Avaliação Comparativa de Estratégias de Particionamento para Dados Raster em Bancos de Dados Multidimensionais

Marco Túlio Alves de Barros[1], **Geovani Pereira dos Santos**[1], **Daniel dos Santos Kaster**[1]

[1] Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – 86057-970 – Londrina – PR – Brasil

{marcotulio.barros, geovani.pereira, dskaster}@uel.br

Abstract. *The chunking method in raster data directly impacts the performance of retrieving data via geospatial tools. This paper compares the regular, horizontal, and vertical chunking strategies, using the systems* `GDAL`*,* `PostGIS`*, and* `RasDaMan`*. To make the comparatives, we analyzed execution time and the proportion of useful pixels retrieved. Tests were conducted using the MODIS, MapBiomas, and Sentinel-2 datasets, covering different resolutions and query scenarios based on real data. The ratio results (useful/total pixels) indicate that* `RasDaMan` *achieved an average of 0.85,* `PostGIS` *0.69, and* `GDAL` *0.17. As an additional contribution, the study also proposes a replicable methodology for handling data based on real-world scenarios.*

Resumo. *O método de particionamento de dados raster impacta diretamente a eficiência da recuperação em bancos de dados espaciais. Este artigo compara estratégias de particionamento regular, horizontal e vertical, usando os sistemas* `GDAL`*,* `PostGIS` *e* `RasDaMan`*. Para as comparações, foram analisados tempo de execução e proporção de pixels úteis recuperados. Foram realizados testes com conjuntos, abrangendo diferentes resoluções e cenários de consulta baseados em dados reais. Os resultados de aproveitamento (pixels úteis/totais) indicam que o* `RasDaMan` *com média de 0,85,* `PostGIS` *com 0,69 e* `GDAL` *com 0,17. Como contribuição adicional, o estudo também propõe uma metodologia replicável para manipulação de dados baseados em cenários reais.*

1. Introdução

O crescimento da coleta e uso de dados geoespaciais, impulsionado por avanços no sensoriamento remoto, tem gerado desafios significativos para o armazenamento e a recuperação eficiente dessas informações. Dados ***raster***, amplamente utilizados em aplicações como monitoramento ambiental, planejamento urbano e análise agrícola [Murodilov et al. 2023, TAQUES and ROCHA 2014, Esri 2012], apresentam estruturas volumosas que exigem estratégias otimizadas para manipulação em bancos de dados [Sveen 2019, Furtado and Baumann 1999, Widmann and Baumann 1999, Hu et al. 2018].

O particionamento de dados *raster* desempenha um papel essencial na eficiência das consultas, permitindo armazenar e recuperar blocos específicos de maneira otimizada (custo computacional). Estratégias como particionamento regular, horizontal, vertical e baseado em regiões de interesse (ROI, do inglês ***R****egion* ***O****f*

***I**nterest*) afetam diretamente o tempo de resposta e a quantidade de dados processados [Furtado and Baumann 1999, Widmann and Baumann 1999]. No entanto, a literatura ainda carece de estudos que avaliem comparativamente essas estratégias em cenários reais e em diferentes sistemas de gerenciamento de dados multidimensionais [Baumann et al. 2021, Vu et al. 2021]. Embora existam pesquisas sobre armazenamento e recuperação de dados multidimensionais [Baumann et al. 2021, Furtado and Baumann 1999], muitas não detalham o impacto quantitativo do particionamento nem fornecem experimentos replicáveis. Este trabalho busca preencher essa lacuna, oferecendo uma análise rigorosa e uma metodologia replicável para otimizar a manipulação de dados *raster* em bancos multidimensionais.

Este artigo apresenta uma análise quantitativa das estratégias de particionamento em três sistemas amplamente utilizados: `GDAL`, `PostGIS` e `RasDaMan` [GDAL/OGR contributors 2024, PostGIS Development Team 2025, Baumann et al. 1998]. Assim, o objetivo principal é avaliar o impacto do tipo e tamanho dos blocos de particionamento em diferentes cenários, identificando características aplicadas em um cenário baseado em cenários reais. Para tal, foi proposta uma metodologia para manipulação de conjuntos de dados e cargas de trabalho sintéticos baseados em dados reais.

Os experimentos demonstraram que o `RasDaMan` apresentou o melhor desempenho quanto ao aproveitamento de pixels, com média de 0,85, seguido por `PostGIS` com 0,69 e `GDAL` com 0,17. Observou-se que a eficiência da recuperação depende da resolução do *raster* e da distribuição das ROIs, com *rasters* de maior precisão resultando em menor desperdício de dados. Além disso, os resultados indicam um equilíbrio entre as diferentes estratégias de particionamento, sem uma orientação claramente superior, variando conforme o cenário e a estrutura dos dados analisados.

2. Fundamentação

Esta seção apresenta os conceitos fundamentais relacionados ao armazenamento e manipulação de dados *raster*, abordando os principais desafios do particionamento e os impactos das diferentes estratégias adotadas. Além disso, descreve as ferramentas utilizadas no estudo, suas características e formas de operação.

2.1. Principais Conceitos

Os dados *raster*, na Figura 1, são representações matriciais de informações espaciais, onde cada pixel contém um valor de intensidade associado a uma característica geográfica específica (e.g. temperatura, altitude, cobertura do solo), e também estão relacionados a uma precisão real em metros. Eles são amplamente utilizados em aplicações de sensoriamento remoto e geoprocessamento devido à sua capacidade de representar variações contínuas no espaço.

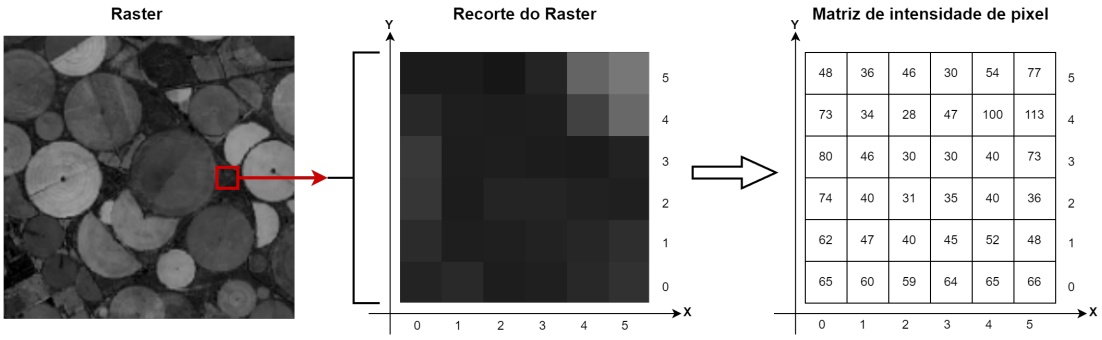

Figura 1. Representação abstrata da estruturação de um dado *Raster*.

Em contrapartida, os dados vetoriais são representados por pontos, linhas ou polígonos e descrevem entidades discretas (e.g. limites políticos, redes viárias, corpos d'água). No contexto deste estudo, os *shapefiles* vetoriais são usados para definir o ROI sobre os *rasters*, servindo como base para as consultas espaciais.

Em diversas aplicações geoespaciais, é comum lidar com séries temporais de dados *raster*. Esse conceito, conhecido como *time slices*, refere-se à organização de múltiplas imagens adquiridas ao longo do tempo, formando um *datacube* multidimensional. Cada camada do cubo (*slice*) representa um instante temporal específico, permitindo análises de evolução e mudanças espaciais ao longo do tempo. Na Figura 2, está representado uma janela temporal de n intervalos (*slices* $T_1 - T_n$) e o ROI definido pelo retângulo em amarelo em cada *slice*, o que permite a inferência de análises temporais sobre essa região.

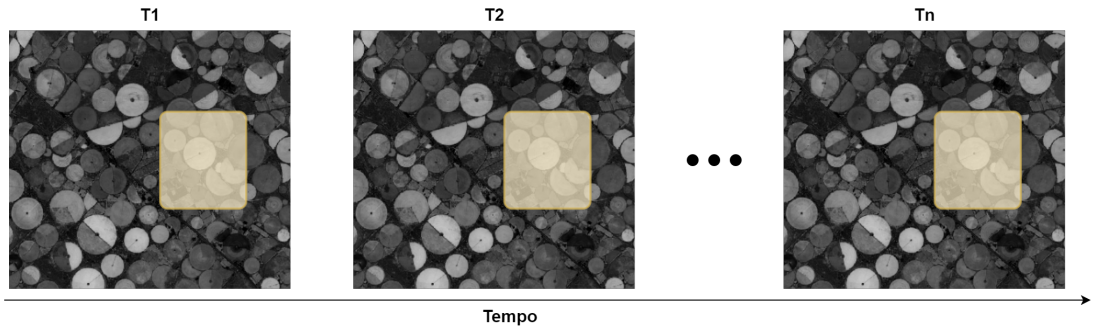

Figura 2. Representação de uma janela temporal.

Outro conceito fundamental nesse artigo, o particionamento de dados *raster*, também conhecido como *chunking*, consiste na divisão dos dados em blocos menores para otimizar armazenamento e recuperação. Diferentes estratégias de particionamento afetam o desempenho das consultas, dependendo do padrão de acesso aos dados. As principais abordagens incluem:

- **Particionamento Regular**: divisão uniforme em blocos de tamanhos fixos, resultando principalmente blocos quadrados;

- **Particionamento Horizontal**: divisão com orientação no bloco resultante, produzindo retângulos horizontais;
- **Particionamento Vertical**: divisão com orientação no bloco resultante, produzindo retângulos verticais;
- **Particionamento por Região de Interesse**: busca adaptar-se a padrões de acesso, minimizando desperdícios, ou seja, associa possíveis regiões de interesse e consultas a blocos específicos no particionamento.

Com as principais possibilidades apresentadas, percebe-se como a escolha da estratégia influencia diretamente a eficiência das consultas e o volume de dados recuperados, sendo um fator crítico para a manipulação eficaz de grandes volumes de dados *raster*.

Para complementar, a Figura 3 ilustra o impacto do particionamento em consultas por ROI. Nela busca evidenciar como consultas pelo conjunto hachurado acarreta recuperação de muitos outros dados desnecessários, aumentando tempo e recursos computacionais para tal. Em especial na Figura 3 (B), observa-se que talvez seja um *chunking* por áreas de interesse, mas sem muita antecipação dos intervalos. Já na Figura 3 (C), é, provavelmente, um *chunking* por área de interesse com preparações prévias, apresentando antecipação das possíveis consultas.

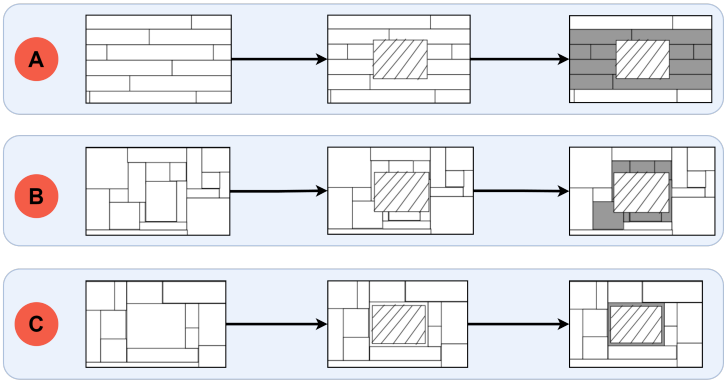

Figura 3. Impacto do particionamento em consultas.

2.2. Ferramentas para Armazenamento e Processamento de Dados *Raster*

Para a realização deste estudo, foram analisadas três ferramentas amplamente utilizadas e consolidadas na literatura para manipulação de dados *raster* em bancos de dados multidimensionais: `GDAL`, `PostGIS` e `RasDaMan`. Cada uma dessas soluções possui características distintas no que se refere ao armazenamento, recuperação e particionamento de dados.

- `GDAL`: é uma biblioteca de código aberto para manipulação de dados geoespaciais, suportando múltiplos formatos *raster*. Embora seja amplamente utilizada, sua abordagem de particionamento é limitada, pois os blocos são definidos pelos metadados do arquivo e variam conforme características do arquivo original, ocasionando, geralmente, em alto desperdício. Sua principal vantagem é a simplicidade, envolvendo programação direta, e compatibilidade com diversos sistemas.
- `PostGIS`: é uma extensão espacial do PostgreSQL que especializada em dados geoespaciais. Oferece suporte a particionamento automático (macro que define

dimensões ideais para os blocos resultantes) ou manual, sendo amplamente utilizado por sua robustez, integração e popularidade. No entanto, seu desempenho está associado a bancos relacionais, estruturas de indexação e características dos dados.

- `RasDaMan`: é um sistema de banco de dados especializado no armazenamento de *arrays* multidimensionais, oferecendo suporte avançado para consultas em grandes volumes de dados *raster*. Sua principal característica é a flexibilidade no particionamento, exigindo ao usuário definir configurações personalizadas para otimizar o desempenho das consultas. Entretanto, sua curva de aprendizado é mais acentuada e a configuração manual dos parâmetros pode ser um desafio.

3. Configuração Experimental e Procedimentos

Nesta seção, são descritos os conjuntos de dados utilizados, o processo de construção das cargas de trabalho e os procedimentos adotados para a realização dos testes. A metodologia foi concebida para garantir a reprodutibilidade dos experimentos e fornecer uma base sólida para a análise quantitativa das estratégias de particionamento.

3.1. Conjuntos de Dados e Carga de Trabalho

A seleção e preparação dos dados são etapas fundamentais para garantir a coerência e relevância dos experimentos propostos. Para assegurar análises representativas, foi realizado um estudo detalhado sobre *rasters* amplamente utilizados e suas características, visando contemplar diferentes resoluções e contextos de aplicação. Da mesma forma, a escolha dos dados vetoriais considerou elementos espaciais recorrentes em estudos geográficos, como delimitações estaduais, municipais e redes viárias, fazendo com que as cargas de trabalho geradas simulem análises em cenários reais.

Os experimentos foram conduzidos utilizando três conjuntos de dados *raster* amplamente utilizados em aplicações geoespaciais: MODIS, MapBiomas e SENTINEL-2. Cada um desses conjuntos de dados possui características distintas de resolução e precisão espacial, exibidas na Tabela 1, permitindo uma análise abrangente sobre os impactos do particionamento [NASA 2024, Souza et al. 2020, Agency 2024].

Dataset	**Dimensão**	**Precisão**
MODIS	$(5566 \times 8020)\ km \leftrightarrow (20mil \times 30mil)\ px$	250 m
MapBiomas	$(4657 \times 4754)\ km \leftrightarrow (155mil \times 155mil)\ px$	30 m
SENTINEL-2	$(110 \times 110)\ km \leftrightarrow (11mil \times 11mil)\ px$	10 m

Tabela 1. Resumo das principais características do *datasets*.

A montagem das cargas de trabalho foi estruturada para garantir relevância prática e adesão aos objetivos do estudo. Para isso, estabelecemos um critério que relaciona os conjuntos a contextos de análise comuns. Assim, os dados vetoriais, utilizados para definir as ROIs, foram obtidos do Instituto Brasileiro de Geografia e Estatística (IBGE) e contemplam *shapefiles* de estradas federais, mesorregiões e municípios brasileiros. Esses dados foram processados para garantir compatibilidade com os *rasters*, passando por etapas de reprojeção de coordenadas, aplicação de grids, recortes, aplicação de buffers, filtros automáticos e manuais, análises de consistência e testes isolados.

Para que as análises e testes fossem coerentes e estivessem conforme a proposta, devido ao escopo e características da combinação de conjuntos de dados com cargas de trabalho, consultas de mesorregiões e rodovias federais foram feitas no MODIS e MapBiomas, enquanto, de forma equivalente, municípios e rodovias locais feitas no SENTINEL-2. Para exibir o produto, a Figura 4 (A) apresenta mesorregiões com *raster* MODIS, a Figura 4 (B) as rodovias federais com o *raster* MapBiomas e a Figura 4 (C) os municípios e rodovias locais com o SENTINEL-2.

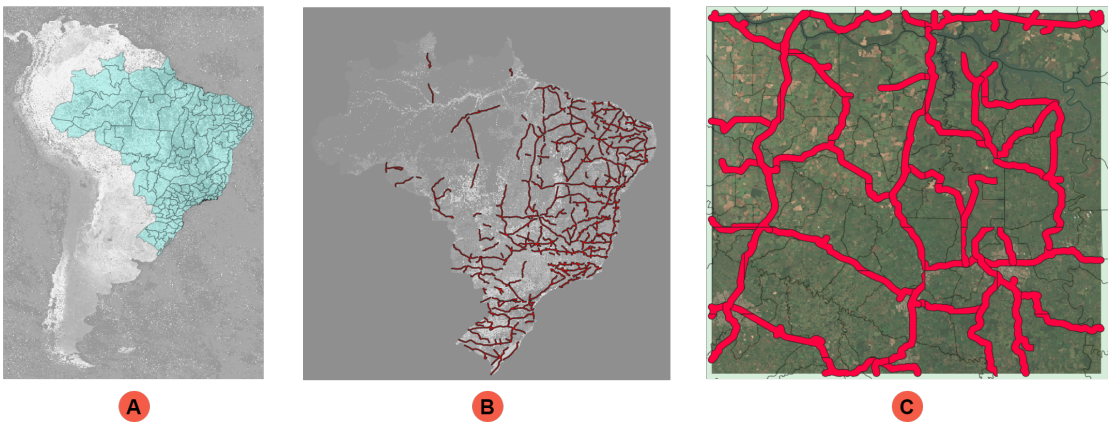

Figura 4. Amostra de conjunto de dados com cargas de trabalho utilizados na comparação quantitativa. (A) Amostra extraída do MODIS com mesorregiões. (B) Amostra extraída do MapBiomas com rodovias federais. (C) Amostra extraída *Sentine-2* com municípios e rodovias locais.

3.2. Particionamentos e Consulta

A definição dos tamanhos e técnicas de partição foram baseadas no equilíbrio entre granularidade e eficiência computacional, garantindo que os testes fossem reproduzíveis e alinhados às estratégias adotadas na literatura [Baumann et al. 2021, Furtado and Baumann 1999, Baumann et al. 1998]. O processo envolveu testes isolados preliminares, revisão de abordagens em sistemas consolidados e análises das métricas coletadas, assegurando que os valores estipulados refletissem cenários realistas e otimizados.

No `RasDaMan`, os blocos foram configurados para áreas entre 4096 e 90000 pixels, alternando as dimensões para analisar os particionamentos regular, horizontal e vertical. Para `PostGIS`, os blocos foram definidos em utilizando a abordagem principal de particionamento automático, portanto as áreas são ajustadas segundo o *raster*. Já no `GDAL`, os blocos seguiram a configuração automática padrão, com área determinada pelos metadados dos arquivos.

Os testes foram planejados para avaliar a eficiência das estratégias de particionamento em diferentes cenários, executando apenas o acesso e recuperação de dados. Para isso, foram sorteadas 40 ROIs em cada cenário e uma definição da janela temporal com 10 intervalos, resultando em cerca de 4400 execuções distribuídas entre os diferentes particionamentos e dimensões.

As métricas analisadas incluem tempo de execução total para acesso e recuperação dos dados; blocos recuperados, calculados através dos conhecimentos da ROI e dimensão

do particionamento em questão; e pixels quantificando a relação entre os dados totais acessados e os pixels úteis efetivamente recuperados.

4. Resultados

Com todos os elementos do artigo já apresentados, esta seção apresenta os principais resultados obtidos nos experimentos, destacando o impacto das estratégias de particionamento na recuperação de dados *raster*. Para manter a objetividade, são exibidas somente as médias gerais das métricas analisadas, acompanhadas de uma discussão sobre os comportamentos observados.

Os gráficos a seguir, exibidos na Figura 5 (A-D), mostram os valores médios das métricas analisadas para cada sistema. Os dados representam o desempenho agregado considerando os diferentes conjuntos de dados e cargas de trabalho utilizados no estudo.

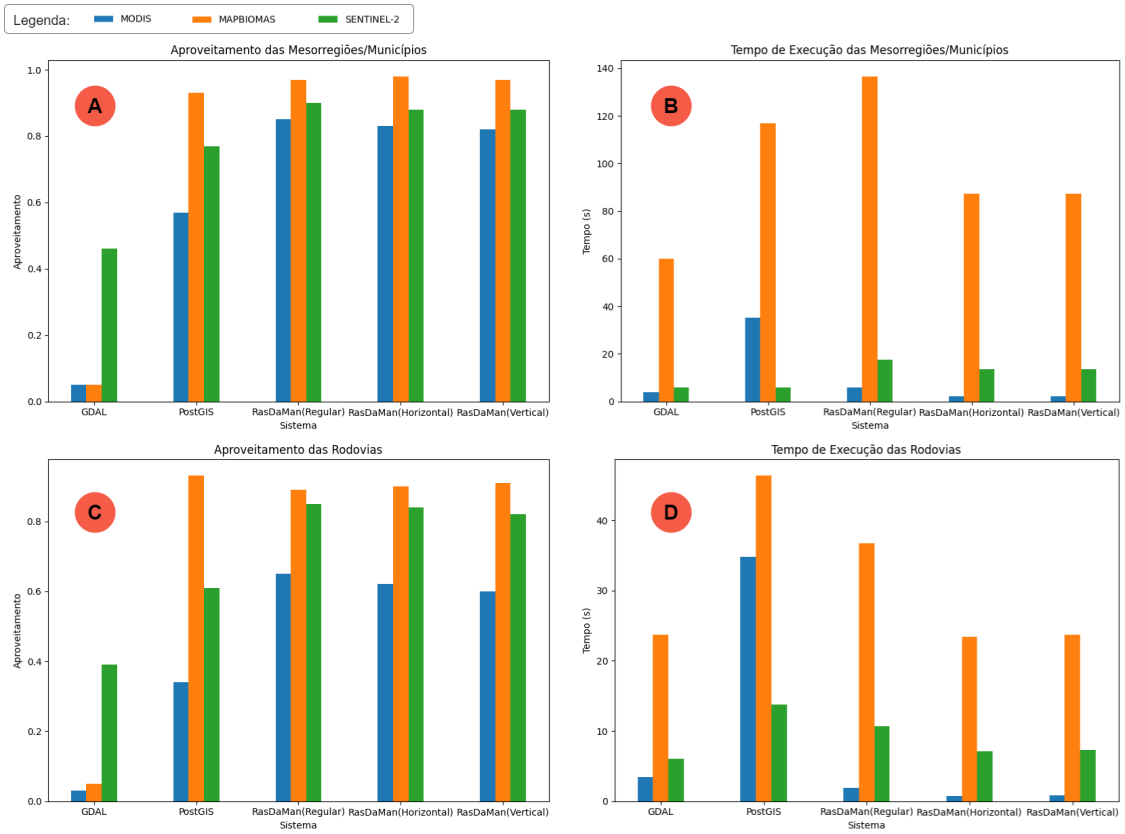

Figura 5. Resultados obtidos. (A) Média Geral de Aproveitamento para Mesorregiões e Municípios. (B) Média Geral de Tempo para Mesorregiões e Municípios. (C) Média Geral de Aproveitamento para Rodovias. (D) Média Geral de Tempo para Rodovias.

Os resultados mostram que a eficiência da recuperação de dados é altamente impactada pela dimensão e precisão do *raster*, da distribuição das ROIs e da orientação do particionamento. *Rasters* de baixa resolução (MODIS) tiveram recuperação rápida, mas com desperdício associado a proporção bloco do particionamento por ROI, enquanto *rasters* de alta precisão (MapBiomas e SENTINEL-2) apresentaram melhor aproveitamento, embora com maior custo computacional.

O RasDaMan permitiu avaliar diferentes opções de particionamento. Nesses testes, o aumento no tamanho do bloco (em número de pixels), resultou em uma piora no aproveitamento, chegando em torno de 15 pontos percentuais. Já o tempo de execução manteve-se relativamente estável com o aumento do tamanho do bloco.

Houve uma pequena superioridade dos particionamentos horizontal e vertical sobre o regular, em termos de aproveitamento. Contudo, o particionamento regular teve um tempo de execução médio consideravelmente maior aos demais. Esse comportamento é devido método de computação implementado para o particionamento regular, que é diferente e menos eficiente que os dos demais particionamentos. O aproveitamento apenas sensivelmente melhor dos particionamentos horizontal e vertical frente ao regular ocorreu mesmo com relação às rodovias. Isto porque o *workload* alterna entre rodovias na vertical e na horizontal e a execução beneficia-se do particionamento quando a consulta coincide com o particionamento (i.e., trecho da rodovia majoritariamente na horizontal é beneficiado do particionamento horizontal), mas é penalizada quando o oposto acontece.

Por ser um estudo baseado em dados reais, não houve dominância de um tipo de consulta específico, portanto, não houve uma orientação superior de particionamento, considerando-se todas as configurações e cargas de trabalho. No entanto, em testes mais específicos e controlados, notou-se uma melhora em alguns casos. No geral, o desempenho se equilibrou, mostrando que a alternância entre estratégias é essencial para otimizar consultas em cenários reais. A flexibilidade na escolha do particionamento, ajustando-se às características específicas de cada conjunto de dados com carga de trabalho e consultas possíveis, foi crucial para alcançar uma eficiência balanceada e robusta.

A comparação entre os sistemas `GDAL`, `PostGIS` e `RasDaMan` revelou diferenças significativas em termos de desempenho e eficiência, que devem ser levadas em consideração em escolhas futuras. O `GDAL` foi muito rápido, mas apresentou um alto desperdício devido aos blocos automáticos de área elevada, resultando em um consumo muito maior de memória. Já `PostGIS`, mostrou-se mediano em termos de aproveitamento e o pior em termos de tempo de execução, em geral, tendo sido competitivo apenas para a carga de trabalho de municípios no SENTINEL-2. Entretanto, é a opção que mais se destaca em termos de funções integradas e possibilidade de índices de consultas. Por fim, o RasDaMan apresentou-se como o mais flexível e com mais variações no contexto desse artigo, demonstrou as vantagens de ser um banco especializado nesse tipo de dado.

5. Conclusão

Este estudo apresentou uma análise quantitativa sobre o impacto das estratégias de particionamento na recuperação de dados *raster* em bancos de dados multidimensionais, aplicado em um cenário sintético baseado em dados reais. Foram comparados `GDAL`, `PostGIS` e `RasDaMan`, avaliando métricas de tempo de execução, blocos recuperados e proporção de pixels úteis.

Os resultados comparando a proporção entre pixels úteis/totais indicam que `RasDaMan` obteve o melhor aproveitamento (0,85 de aproveitamento médio), seguido por `PostGIS` (0,69) e `GDAL` (0,17). A alternância entre particionamento horizontal e vertical mostrou que não há uma estratégia superior absoluta, devido à alternância entre diferentes orientações na ROI baseadas em dados reais, bem como identificação e explicações do impacto de características específicas dos dados nas consultas.

Agradecimentos

Agradecemos à Fundação Araucária e ao NAPI CIA-AGRO pelo apoio e fomento dessa pesquisa.

Referências

Agency, E. S. (2024). Sentinel-2. Accessed: 2024-09-03.

Baumann, P., Dehmel, A., Furtado, P., Ritsch, R., and Widmann, N. (1998). The multidimensional database system rasdaman. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 575–577.

Baumann, P., Misev, D., Merticariu, V., and Huu, B. P. (2021). Array databases: Concepts, standards, implementations. *Journal of Big Data*, 8:1–61.

Esri (19 de fevereiro de 2012). Mapa topográfico mundial. ArcGIS Online. Escala Não Fornecida.

Furtado, P. and Baumann, P. (1999). Storage of multidimensional arrays based on arbitrary tiling. In *Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337)*, pages 480–489. IEEE.

GDAL/OGR contributors (2024). *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation.

Hu, F., Xu, M., Yang, J., Liang, Y., Cui, K., Little, M. M., Lynnes, C. S., Duffy, D. Q., and Yang, C. (2018). Evaluating the open source data containers for handling big geospatial raster data. *ISPRS International Journal of Geo-Information*, 7(4).

Murodilov, K. T., Muminov, I., and Abdumalikov, R. (2023). Using geospatial data to optimize agricultural production in region/country. *Educational Research in Universal Sciences*, 2(4):115–117.

NASA (2024). Modis (moderate resolution imaging spectroradiometer). Accessed: 2024-09-03.

PostGIS Development Team (2025). *PostGIS Raster Reference*. Accessed: 2025-01-11.

Souza, C. M., Z. Shimbo, J., Rosa, M. R., Parente, L. L., A. Alencar, A., Rudorff, B. F. T., Hasenack, H., Matsumoto, M., G. Ferreira, L., Souza-Filho, P. W. M., de Oliveira, S. W., Rocha, W. F., Fonseca, A. V., Marques, C. B., Diniz, C. G., Costa, D., Monteiro, D., Rosa, E. R., Vélez-Martin, E., Weber, E. J., Lenti, F. E. B., Paternost, F. F., Pareyn, F. G. C., Siqueira, J. V., Viera, J. L., Neto, L. C. F., Saraiva, M. M., Sales, M. H., Salgado, M. P. G., Vasconcelos, R., Galano, S., Mesquita, V. V., and Azevedo, T. (2020). Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(17).

Sveen, A. F. (2019). Efficient storage of heterogeneous geospatial data in spatial databases. *Journal of Big Data*, 6(1):102.

TAQUES, R. and ROCHA, M. (2014). Aptidão agrícola para a cultura da mamoneira (ricinus communis l.) no estado do espírito santo. In: CONGRESSO BRASILEIRO DE AGRONOMIA, 25., Vitória, ES.[Anais...] Vitória ....

Vu, T., Eldawy, A., Hristidis, V., and Tsotras, V. (2021). Incremental partitioning for efficient spatial data analytics. *Proceedings of the VLDB Endowment*, 15(3):713–726.

Widmann, N. and Baumann, P. (1999). Performance evaluation of multidimensional array storage techniques in databases. In *Proceedings. IDEAS'99. International Database Engineering and Applications Symposium (Cat. No. PR00265)*, pages 385–389. IEEE.