Uso de Embeddings para Detecção de Distúrbios Florestais

Carla Almeida

Instituto Nacional de Pesquisa Espaciais (INPE)
Avenida dos Astronautas, 1758, Jardim da Granja, São José dos Campos -SP -Brasil

carla.paula@inpe.br

Abstract. *This work presents a methodological approach for the automated structuring and extraction of spectral attributes (Embeddings), aiming at the classification of forest disturbances in the Amazon using the Random Forest algorithm. The methodology consisted of selecting polygons in a GIS environment (QGIS) from data from the AMAZONIA-1 satellite. The vectors were integrated into the Google Earth Engine (GEE) platform.*

Resumo. *Este trabalho apresenta uma abordagem metodológica para a estruturação e extração automatizada de atributos espectrais (Embeddings), visando a classificação de distúrbios florestais na Amazônia com o algoritmo Random Forest. A metodologia consistiu na seleção de polígonos em um ambiente SIG (QGIS) a partir de dados do satélite AMAZONIA-1. Os vetores foram integrados à plataforma Google Earth Engine (GEE).*

1. Introdução

A Amazônia é o maior bioma brasileiro, abrigando a maior biodiversidade do planeta com uma vasta variedade de espécies de animais, plantas e microrganismos. Por essa razão, a preservação desse ecossistema é fundamental para garantir a sustentabilidade climática e ambiental em âmbito global. Todavia, a integridade da região tem sido severamente ameaçada pelo avanço do desmatamento, um processo intensificado a partir das políticas de integração e expansão territorial implementadas na década de 1970 [1].

Diante desse cenário, os sistemas de monitoramento por satélite tornaram-se ferramentas de extrema relevância. Um desses sistemas é o DETER, implementado em 2004, que atua no mapeamento ágil da supressão e degradação florestal na Amazônia Legal Brasileira (ALB), bem como na perda de vegetação primária nas formações florestais [2]. O sistema emite alertas diários de alterações na vegetação nativa,

caracterizadas como distúrbios. Distúrbios definem-se como qualquer evento de origem natural (como deslizamentos) ou antrópica (como queimadas e corte raso) que altere a estrutura biológica e a dinâmica da vegetação em diferentes níveis ecológicos.

Para garantir a precisão dos alertas e mitigar a ocorrência de falsos positivos, o DETER utiliza uma máscara de exclusão. A máscara de exclusão é uma camada digital, em formato vetorial ou *raster*, sobreposta às imagens orbitais para indicar as áreas que devem ser ignoradas na análise corrente, por já terem sido descaracterizadas anteriormente. A partir disso, o monitoramento das mudanças na cobertura vegetal baseia-se na análise visual de imagens de satélite, examinando características como cor, textura e geometria das feições afetadas.

No caso do desmatamento por corte raso, o processo caracteriza-se pela remoção completa da cobertura florestal em um curto período. Na Amazônia, essa dinâmica é planejada de acordo com a sazonalidade climática: inicia-se no final do período chuvoso com a "broca" (corte da vegetação subarbustiva), seguida pela derrubada das árvores de grande porte no início da estação seca, que ocorre majoritariamente entre os meses de julho e setembro. Como consequência, esse distúrbio antrópico resulta na completa exposição do solo, deixando cicatrizes altamente evidentes e de fácil detecção nas imagens orbitais.

Todo esse fluxo operacional de processamento e interpretação é executado no ambiente *TerraAmazon*, um *software* desenvolvido pelo Instituto Nacional de Pesquisas Espaciais (INPE) e utilizado ativamente por sistemas como o DETER e o PRODES. A plataforma integra um conjunto de algoritmos de processamento de imagens e dados vetoriais baseados em técnicas de computação de alto desempenho, o que viabiliza o processamento de grandes volumes de dados geoespaciais em tempo hábil [3].

Contudo, a crescente demanda por monitoramento em tempo real impõe a necessidade de otimizar essas detecções, visando reduzir o volume de dados manipulados e aumentar a precisão na identificação dos distúrbios. Nesse contexto, o uso de embeddings (ou vetores de incorporação) surge como uma alternativa promissora. Essa abordagem metodológica consiste na conversão de dados complexos e não estruturados, como imagens de sensoriamento remoto, em uma sequência organizada de números de

menor dimensionalidade. Ao contrário de uma compressão digital convencional, essa transformação preserva os atributos semânticos e as características fundamentais do objeto original. Coletivamente, esses números atuam como coordenadas matemáticas em um espaço multidimensional, onde a proximidade geométrica entre os vetores reflete a similaridade real entre os alvos mapeados.

A geração desses embeddings é viabilizada por meio de [4] Modelos de Fundação (Foundation Models), que atuam como extratores de características espaciais e temporais das imagens orbitais. Um exemplo é o AlphaEarth Foundations, capaz de transformar o dado bruto de satélite nesses vetores estruturados. Posteriormente, algoritmos de aprendizado de máquina supervisionados, como o Random Forest— disponível na plataforma Google Earth Engine (GEE), conseguem ler esses embeddings e classificar eficientemente as diferentes classes de interesse, distinguindo, por exemplo, floresta preservada de cicatrizes de queimada. Como consequência direta dessa redução de dimensionalidade, classificadores tradicionais e computacionalmente mais leves tornam-se capazes de processar as informações com menor demanda de infraestrutura e máxima eficiência preditiva.

2. Metodologia

O fluxo metodológico deste trabalho foi estruturado em três etapas principais: seleção e preparação das amostras espaciais no ambiente SIG; extração automatizada de características (embeddings) através de um Modelo de Fundação e, por fim, classificação supervisionada para a identificação do distúrbio florestal utilizando o algoritmo Random Forest.

A primeira etapa consistiu na coleta de amostras de referência para o treinamento do modelo. Utilizando o software QGIS (versão 3.34 LTR), foi isolada uma grade espacial específica (tile C56L49). Dentro dessa área, foram selecionados 30 polígonos de controle delimitando de forma precisa a classe de desmatamento por corte raso (CR) e áreas de floresta preservada, conforme apresentado na Figura 1.

Ao realizar a triagem dos dados geoespaciais no QGIS, observou-se a distribuição inicial de amostras por sensor: o satélite AMAZONIA-1 contava com 32 feições amostrais, o CBERS-4 com 18 amostras, e o CBERS-4A com 15 amostras. Para este

estudo, optou-se pela seleção exclusiva dos dados associados ao satélite AMAZONIA-1. Essa escolha justificou-se pelo maior volume de amostras disponíveis para o sensor, critério essencial para garantir a robustez estatística durante a etapa de aprendizado do classificador Random Forest e mitigar potenciais erros de generalização do modelo.

Figura 1: Seleção do Tile e Polígonos de Controle no QGIS

Com os vetores estruturados e validados, foi realizado o upload desses arquivos para a plataforma Google Earth Engine(GEE), onde foram convertidos em coleções de feições (FeatureCollections). Essa integração em nuvem foi necessária para permitir o cruzamento ágil e o processamento de alto desempenho entre os polígonos de referência e os volumes massivos de imagens de satélite.

Na etapa final, os embeddings extraídos foram utilizados como atributos de entrada (features) para o algoritmo de classificação supervisionada Random Forest. O classificador foi executado de forma nativa dentro do ambiente de computação em nuvem do GEE para a obtenção do mapa final de distúrbio florestal. Embora a plataforma permita o uso de rotinas em Python em ambientes externos (como via API), todo o desenvolvimento computacional deste trabalho foi centralizado em linguagem JavaScript, utilizando a interface principal (Code Editor) e a documentação nativa da ferramenta.

3. Resultado

A partir do processamento e da extração automatizada de atributos na plataforma [5] Google Earth Engine (GEE), foi gerada uma base de dados estruturada em formato de matriz de atributos (embeddings). Nesta estrutura, cada linha corresponde a um pixel amostral individual, enquanto as colunas representam a variável dependente (alvo) denominada CLASS, seguida pelas variáveis preditoras compostas pelas bandas espectrais do visível e do infravermelho próximo (SR_B2, SR_B3, SR_B4 e SR_B5).

No total, o mapeamento e a amostragem espacial baseados nos polígonos de controle resultaram em 129.188 registros (pixels). A distribuição quantitativa desses dados de entrada está detalhada na Tabela 1.

Classe de Uso do Solo	Quantidade de Dados (Pixels)	Proporção (%)
Floresta	100.722	77,97%
Desmatamento por Corte Raso (CR)	28.466	22,03%
Total Geral	129.188	100,00%

Tabela 1: Distribuição absoluta e proporcional dos dados estruturados por Classe.

A análise da matriz indica uma predominância de dados associados à classe Floresta (77,97%), enquanto os vetores da classe desmatamento por corte raso representam 22,03% do volume amostral. Do ponto de vista computacional, a configuração deste conjunto de dados fornece uma densidade amostral altamente robusta.

No entanto, cabe ressaltar que a elevada densidade e robustez volumétrica deste conjunto de atributos não implicam, necessariamente, em uma garantia antecipada de alta acurácia na classificação final. A capacidade de generalização e a exatidão do algoritmo dependem intrinsecamente da separabilidade estatística dos Embeddings nas regiões de transição entre as classes. Portanto, a eficiência real desta configuração amostral permanece como uma hipótese a ser validada numericamente nas etapas subsequentes do

experimento, por meio da aplicação de métricas rigorosas de validação cruzada e matrizes de erro ajustadas.

4. Discussão

A estruturação da matriz de dados contendo 129.188 registros evidencia o potencial de integração entre ferramentas de ambiente SIG local (QGIS) e o processamento massivo em nuvem (GEE). O desbalanceamento amostral observado (77,97% para Floresta e 22,03% para Desmatamento por corte raso) assemelha-se à distribuição real das classes na paisagem amazônica, o que é computacionalmente desejável para que o classificador Random Forest aprenda a ponderar a probabilidade a priori das ocorrências no território. Discussões futuras deverão focar na avaliação de desempenho do classificador sob diferentes cenários de amostragem.

5. Conclusão

Este estudo viabilizou a modelagem metodológica, a estruturação e a extração automatizada de atributos espectrais a partir de polígonos de referência, permitindo a consolidação de uma base de dados robusta e padronizada para o aprendizado de máquina. A estratégia de centralizar a extração de embeddings no ambiente Google Earth Engine demonstrou-se operacionalmente eficiente para o manejo de volumes massivos de dados geográficos.

Por tratar-se de uma pesquisa em andamento, as inferências sobre a precisão e o desempenho definitivo do algoritmo *Random Forest* permanecem em fase experimental. Os próximos passos deste trabalho preveem a execução de rodadas de classificação e o cálculo de indicadores de acurácia ajustados pela área, além da incorporação de novas amostras espaciais para verificar a sensibilidade do modelo frente a diferentes tipologias de cobertura vegetal. Espera-se, ao término das validações estatísticas, consolidar um fluxo de trabalho metodológico seguro e replicável para o monitoramento de distúrbios e suporte aos alertas de desmatamento na Amazônia.

Referências Bibliográficas

[1] SOUZA, Bruno Campos de *et al*. Análise da complementaridade de alertas de desmatamento baseados em SAR do DETER-R aos alertas ópticos do DETER na

Amazônia. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 21., 2025, Salvador. **Anais [...]**. São José dos Campos: INPE, 2025. p. [1485 - 1487].

[2] FEITOSA, Jeremias Vitório Pinto *et al*. Identificação analítica de classes de desflorestamento com o sensor WFI-Amazonia-1, no projeto DETER-Amazonia. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 20., 2023, Florianópolis. **Anais [...]**. São José dos Campos: INPE, 2023. p. 1834-1836.

[3] INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. Metodologia utilizada nos sistemas PRODES e DETER. 2. ed. São José dos Campos: Dibib, 2022. 50 p. Disponível em: http://urlib.net/8JMKD3MGP3W34T/47GAF6S. Acesso em: 5 jun. 2026

[4] NOTEBOOKLM. **The Mechanics of Earth Observation Foundational Models**: Decoding the shift from task-specific machine learning to universal embeddings. [S. l.]: Google, 2026.

[5] SPATIALTHOUGHTS. **Introdução ao conjunto de dados de incorporação de satélite**. *In*: Google Earth Engine: Comunidade. Google for Developers, 11 out. 2025. Disponível em: https://developers.google.com/earth-engine/tutorials/community/satellite-embedding-01-introduction?hl=pt-br. Acesso em: 5 jun. 2026.

[6] ROCHA, Alby Duarte; GOMES, Alessandra Rodrigues; SADECK, Luis Waldyr Rodrigues; FERREIRA, Karine Reis. Beyond Alerts: spatiotemporal trade-offs in near-real-time detection systems for forest disturbance in the Brazilian Amazon. São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), 2024.

[7] Instituto de Pesquisas Espaciais (INPE). **DETER**: Detecção de Desmatamento em Tempo Real. São José dos Campos: INPE. Disponível em: http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/deter/deter. Acesso em: 5 jun. 2026.