

CAFÉ PARA TODOS

Luís Eduardo Damasceno¹ (IC), Melise Maria Veiga de Paula¹, (PQ)¹Universidade Federal de Itajubá**Palavras-chave:** BDC. Cafeicultura. MongoDB. Sentinel-2A/2B. YCSB.**Introdução**

O café é um dos produtos mais consumidos mundialmente. O Brasil ocupa a liderança na produção global, com Minas Gerais representando 50,8% dessa produção (FAEB, 2024). Nesse contexto, o uso de tecnologias digitais, como sensores remotos e inteligência artificial, pode impulsionar a agricultura, trazendo eficiência e sustentabilidade ao setor (BASSOI et al., 2019). Em particular, os sensores remotos orbitais permitem o monitoramento eficiente das áreas cafeeiras, otimizando o uso de recursos agrícolas, como terra, água e insumos (MOREIRA et al., 2008).

Embora a agricultura digital tenha ampliado a capacidade de aquisição de dados, muitos agricultores ainda enfrentam desafios no acesso e na interpretação de informações agrônômicas (BASSOI et al., 2019).

Este estudo faz parte de projeto de pesquisa que visa reduzir essa lacuna por meio da disponibilização de dados de maneira mais acessível. Para tanto, dados de sensores remotos serão consumidos de uma base de dados em forma de cubo geoespacial, processados com objetivo de atender à cafeicultura mineira e, posteriormente, disponibilizado à cadeia produtiva do café de forma acessível, tanto digitalmente, quanto semanticamente.

A fonte de dados considerada foi o projeto *Brazil Data Cube* (BDC) que é uma iniciativa do Instituto Nacional de Pesquisas Espaciais (INPE) voltada à organização e disponibilização de dados geoespaciais em forma de cubos de dados. Uma das características mais importantes do cubo de dados é que cada pixel é associado a uma série temporal. Dessa forma, é possível acessar o valor do pixel em diferentes datas, o que facilita análises em regiões específicas (GOMES, 2024).

Especificamente, neste estudo, a proposta foi criar uma base de dados com informações espectrais coletadas pelo satélite Sentinel-2A/2B para a região do Sul de Minas Gerais, com foco na cidade de Três Pontas. Os dados espectrais gerados pelo BDC foram processados e armazenados em um banco de dados orientado a documentos, o MongoDB, que é ideal para lidar com grandes volumes de dados não estruturados e temporais, como as imagens de sensoriamento remoto

(HACHIMI, 2022).

O estudo realizado é o primeiro passo para permitir que, tanto pesquisadores, cafeicultores e demais interessados acessem facilmente informações cruciais para o manejo das plantações de café, como a identificação de áreas com pragas ou deficiência de nutrientes (FORMAGGIO, 2017).

Este trabalho se justifica pela necessidade de democratizar o acesso às informações geoespaciais sobre a cafeicultura em Minas Gerais, fornecendo uma ferramenta acessível para que pequenos e médios produtores possam tomar decisões mais informadas e estratégicas. Ao disponibilizar dados espectrais e meteorológicos, por exemplo, de maneira simples e compreensível, espera-se contribuir para o aumento da produtividade e a sustentabilidade na cadeia produtiva do café.

Metodologia

Inicialmente, todos os arquivos de códigos feitos estão no repositório do [Github](#).

A primeira etapa do trabalho foi estudar temas importantes para desenvolvimento do trabalho, tais como geoprocessamento, geotecnologia, sensoriamento remoto e bandas espectrais. Em seguida, foi necessário investigar e definir a estratégia usada para a coleta dos dados. Como a API de acesso ao BDC está em R, essa foi a linguagem utilizada. Para o armazenamento, foi analisada a documentação do MongoDB. Esse servidor de banco de dados se adequa ao projeto uma vez que é nativamente espacial e porque permite o armazenamento otimizado de séries temporais (CABRAL, 2023).

O escopo do projeto foi definido com base na importância socioeconômica da cafeicultura na cidade de Três Pontas, localizada no sul de Minas Gerais, uma das principais regiões produtoras de café do Brasil. Três Pontas se destaca no cenário nacional devido à sua longa tradição no cultivo de café de alta qualidade, o que a torna uma área estratégica, conectada ao mercado global (SOUSA, 2023).

Os dados utilizados no projeto foram obtidos a partir do cubo de dados "S2-16D-2" do satélite Sentinel-2A/2B, que oferece uma resolução espacial de 10 metros por pixel, adequada para detalhar as lavouras

de café. Essa resolução permite capturar informações mais precisas, essenciais para a análise do terreno agrícola (KIM, 2019).

A extração dos dados foi realizada por meio da API STAC do *Brazil Data Cube*, que organiza as informações em quatro entidades principais. A primeira é o *Item*, que contém metadados em formato GeoJSON e links para os *Assets*. Os *Assets* são os arquivos reais que contêm os dados referentes a um espaço e tempo específicos. O *Catalog* conecta os *Itens* e as *Collections*, permitindo a navegação entre as duas entidades. A *Collection* é uma versão especializada do *Catalog*, que disponibiliza detalhes sobre os dados espaciais e temporais (FERREIRA, 2020).

Dentre os dados disponíveis, as bandas espectrais *Red*, *Green*, *NIR*, e o índice de vegetação (IVs) *NVDI* foram escolhidos para compor a primeira versão da base de dados. Essa escolha se justifica pelo fato de que esses dados estão diretamente relacionados às variáveis agronômicas relevantes, podendo auxiliar na análise e tomada de decisões nas lavouras de café.

Foi desenvolvida uma biblioteca chamada "helperSTAC.R" para simplificar o acesso aos dados do Brazil Data Cube (BDC) por meio da API STAC. O funcionamento desta biblioteca é baseado em um processo encapsulado que gerencia de forma eficiente as requisições. Dessa forma, o código detalhado dessa biblioteca está disponível no repositório GitHub.

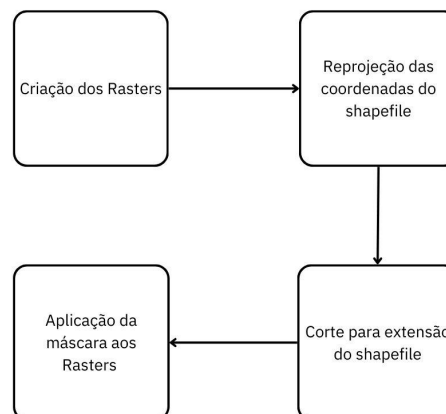
Para a obtenção dos dados, o primeiro passo envolve a conexão para realizar a requisição, que prepara o sistema para acessar os dados disponíveis na plataforma STAC.

Na etapa seguinte, ocorre o resgate dos itens da região de Três Pontas, dentro do período requisitado. Aqui, são obtidos todos os dados para a área de Três Pontas no intervalo de tempo definido por todas as datas disponíveis até 2022. No qual, representa o último ano de dados disponibilizado pelo BDC.

Por fim, é feita a etapa de resgatar os *assets* das bandas requisitadas, extraindo os dados das bandas espectrais e índices de vegetação de interesse, como *Red*, *Green*, *NIR* e *NDVI*.

No tratamento dos dados, foram realizadas 4 etapas, ilustradas na Figura 1.

Figura 1 - Fluxo de etapas para o tratamento dos dados



Dessa forma, foram criados rasters para os *assets*, cobrindo todas as datas. Um raster é uma matriz de pixels que representa informações geoespaciais (HIJMANS, 2022). O BDC utiliza a projeção Albers Equal-Area Conic, apropriada para grandes áreas. Contudo, o projeto foca na região de Três Pontas, cuja projeção está em WGS84, como ilustrado na primeira etapa da Figura 1. Para alinhar as coordenadas, o shapefile da área foi adaptado para a projeção do BDC utilizando a biblioteca Terra em R, na segunda etapa da Figura 1.

Além disso, os rasters abrangem uma área maior que Três Pontas, então foi necessário realizar corte para ajustar a região. Em seguida, foi aplicada uma máscara para focar apenas na área delimitada por Três Pontas, ilustrados nas etapas nas duas últimas etapas da Figura 1, respectivamente.

Após o tratamento dos dados, iniciou-se a definição das estratégias de armazenamento. Foram definidas duas abordagens para análise. A principal diferença entre as modelagens reside na organização dos dados.

Na Modelagem 1 da Figura 2, todas as informações de um pixel, incluindo as diferentes datas com seus valores de bandas espectrais e IVs, estão armazenadas em um único documento. Na Modelagem 2 da Figura 2, as informações de um pixel são distribuídas em múltiplos documentos, um para cada data, sendo armazenada na coleção Séries Temporais do MongoDB.

A primeira modelagem favorece análises centradas em um pixel, enquanto a segunda facilita a análise temporal, aproveitando a estrutura eficiente da coleção Time Series para séries temporais do MongoDB. As duas modelagens estão ilustradas na Figura 2.

Figura 2 - Estratégias de armazenamento

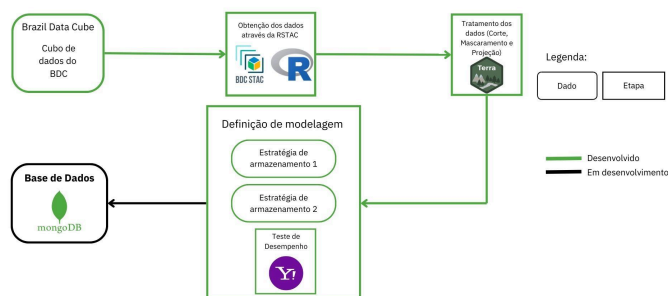


Com base nas modelagens de persistência dos dados, será realizada uma análise de desempenho para avaliar o custo de escrita (inserção de dados) e leitura (consulta). Essa análise é essencial para determinar qual modelagem oferece maior eficiência e desempenho nas futuras aplicações.

Para isso, está sendo utilizada a ferramenta de benchmark YCSB, conhecida por sua estrutura adequada para ambientes não relacionais, como o MongoDB, e por oferecer cargas de trabalho definidas para testar limites de desempenho (KIM, 2019). Optou-se pelo uso de workloads padrão da ferramenta, que já fornecem uma boa base para a análise.

Duas interfaces de dados foram desenvolvidas que abordam estratégias de armazenamento de dados geoespaciais, já que a interface padrão do YCSB não atende ao contexto específico deste trabalho. Atualmente, o projeto está na fase de análise. Está sendo utilizada a versão do YCSB desenvolvida em GO Lang, chamada GO-YCSB, que foi a mais adequada aos requisitos do estudo, apesar das limitações na documentação e da complexidade do uso da ferramenta. A Figura 3 ilustra os estágios que foram desenvolvidos e os que estão em desenvolvimento.

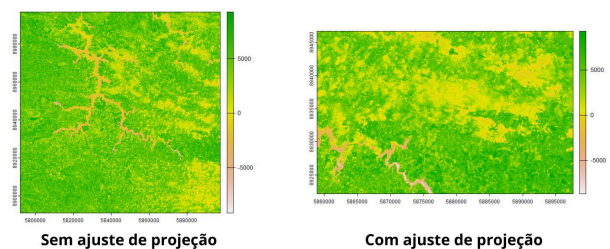
Figura 3 - Fluxo de desenvolvimento da pesquisa



Resultados e discussão

Durante o estágio inicial da análise, que está sendo descrito neste artigo, foi possível constatar alguns aspectos que influenciam a análise dos dados e que podem contribuir para futuras pesquisas. Inicialmente, esperava-se que os dados pudessem ser requisitados diretamente na extensão territorial do shapefile. Contudo, foi necessário recuperar os dados em uma extensão predefinida que incluía o shapefile, para depois realizar a transformação. Esse processo de corte para ajustar a projeção levou 6 minutos e 38 segundos para 18 datas do período analisado. Esse tempo tende a aumentar conforme o intervalo dos dados se expande, e não foi possível encontrar alternativas para otimizar essa operação com os recursos utilizados, como as bibliotecas Terra e STAC. Na Figura 4, é ilustrada um IV ndvi sem o ajuste e com o reajuste de projeção.

Figura 4 – Comparação entre as projeções do IV ndvi



Outro resultado relevante está relacionado ao uso do benchmark GO-YCSB. Durante o desenvolvimento das interfaces de teste, foi identificado que a modelagem que utilizava a coleção Time Series do MongoDB necessitava de um tratamento específico, como a definição do campo temporal (timeField) e a configuração correta no YCSB (KIM, 2019). Essa mudança foi crucial para garantir que a coleção fosse testada com os parâmetros adequados, maximizando sua eficiência.

Com a realização dos testes de desempenho, espera-se comparar as estratégias de armazenamento, analisando os resultados por meio das cargas de trabalho, o que possibilitará a criação de gráficos para facilitar a comparação entre as modelagens.

Conclusões

Neste artigo, foi descrito um estudo cujo objetivo principal foi coletar dados do projeto Brazil Data Cube (BDC). Após a captura dos dados, foi necessário realizar o tratamento, incluindo projeção, recorte e mascaramento. Duas estratégias de

armazenamento foram propostas, com potencial de sucesso. A expectativa é que, após os testes de desempenho, seja possível determinar qual modelagem oferece a melhor eficiência.

Estudos como este evidenciam a relevância do BDC, que é um projeto brasileiro iniciado em 2019. Esse trabalho oferece suporte a outras iniciativas que utilizam ferramentas do BDC e apresenta contribuições preliminares importantes para as áreas de sensoriamento remoto, processamento de dados geoespaciais e modelagem de bancos de dados não relacionais.

A análise comparativa das estratégias de armazenamento está em andamento, e embora ainda não se tenha concluído qual é mais eficiente, o estudo já gerou *insights* valiosos sobre o processamento e armazenamento dos dados em um banco não relacional.

Ao final do projeto, esses dados serão entregues aos cafeicultores de forma interpretada no contexto da cafeicultura. E, com isso, espera-se democratizar o acesso dessas informações tão relevantes aos pequenos agricultores, em especial. No entanto, toda cadeia produtiva se beneficiará, incluindo técnicos e pesquisadores.

Agradecimentos

Agradecimentos à orientadora Melise Maria Veiga de Paula pela valiosa oportunidade de participar deste projeto, que trouxe contribuições significativas em termos de conhecimento e desenvolvimento acadêmico. Um agradecimento especial é direcionado a Vanessa Cristina Oliveira de Souza, por todo o suporte e orientação ao longo do processo, cujo auxílio foi fundamental para a realização deste trabalho. Reconhece-se, também, o apoio indispensável do CNPq, que forneceu os recursos necessários para a condução deste estudo, e à Universidade Federal de Itajubá (UNIFEI), por todo o suporte institucional oferecido ao longo da pesquisa.

Referências

FAEB. Conheça os maiores produtores de café do mundo. Disponível em: <https://sistemafaeb.org.br/conheca-os-maiores-produtores-de-cafe-do-mundo/#:~:text=Brasil%20ocupa%20o%20topo%20d>.
BASSOI, Luís Henrique; INAMASU, Ricardo Yassushi; BERNARDI, Alberto Carlos de Campos; VAZ, Carlos Manoel Pedro; SPERANZA, Eduardo Antonio; CRUVINEL, Paulo Estevão. Agricultura de precisão e agricultura digital. In: TECCOGS – Revista Digital de Tecnologias Cognitivas, n. 20, jul./dez. 2019, p. 17-36.

MOREIRA, M. A.; BARROS, M. A.; RUDORFF, B. F. T.. Geotecnologias no mapeamento da cultura do café em escala municipal. *Sociedade & Natureza*, v. 20, n. 1, p. 101–110, jun. 2008.

GOMES, V. C. F. et al. Brazil Data Cube Workflow Engine: a tool for big Earth observation data processing. *International journal of digital earth*, v. 17, n. 1, 2024.

HACHIMI, Chouaib El et al. Smart weather data management based on artificial intelligence and big data analytics for precision agriculture. *Agriculture*, v. 13, n. 1, p. 95, 2022.

ANTONIO ROBERTO FORMAGGIO; IEDA DEL'ARCO SANCHES. Sensoriamento remoto em Agricultura. [s.l.] Oficina de Textos, 2017.

COLOSI, M.; MARTELLA, F.; PARRINO, G.; CELESTI, A.; FAZIO, M.; VILLARI, M. Time Series Data Management Optimized for Smart City Policy Decision. In: *Proceedings of the 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, Taormina, Italy, 2022. p. 585-594. doi: 10.1109/CCGrid54584.2022.00068.

SOUSA, G.; LEANDRO, M.; FERMINO, H. O papel da cafeicultura no município de Três Pontas (MG). [s.l.: s.n.]. 2023

KIM, S.; KANWAR, Y. S. GeoYCSB: A benchmark framework for the performance and scalability evaluation of NoSQL databases for geospatial workloads. 2019 IEEE International Conference on Big Data (Big Data). *Anais...IEEE*, 2019.

HIJMANS, Robert J. et al. Package 'terra'. Maintainer: Vienna, Austria, 2022.

MongoDB Documentation. Disponível em: <<https://www.mongodb.com/pt-br/docs/>>. Acesso em: 6 out. 2024.

CABRAL, J. V. L. et al. Enabling schema-independent data retrieval queries in MongoDB. *Information systems*, v. 114, n. 102165, p. 102165, 2023.

FERREIRA, Karine R. et al. Earth observation data cubes for Brazil: Requirements, methodology and products. *Remote Sensing*, v. 12, n. 24, p. 4033, 2020.