodologia proposta pode exigir outras abordagens de tratamento. Por exemplo, em regiões com maior incidência de nuvens, as séries temporais apresentam mais observações inválidas. Métodos alternativos também podem ser necessários, como foi possível observar no caso de amostras da classe ‘Água’ deste estudo.

As amostras dessa classe apresentam diferenças significativas das demais em relação à qualidade dos dados conforme a banda CLEAROB da coleção usada. Este caso específico pode ser objeto de maior investigação.

A banda CLEAROB é de grande relevância na avaliação da qualidade das séries temporais obtidas do BDC. A banda SCL, que, em princípio, não era considerada relevante para o contexto deste trabalho, por se tratar de uma classificação mais geral do que a desejada, revelou-se de grande importância para o tratamento das amostras da classe ‘Água’.

O tratamento das séries temporais com o uso dessas bandas para o ajuste dos valores das observações propicia avanços na qualidade dos dados. Uma fase adicional de remoção de amostras consideradas ruins pode gerar representações ainda melhores das classes envolvidas. Porém, a eliminação dessas amostras deve ser analisada de acordo com a aplicação, o que depende, por exemplo, da quantidade de amostras disponíveis e do ganho efetivo na representatividade das classes. Os valores dos parâmetros que determinam a qualidade não devem ser tão rígidos a ponto de provocar o descarte desnecessário de amostras e nem tão elásticos que descaracterizem as classes que elas representam.

5. CONCLUSÕES

Na busca de maior acurácia em modelos de inteligência artificial para a classificação do uso e cobertura da terra, é importante que os dados de sensoriamento remoto sejam tratados antes de serem efetivamente usados.

A metodologia de tratamento proposta neste trabalho recomenda a utilização de bandas disponibilizadas pelo próprio cubo de dados utilizado para fazer os ajustes das séries temporais de interesse. Para a coleção S2-16D-2 do cubo de dados BDC, a banda CLEAROB é fundamental nesse contexto, assim como a banda SCL no caso em que a classe ‘Água’ é envolvida. Para este caso, indica-se a classificação desta classe antes das outras.

O ganho de qualidade dos dados é significativo quando tratados com a metodologia desenvolvida neste trabalho.

6. REFERÊNCIAS

[1] C. Zhang, and X. Li. Land use and land cover mapping in the era of big data, *Land*, v. 11, 1692, 2022.

[2] M. Arpitha, S. A. Ahmed, and N. Harishnaika. Land use and land cover classification using machine learning algorithms in Google Earth Engine, *Earth Science Informatics*, v. 16, n. 4, p. 3057-3073, 2023.

[3] C. M. Souza Jr. et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine, *Remote Sensing*, v. 12, n. 17, 2020.

[4] TerraClass. 2024. Disponível em: <https://www.terraclass.gov.br>. Acesso em: 23 out. 2024.

[5] A. E. Maxwell, T. A. Warner, and F. Fang. Implementation of machine-learning classification in remote sensing: An applied review, *International Journal of Remote Sensing*, v. 39, n. 9, p. 2784-2817, 2018.

[6] K. R. Ferreira et al. Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products, *Remote Sensing*, v. 12, n. 24, 4033, 2020.

[7] E. Rosa, M. Rosa, M. Dias, T. Azevedo, and J. Shimbo. *Conservação da Planície e do Planalto na Bacia Hidrográfica do Alto Paraguai*, Nota técnica, MapBiomas, 2024.

[8] G. J. Vaz, A. S. Tavares. J. F. G. Antunes, A. C. Coutinho, and J. C. D. M. Esquerdo. Obtenção de dados para o treinamento de modelos de aprendizado de máquina para o mapeamento do uso e cobertura da terra na região amazônica da Bacia do Alto Paraguai. *Simpósio de Geotecnologias no Pantanal (GeoPantanal)*, 8, 2024, Poconé - MT. Atigo a ser apresentado.