UNIVERSIDADE FEDERAL DE JATAÍ (UFJ)
INSTITUTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS (ICET)
CURSO DE CIÊNCIA DA COMPUTAÇÃO

Leonardo Paiva Vieira

Caracterização da Precisão da Estimativa de Seletividade em Consultas Distribuídas de Multijunção Espacial

Jataí-Goiás

2024

Leonardo Paiva Vieira

Caracterização da Precisão da Estimativa de Seletividade em Consultas Distribuídas de Multijunção Espacial

Trabalho de Conclusão de Curso apresentado ao curso de Ciência da Computação do Instituto de Ciências Exatas e Tecnológicas da Universidade Federal de Jataí (UFJ), como requisito para obtenção do título de Bacharel em Ciência da Computação.

Orientador(a): Prof. Dr. Thiago Borges de Oliveira

Jataí-Goiás

2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFJ.

Vieira, Leonardo Paiva
Caracterização da Precisão da Estimativa de Seletividade em Consultas Distribuídas de Multijunção Espacial / Leonardo Paiva Vieira. - 2024.
L, 50 f.: il.

Orientador: Prof. Dr. Thiago Borges de Oliveira.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Jataí, Instituto de Ciências Exatas e Tecnológicas, Ciência da Computação, Jataí, 2024.
Bibliografia. Apêndice.
Inclui siglas, abreviaturas, símbolos, gráfico, tabelas, lista de figuras, lista de tabelas.

1. Multijunção Espacial. 2. Estimativa de Seletividade. 3. Propagação de Erro. 4. Processamento Distribuído. I. Oliveira, Thiago Borges de, orient. II. Título.

CDU 004

UFJ

UNIVERSIDADE FEDERAL DE JATAÍ
INSTITUTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

DECLARAÇÃO DE APROVAÇÃO DA VERSÃO FINAL

Declaro que o(a) discente Leonardo Paiva Vieira do curso de Bacharelado em Ciência da Computação foi aprovado(a) na defesa do Trabalho de Conclusão de Curso (TCC) com o título final Caracterização da Precisão da Estimativa de Seletividade em Consultas Distribuídas de Multijunção Espacial na data de 08/03/2024 e efetuou todas as correções pertinentes sugeridas pela banca examinadora, composta pelo seguintes membros:

Orientador(a)	Thiago Borges de Oliveira
Membro 1	Ariadne de Andrade Costa
Membro 2	Franciny Medeiros Barreto

Declaro ainda que a versão final anexada a este processo está adequada para ser devidamente depositada em repositório institucional.

Thiago Borges de Oliveira

Professor Orientador

Observação

Esta declaração deve ser assinada pelo(a) orientador(a)

Documento assinado eletronicamente por **THIAGO BORGES DE OLIVEIRA**, **Professor do Magistério Superior**, em 08/04/2024, às 10:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufj.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0258435** e o código CRC **5ADFE57B**.

Dedico este trabalho ao meu irmão Venâncio, pelos momentos que compartilhamos juntos desde o seu nascimento. Aos meus amigos de quarto, Elias e Gustavo, pelas risadas e reflexões, fundamentais para me manter motivado e forte em cada etapa deste percurso acadêmico.

Agradecimentos

Agradeço aos meus pais por terem dado todo o apoio necessário, aos meus amigos que me acompanharam durante essa jornada desafiadora, aos meus professores pelo o conhecimento compartilhado e, em especial, ao meu orientador pela a ajuda e paciência na realização deste trabalho.

"As pessoas que são loucas o suficiente para pensar que podem mudar o mundo são as que realmente o fazem.
(Steve Jobs)

Resumo

Os dispositivos capazes de coletar dados espaciais estão cada vez mais presentes em nossa vida cotidiana. No entanto, os sistemas encarregados de manipular esses dados não acompanharam o mesmo ritmo de evolução. Estes dados e sistemas têm um papel crucial no processo de tomada de decisões em diversas áreas da atividade humana. Na medicina, por exemplo, eles podem ser empregados para verificar se um determinado câncer regrediu quanto ao uso de terapias oncológicas. A multijunção é um tipo importante e complexo de consulta espacial usada para processar tais dados, e envolve uma série de etapas cuja execução eficiente é fundamental para reduzir o uso de recursos computacionais. Assim como nos sistemas de bancos de dados relacionais comuns, os espaciais possuem uma parte chamada otimizador de consultas, que leva em consideração a seletividade estimada para escolher o melhor plano de execução. Neste trabalho foi avaliada a precisão da estimativa de seletividade ao longo das diversas etapas envolvidas na multijunção espacial. Para isso, foi realizado um experimento envolvendo cinco consultas, cada uma composta por dez conjuntos de dados reais. Também foi calculada a estimativa de seletividade para cada uma das consultas utilizando dois métodos propostos na literatura. Os resultados foram aplicados a uma métrica, seguida da visualização por meio de gráficos para facilitar a análise. Constatou-se que o erro de seletividade é propagado com crescimento exponencial ao longo das etapas quando o método de estimativa possui baixa assertividade. O mesmo não ocorre quando o método é mais assertivo e a propagação tende a se manter mais controlada. Como o erro de estimativa provoca a escolha de formas menos eficientes de execução das consultas, espera-se que os resultados da pesquisa esclareçam o comportamento da estimativa de seletividade, fornecendo informações valiosas para estudos futuros nessa área.

Palavras-chave: *Multijunção Espacial; Estimativa de Seletividade; Propagação de Erro; Processamento Distribuído.*

Abstract

Devices capable of collecting spatial data are increasingly present in our daily lives. However, the systems responsible for processing this data have not kept up with the same pace of evolution. These data and systems play a crucial role in the decision-making process across various areas of human activity. In medicine, for example, they can be used to check whether a certain cancer has regressed in relation to the use of oncological therapies. The multiway join is an important and complex type of spatial query used to process such data and it involves a series of steps whose efficient execution is fundamental to reduce the use of computing resources. Similar to common relational database systems, the spatial have a component called query optimizer, which takes into account the estimated selectivity to choose the best execution plan. This work evaluated the accuracy of selectivity estimation throughout the various steps involved in multiway spatial join. For this purpose, an experiment involving five queries, each composed of ten sets of real data, was conducted. The selectivity estimation was also calculated for each query using two methods proposed in the literature. The results were applied to a metric, followed by visualization through graphs to facilitate analysis. It was concluded that the selectivity error is propagated exponentially along the query steps when the estimation method provides poor estimates. The same doesn't occurs when the estimation method is more accurate and the propagation tends to be more controlled. As the propagation of errors along query steps causes bad execution plan selection, we expect that this research results clarify the behavior of selectivity estimation, providing valuable insights for future studies in this area.

Keywords: *Multiway Spatial Join; Selectivity Estimation; Error Propagation; Distributed Processing.*

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

GPS	*Global Positioning System*
SIG	Sistema de Informação Geográfica
MBR	*Minimum Bounding Rectangle*
SBDR	Sistema de Banco de Dados Relacional
SBDE	Sistema de Banco de Dados Espaciais
SGBDR	Sistema Gerenciador de Banco de Dados Relacional
SGBDE	Sistema Gerenciador de Banco de Dados Espaciais
SGBDD	Sistema Gerenciador de Banco de Dados Distribuído
SQL	*Structured Query Language*
IHWAF	*Intermediate Histogram With Average Length Fix*
TCP	*Transmission Control Protocol*
GDAL	*Geospatial Data Abstraction Library*
GEOS	*Geometry Engine*
LGPL	*GNU Lesser General Public License*
MIT	*Massachusetts Institute of Technology*
INDE	Infraestrutura Nacional de Dados Espaciais
LAPIG	Laboratório de Processamento de Imagens e Geoprocessamento
UFJ	Universidade Federal de Jataí

Lista de símbolos

θ	Predicado espacial
$\in$	Pertence
$\wedge$	Operador lógico “e”
$\bowtie$	Junção espacial
$\forall$	Para todo
λ	Letra grega minúscula Lambda, usada neste trabalho para representar o erro relativo, em percentual, da estimativa para cada etapa da multijunção espacial.

Sumário

1 Introdução

1.1 Motivação

A proliferação de dispositivos equipados com o Sistema de Posicionamento Global (*Global Positioning System* - GPS) foi o catalisador para o aumento crescente na quantidade de dados espaciais que são coletados diariamente. Estes conjuntos de dados (*datasets*) são armazenados e manipulados através de um Sistema de Informação Geográfica (SIG) com o objetivo de facilitar o processo de tomada de decisões (BOUROS; MAMOULIS, 2019).

Em banco de dados relacionais, ao realizar uma consulta de junção, as tabelas normalmente são unidas através de um único atributo chave utilizando o operador lógico de igualdade (MAMOULIS; PAPADIAS, 2001a). Já em uma junção espacial (*spatial join*), é aplicado um predicado θ, usualmente a intersecção, a fim de identificar objetos correlacionados entre dois *datasets* (BRINKHOFF; KRIEGEL; SEEGER, 1996). Caso for envolvido mais de dois *datasets* na mesma consulta, ela é denominada de multijunção espacial (*multiway spatial join*) (MAMOULIS; PAPADIAS, 2001b).

A multijunção espacial possui aplicação em vários domínios da ação humana. Pode ser aplicada na geografia (ex., encontrar espécies de animais que vivem em áreas de preservação ambiental que foram destruídas por queimadas) (OLIVEIRA; COSTA; RODRIGUES, 2015), na construção de placas de circuito impresso (ex., encontrar circuitos lógicos que formam uma configuração topológica específica) (MAMOULIS; PAPADIAS, 2001a) e no diagnóstico assistido por imagem (ex., analisar imagens tomográficas de pacientes para confirmar se um câncer regrediu quando do uso de medicamentos ou procedimentos de quimioterapia, radioterapia, etc.) (AJI; WANG; SALTZ, 2012).

O processamento de uma consulta de multijunção espacial envolve complexos algoritmos de geometria computacional, o que implica em um longo tempo de resposta. Logo, uma alternativa para diminuir este tempo é executar a consulta em um ambiente distribuído, já que o custo total é dividido entre várias máquinas que formam o *cluster* (OLIVEIRA et al., 2023).

A execução de uma consulta de multijunção espacial pode ser realizada de diversas formas, denominadas de planos de execução. Cada um desses planos, embora sejam semanticamente equivalentes, se diferenciam em termos de utilização de recursos computacionais. É importante, portanto, escolher um plano de execução adequado para cada consulta com intuito de obter as respostas rapidamente e economizar recursos computacionais. Essa escolha é feita atualmente através de estimativa de custo computacional, frequentemente usando histogramas espaciais (OLIVEIRA, 2017).

Com o objetivo de simplificar os *datasets*, a literatura propõe a criação dos his-

togramas espaciais, sendo eles estruturas de dados que dividem a totalidade do espaço geográfico do *dataset* em uma grade formada por várias células, com as dimensões fixas ou variáveis, dependendo do tipo de construção escolhido. A Figura 1 ilustra um histograma de grade para um *dataset* com alertas de desmatamento do cerrado brasileiro. Cada célula armazena informações a respeito dos objetos que intersectam a sua estrutura (OLIVEIRA; COSTA; RODRIGUES, 2015). Uma importante informação é a cardinalidade, a qual é estabelecida a partir da quantidade de objetos que possuem o seu retângulo delimitador mínimo (*Minimum Bounding Rectangle* - MBR) dentro dos limites de uma célula. Sendo o MBR uma simplificação do objeto espacial (MAMOULIS; PAPADIAS, 2001a).

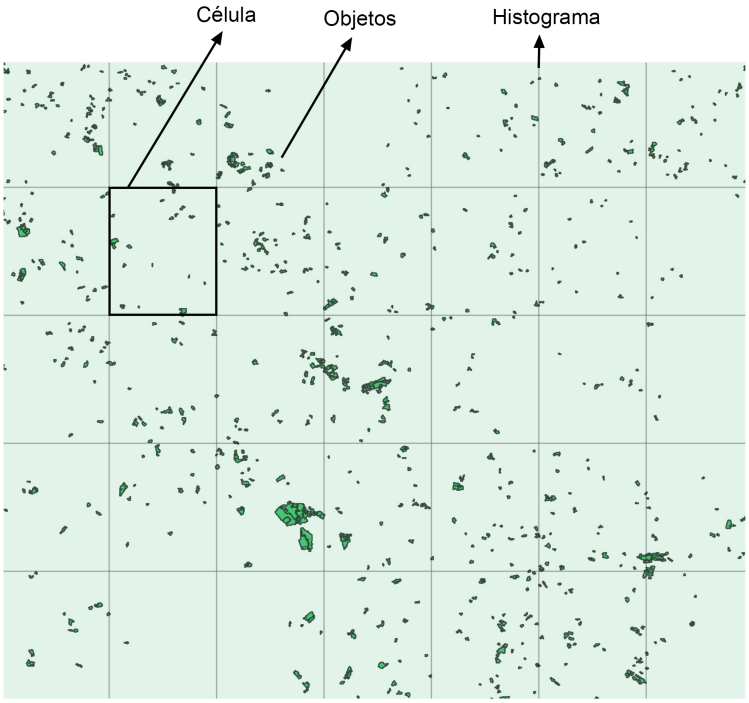


Figura 1 – Exemplo do Histograma de Grade para o *dataset* Alertas.

A estimativa de seletividade se destaca como uma métrica essencial para avaliar o custo computacional de uma junção espacial, já que refere-se a quantidade aproximada de objetos (cardinalidade) que serão retornados pela consulta. Desta forma, ao executar uma multijunção espacial distribuída, o otimizador leva em consideração a estimativa de seletividade para determinar qual nó do *cluster* será responsável por executar cada fragmento da consulta, a fim de diminuir o tempo total de execução e evitar nós ociosos durante a consulta (OLIVEIRA, 2017).

Devido à incorporação de vários *datasets* em sua formulação, é natural que, ao estimar o custo computacional da multijunção espacial, sejam empregados histogramas espaciais intermediários (SANTOS; OLIVEIRA, 2019), gerados a partir dos histogramas dos *datasets* das etapas iniciais da consulta. Naturalmente, o uso de uma estrutura derivada, construída a partir de uma aproximação do dado original, resulta em perda de precisão nas estimativas (MAMOULIS; PAPADIAS, 2001a).

A Figura 2 ilustra uma situação em que o escalonador utiliza estimativas de custo com baixa precisão para alocar quatro tarefas distintas (J1, J2, J3 e J4) em dois nós do *cluster*. Na Figura 2 (a), são apresentadas as estimativas incorretas para cada uma das tarefas. Com isso, o escalonamento é realizado conforme mostrado na Figura 2 (b). Os custos reais de cada tarefa estão representados na Figura 2 (c). Como resultado, ocorre um escalonamento deficiente, retratado na Figura 2 (d). Assim, um dos nós do *cluster* ficará ocioso, gerando um desperdício de recursos.

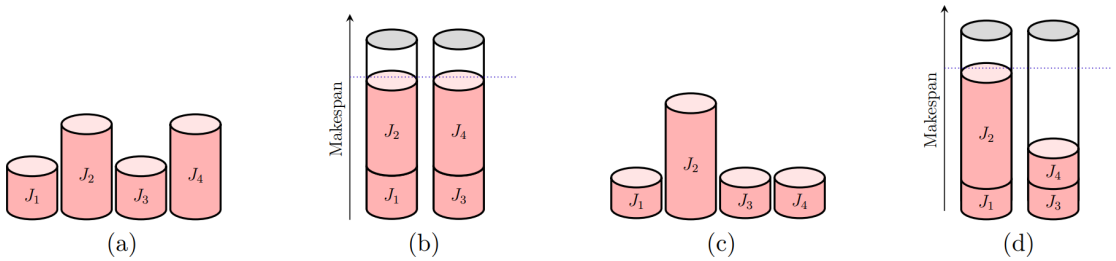


Figura 2 – Cenário no qual a imprecisão da estimativa resulta em um escalonamento deficiente. Fonte: França (2018).

No entanto, até o momento deste trabalho ainda não se entendia completamente qual a intensidade desta imprecisão e a partir de quantas iterações (ou a partir de qual etapa da consulta) a estimativa se degrada ao ponto de não ser mais útil. Resta também entender quais são as principais características dos dados originais que acentuam esse fenômeno. Neste trabalho, foi realizada a experimentação necessária para se obter essa caracterização e entendimento.

Mamoulis e Papadias (2001b) investigaram o erro na estimativa de seletividade em relação ao tamanho da consulta de multijunção, onde o tamanho, nesse contexto, se refere à quantidade de objetos contidos nos *datasets*. Neste trabalho, nossa abordagem difere da anterior devido considerarmos que o tamanho da consulta é determinado pela quantidade de *datasets* da consulta.

1.2 Objetivo do Trabalho

Este trabalho teve como objetivo analisar a estimativa de seletividade ao longo das etapas presentes nas consultas de multijunção espacial. O objetivo principal foi caracterizar a precisão dessa estimativa e identificar os tipos de dados espaciais que influenciam de maneira significativa. Os objetivos específicos são:

1. Elaborar um conjunto experimental envolvendo *datasets* espaciais heterogêneos e consultas de multijunção espacial com várias etapas;

2. Executar as consultas de multijunção espacial e comparar a seletividade real com a estimada;

3. A partir dos resultados, estabelecer o ponto de imprecisão limite, para o qual a estimativa não é mais útil na determinação do custo computacional;

4. Identificar os fatores e características dos dados que acentuam a degradação da precisão.

1.3 Contribuição do Trabalho

Os resultados dos experimentos conduzidos neste estudo permitem avaliar a eficácia dos métodos atualmente empregados no cálculo da estimativa de seletividade. Esta análise crítica, por sua vez, permite a identificação de oportunidades de melhoria nos métodos existentes, abrindo caminho para futuras pesquisas neste domínio. Como a estimativa de seletividade é uma métrica fundamental para avaliar o custo computacional de consultas de multijunção espacial, melhorar sua precisão tem o potencial de reduzir o uso de recursos computacionais no processamento de dados espaciais.

1.4 Organização da Monografia

O trabalho está estruturado em cinco capítulos, descritos brevemente a seguir: o Capítulo 1 apresenta o referencial teórico, onde ocorre a contextualização do objeto de estudo e suas áreas fundamentais, sendo de suma importância para a compreensão deste referido estudo. São abordados conceitos como banco de dados espacias, tipos de consultas existentes e técnicas empregadas para performá-las. No Capítulo 3, detalha-se o processo de busca por trabalhos correlatos, destacando a metodologia adotada, que inclui as strings de busca e os critérios de seleção utilizados. Em seguida, o Capítulo 4 discorre sobre a metodologia empregada na execução dos experimentos, incluindo a configuração do ambiente, os dados utilizados e as consultas de multijunção escolhidas, além de apresentar uma análise dos dados obtidos. Por fim, o Capítulo 5 engloba as conclusões alcançadas neste trabalho e sugere direções para pesquisas futuras.

2 Referencial Teórico

2.1 Introdução

É apresentado neste capítulo assuntos abordados neste trabalho, a fim de facilitar a sua compreensão. A seção 2.2 explica o que são dados espaciais e como são coletados. Na seção 2.3, é explorado o conceito de Sistema de Informação Geográfica (SIG). A seção 2.4 trata das Consultas Espaciais, abordando tanto a Junção quanto a Multijunção. A seção 2.5 discute os Histogramas Espaciais, detalhando o Histograma de Grade e IHWAF. Em seguida, na seção 2.6 é apresentado o Processamento Distribuído de Consultas Espaciais, destacando o porque ele é importante para as consultas de Multijunção Espacial. Por fim, na seção 2.7 é abordado o DGEO, sistema responsável pelo processamento das consultas espaciais e utilizado para realizar os experimentos descritos neste documento.

2.2 Dados Espaciais

Dados espaciais referem-se aos dados (posições, atributos, relações, etc.) pertencentes a entidades de um determinado espaço e que podem ser representados de forma gráfica (FITZ, 2018). Por exemplo, hotéis (objetos espaciais) são comumente vinculados a coordenadas geográficas, como latitude e longitude, além de possuírem diversos atributos de qualidade, como preço e classificação por estrelas (LU; YIU; XIE, 2018).

Estes dados espaciais são adquiridos por meio de dispositivos preparados para captar a energia refletida ou emitida por uma superfície e armazená-la em forma de dados digitais – técnica conhecida como Sensoriamento Remoto. Tais dados podem ser representados em uma estrutura matricial ou vetorial (FITZ, 2018).

A Figura 3a ilustra a estrutura matricial (*raster structure*), a qual consiste em uma matriz M(n, m) com n linhas e m colunas, onde cada célula, conhecida como pixel, contém um valor z que representa uma cor ou um tom de cinza. A Figura 3b, por sua vez, apresenta uma imagem gerada a partir de estruturas vetoriais (*vector structure*), composta por três primitivas gráficas (pontos, linhas e polígonos) e faz uso de um sistema de coordenadas para sua representação. Os pontos são formados por apenas um par de coordenadas, enquanto linhas e polígonos necessitam de um conjunto de pares de coordenadas (FITZ, 2018).

(a) Representação Matricial

(b) Imagem construída a partir de Representação Vetorial

Figura 3 – Exemplos de representações espaciais. Fonte: Google Maps.

2.3 Sistema de Informação Geográfica

Conforme afirmado por Fitz (2018), um Sistema de Informação Geográfica (SIG) é um sistema baseado em computação capaz de capturar, armazenar, modificar, recuperar e visualizar dados espaciais. Ele é composto pelos seguintes elementos:

- **Hardware:** refere-se ao conjunto de dispositivos físicos;
- **Software:** engloba os programas, módulos e sistemas que permitem o processamento e análise dos dados;
- **Dados:** é a matéria-prima resultante do Sensoriamento Remoto;
- **Peopleware:** são os profissionais e usuários envolvidos durante todo o processo.

Após os dados serem coletados, eles são tratados e agrupados com base em suas características. Esse procedimento é essencial para aprimorar a compreensão e utilidade dos mesmos. Imaginamos a camada mais inferior da Figura 4 como um atlas do mundo real. Cada parte desse atlas, por sua vez, corresponde a um conjunto de dados (*dataset*), sendo eles terreno, elevação, divisões, ruas e pessoas. Além de ocuparem menos espaço de armazenamento, torna notavelmente mais ágil o processo de combinar informações (consultas espaciais). Isso resulta em um processamento mais eficiente, evitando assim, um dispêndio excessivo de recursos computacionais (CAMPBELL; SHIN, 2012).

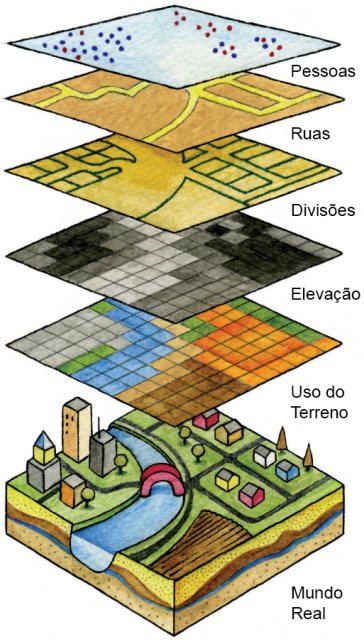


Figura 4 – Divisão do mundo real em vários *datasets*. Fonte: Campbell e Shin (2012).

2.4 Consultas Espaciais

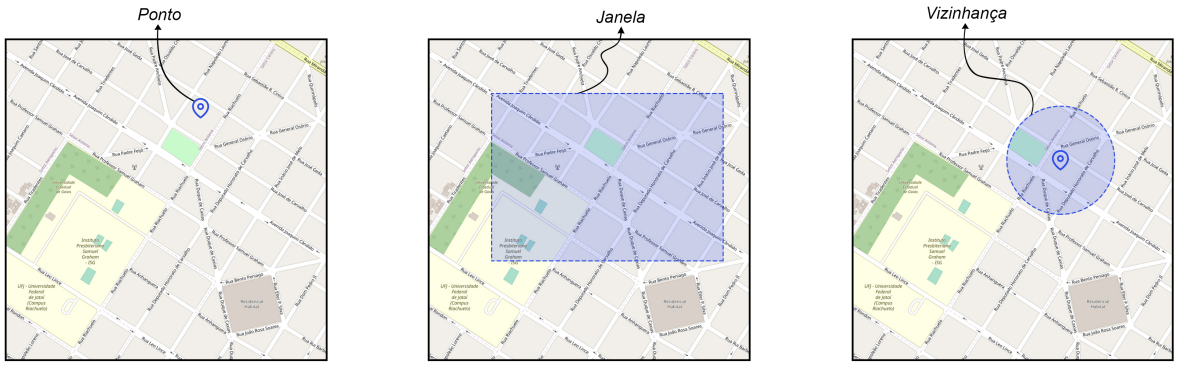
A Análise Espacial consiste na habilidade de manipular dados espaciais, com o propósito de enriquecê-los, revelar informações não visíveis, bem como auxiliar na tomada de decisões. Em essência, trata-se do processo de converter dados espaciais brutos em informações valiosas (LONGLEY et al., 2010).

Consultas Espaciais (*Spatial Queries*) representam as operações essenciais na análise espacial. Elas possibilitam uma investigação detalhada dos atributos topológicos e geométricos de objetos espaciais a partir de conjuntos individuais ou múltiplos de dados (CAMPBELL; SHIN, 2012).

Uma parte do SIG denominada de Sistema Gerenciador de Banco de Dados Espaciais (SGBDE) é responsável pelo armazenamento, atualização e recuperação dos dados espaciais (FITZ, 2018). Um SGBDE difere de um SGBD Relacional (SGBDR) comum não apenas porque necessita manipular objetos espaciais mas também pelo grande conjunto de comandos que estende o SQL (*Structured Query Language*) padrão (HUISMAN; BY, 2009).

Os SGBDEs são capazes de realizar consultas de tipos diferentes. Entretanto, as principais consultas são as de: Ponto, Janela, Vizinhança, Junção e Multijunção (CAMPBELL; SHIN, 2012). Na Figura 5a, é apresentada uma consulta de ponto, a qual recupera o objeto cuja coordenada corresponda exatamente à do ponto em questão. Por sua vez, na Figura 5b, é ilustrada uma consulta de janela, a qual, ao receber uma seleção geométrica, comumente retangular, retorna um conjunto de objetos contidos nessa área específica.

Por fim, na Figura 5c, é exemplificada uma consulta de vizinhança, que retorna todos os objetos situados dentro de um limite de distância, também conhecido como raio. Ambas as consultas são aplicadas a um único *dataset* (ACHARYA; POOSALA; RAMASWAMY, 1999).



(a) Consulta de Ponto (b) Consulta de Janela (c) Consulta de Vizinhança

Figura 5 – Exemplos de consultas espaciais.

2.4.1 Junção

Em um banco de dados relacional as tabelas são unidas através de chaves pré determinadas. A chave primária (*primary key*) que identifica exclusivamente um registro (linha) de uma tabela corresponde exatamente à chave estrangeira (*foreign key*) de um registro da segunda tabela. Logo, ao realizar uma consulta de junção, basta o SGBD comparar os valores dessas chaves (CAMPBELL; SHIN, 2012).

Entretanto, uma Junção Espacial (*Spatial Join*) é mais complexa de ser executada quando comparada à uma junção em banco de dados relacional, já que aplica-se um predicado θ, usualmente a intersecção, a fim de identificar objetos correlacionados entre dois *datasets* (MAMOULIS; PAPADIAS, 2001b). A formalização da junção espacial pode ser encontrada na Definição 1 (BRINKHOFF; KRIEGEL; SEEGER, 1996).

Definição 1 (Junção) *Sejam* $A = \{a_1, a_2, \ldots, a_n\}$ *e* $B = \{b_1, b_2, \ldots, b_n\}$ *dois* datasets *distintos de objetos multidimensionais, então se existem objetos* $a \in A$ *e* $b \in B$ *que satisfazem um predicado* θ*, é possível realizar uma* θ*-junção espacial entre* A*,* B*. Formalmente uma junção espacial pode ser definida como uma função*

$$A \bowtie B : A \times B \rightarrow R \tag{2.1}$$

sendo $R = \{(a, b) : a \in A \wedge b \in B\}$ *os objetos resultantes da junção, em outras palavras*

$$A \bowtie B = \{(a, b) \mid a \in A \wedge b \in B \wedge a\theta b\}. \tag{2.2}$$

Aqui estão alguns exemplos de predicados espaciais que desempenham um papel fundamental na análise espacial (CAMPBELL; SHIN, 2012):

- **Intersecção:** retorna os objetos do *dataset* de destino que compartilham uma localização em comum com o *dataset* de origem;
- **Contido:** retorna os objetos que se encontram completamente contidos no *dataset* de origem, logo, objetos com limites coincidentes não são selecionados por este predicado;
- **Idêntico:** retorna os objetos que possuem exatamente a mesma localização.

Para otimizar o uso de recursos computacionais, a literatura propôs dividir a junção espacial em dois estágios distintos: filtro e refinamento. Ambos empregam algoritmos de geometria computacional para atingir seus objetivos. No estágio de filtro, cada objeto espacial é aproximado pelo menor retângulo em que está contido, conhecido como MBR (*Minimum Bounding Rectangle*), o qual é formado pelas as coordenadas mínima e máxima do objeto original, conforme ilustrado na Figura 6. Contudo, é importante salientar que no estágio de filtragem podem surgir falsos positivos. Isso ocorre porque o MBR de um objeto pode se sobrepor ao MBR de outro, embora na realidade isso não seja verdade. Esse fenômeno é denominado de *área morta* e está ilustrado na Figura 6.

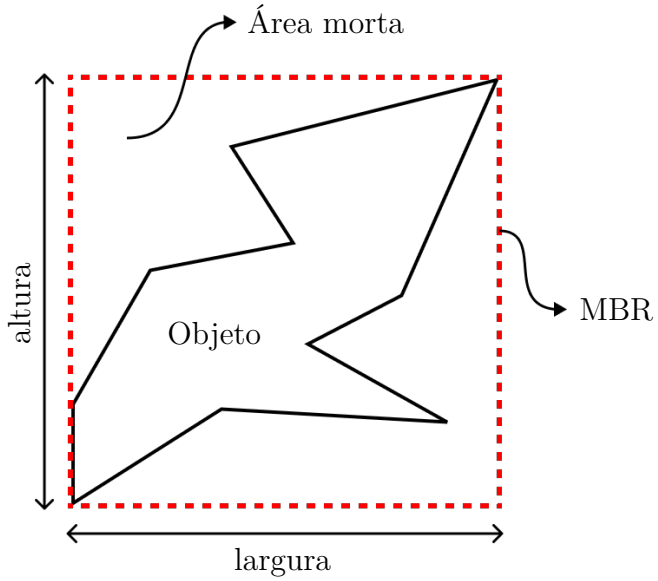


Figura 6 – Exemplo de MBR.

Dessa forma, o estágio de refinamento entra em cena, fazendo uso de algoritmos específicos para identificar e remover objetos falsos que foram erroneamente incluídos durante o estágio de filtragem. Portanto, o mesmo atua como uma camada adicional de

validação, assegurando a confiabilidade dos resultados finais (JACOX; SAMET, 2007). Isso implica em um alto custo computacional quando comparado ao estágio anterior. Logo, para aprimorar a eficiência no processamento das operações de junção, é viável considerar a implementação de sistemas distribuídos (BRINKHOFF; KRIEGEL; SEEGER, 1996).

Considerando a Figura 7, a qual ilustra dois *datasets*, $R = \{r1, r2, r3\}$ e $S = \{s1, s2, s3\}$. Ao realizar uma consulta de junção para identificar todos os pares de objetos que fazem interseção, ou seja, para cada objeto $r \in R$, encontre os objetos $s \in S$ que se sobrepõe, o seguinte conjunto de tuplas será retornado: $Q = \{(r1, s2), (r2, s2), (r2, s3), (r3, s2)\}$.

Um exemplo de aplicação real da junção espacial é descobrir as pontes de uma determinada região. Para isso, basta aplicar um predicado de intersecção nos *datasets* de rodovias e rios (BOUROS; MAMOULIS, 2019).

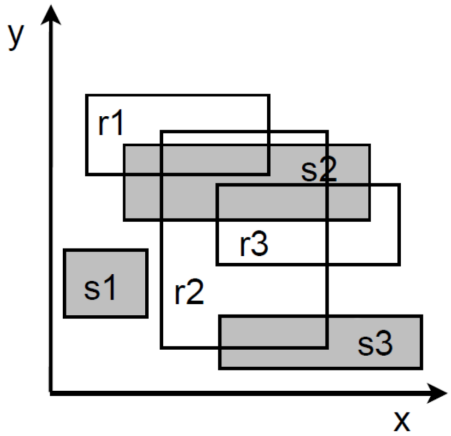


Figura 7 – Exemplo de Junção Espacial. Fonte: Jacox e Samet (2007).

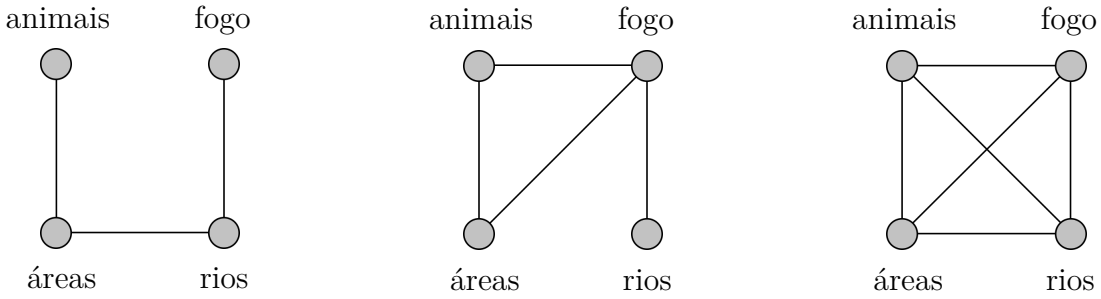
2.4.2 Multijunção

Ao contrário da junção espacial, que envolve apenas duas entradas, a Multijunção Espacial (*Multiway Spatial Join*) permite utilizar mais de dois *datasets* na mesma consulta (MAMOULIS; PAPADIAS, 2001b). Um exemplo característico desse tipo de consulta é: "Encontrar todas as espécies de animais que moram em áreas de preservação que foram danificadas por um incêndio na beira de um rio". Para realizar esta consulta é necessário combinar quatro *datasets*, sendo eles: Animais, Áreas de Preservação, Queimadas e Rios (OLIVEIRA, 2017). Formalmente, a multijunção espacial pode ser expressa conforme a Definição 2.

Definição 2 (Multijunção) *Dado um conjunto de* datasets $D = \{D_1, ..., D_n\}$*, cada um contendo um conjunto de registros* $r^i_1, ..., r^i_{m_i}$*,* $1 \leqslant i \leqslant n$ *e* m_i *sendo a cardinalidade de* D_i*, e um conjunto de predicados espaciais* $P = \{\theta_{ij} \mid \forall i, j,\ 1 \leqslant i, j \leqslant n\}$*, a consulta recupera todas as* $n-$*tuplas* $(r^1_p, ..., r^i_k, ..., r^j_l, ..., u^n_r)$ *tais que cada predicado* θ_{ij} *é válido*

quando aplicado aos seus respectivos elementos na n−tupla, com p, k, l *e* r *referindo-se a registros específicos de seus respectivos* datasets*,* $1 \leqslant k \leqslant m_i$*,* $1 \leqslant l \leqslant m_j$ *e analogamente para* p *e* r *(OLIVEIRA et al., 2023).*

A representação da multijunção espacial pode ser concebida como um grafo $G = (V, E)$, sendo V o conjunto de vértices e E o conjunto de arestas, onde os vértices representam os *datasets* e as arestas os predicados. A Figura 8 ilustra três exemplos de grafos de consulta. Cada sub-figura representa um tipo distinto de consulta, as quais estão classificadas de acordo com as propriedades dos grafos. Na Figura 8a, é apresentada uma consulta em forma de árvore ou cadeia, um tipo comum de consulta na análise espacial, onde todos os conjuntos de dados são combinados aos pares, sem repetições. Já a Figura 8b ilustra uma consulta de ciclo, sendo que o grafo possui as propriedades que o tornam cíclico. Por fim, a Figura 8c demonstra uma consulta do tipo clique, cuja característica é relacionar todos os *datasets* entre si.

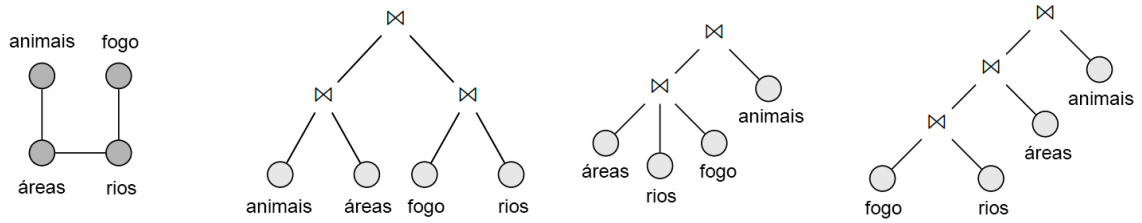


(a) Árvore ou Cadeia (b) Ciclo (c) Clique

Figura 8 – Tipos de consulta de multijunção espacial. Fonte: (OLIVEIRA, 2017).

Uma consulta de multijunção espacial pode ser processada de várias maneiras distintas, conhecidas como planos de execução. Cada plano de execução estabelece uma sequência específica de etapas, bem como os algoritmos e *datasets* que serão utilizados para calcular o resultado intermediário em cada etapa.

A Figura 9 ilustra o grafo e três planos alternativos para a seguinte consulta de cadeia (*chain*): "encontrar espécies de animais que vivem em áreas de preservação ambiental cortadas por rios e que foram destruídas por queimadas". Na Figura 9b, a etapa inicial da consulta combina os *datasets* em pares, gerando dois resultados intermediários que são posteriormente unificados em uma segunda etapa. Na Figura 9c, três *datasets* são unidos com objetivo de gerar um resultado intermediário, o qual será combinado na segunda etapa com um quarto *dataset*. Já na Figura 9d, dois *datasets* são unidos na etapa inicial e posteriormente a cada etapa um novo *dataset* é combinado com o resultado da junção anterior.



(a) Grafo (b) Plano 1 (c) Plano 2 (d) Plano 3

Figura 9 – Planos de execução alternativos para a multijunção espacial (OLIVEIRA, 2017).

Todos os planos para uma consulta são equivalentes semanticamente, ou seja, retornam o mesmo conjunto de respostas como resultado. Porém, diferem entre si em relação ao uso de recursos computacionais como, por exemplo, tempo de processamento e tráfego de rede (OLIVEIRA, 2017).

Mamoulis e Papadias (2001b) analisaram o número de planos de execução em um contexto serial (não paralelo, não distribuído). Eles demonstraram que essa variedade é uma função que depende de três fatores: da quantidade de *datasets* de entrada, do tipo de consulta e do conjunto de algoritmos de junção que podem ser aplicados em cada etapa. Um exemplo que deram é que considerando três algoritmos de junção, para uma consulta do tipo clique com cinco *datasets* de entrada, existem 100 combinações possíveis de planos de execução.

2.5 Histogramas Espaciais

Os SGBDRs contam com um componente crucial conhecido como otimizador de consulta. Esse mecanismo utiliza a estimativa de seletividade, ou seja, a quantidade de registros que serão retornados pela operação, para determinar a forma mais eficiente de executar as consultas (escolha do plano de execução). O tamanho estimado da consulta é também utilizado para informar o usuário sobre o tempo necessário para executar a consulta, mesmo antes de sua realização. Dado que executar a consulta completa para calcular a quantidade de resultados seria impraticável, a maioria dos sistemas recorre a métodos de aproximação e estimativas para avaliar o tamanho esperado da consulta.

Essa abordagem não é diferente para os SGBDEs. No entanto, estimar a seletividade em bancos de dados espaciais é uma tarefa desafiadora, dado que os objetos estão distribuídos de forma não uniforme no espaço. Portanto é fundamental simplificar os *datasets*. A literatura apresenta uma variedade de técnicas com esse intuito, porém, a mais frequentemente empregada é o histograma espacial (MAMOULIS; PAPADIAS, 2001a), já que, segundo Acharya, Poosala e Ramaswamy (1999), é de fácil construção e consome pouca memória.

Essencialmente, um histograma é uma estrutura de dados que fragmenta o espaço geográfico do *dataset* em células e aloca *buckets* para cada uma delas. Dentro de cada *bucket* são mantidos dados sobre os objetos contidos no fragmento de espaço representado por aquela célula, como por exemplo, a quantidade de objetos (cardinalidade) e o comprimento médio deles (OLIVEIRA, 2017).

2.5.1 Histograma de Grade

Mamoulis e Papadias (2001a) em sua pesquisa sobre a seletividade em consultas espaciais complexas, propuseram o uso de histogramas uniformes para lidar com múltiplas junções espaciais e seleções. Esses histogramas dividem a superfície do *dataset* em uma grade constituída por várias células de tamanho fixo, sendo que cada célula representa uma pequena região do *dataset*.

O Histograma de Grade oferece vantagens significantes, incluindo uma construção simplificada, fácil manutenção e eficiência de tempo para estimar a saída das consultas (OLIVEIRA, 2017). No entanto, uma desvantagem notável é a baixa precisão da estimativa de seletividade para dados distorcidos. Embora aumentar o número de células possa melhorar a precisão, também aumenta a quantidade de memória necessária para armazenar a estrutura do histograma (MAMOULIS; PAPADIAS, 2001b).

Uma questão crucial na construção de histogramas multidimensionais é como contabilizar os objetos que estão em cada célula, um processo conhecido como *hashing*. Mamoulis e Papadias (2001a) definiram o valor da cardinalidade de cada célula com base na quantidade de objetos espaciais que possuem o centro do seu MBR dentro dos limites da célula.

Para aprimorar a precisão da estimativa de custos com histogramas de grade, Mamoulis e Papadias (2001b) sugeriram o uso do comprimento médio dos objetos como um metadado adicional para cada célula do histograma, considerando apenas a área dentro dos limites da célula para objetos que se sobrepõem várias células.

2.5.2 IHWAF

A construção desse histograma utiliza como base o que foi proposto por Mamoulis e Papadias (2001a), porém com a implementação de novas técnicas. Uma delas é o método da sobreposição proporcional, que ao contrário do histograma de grade convencional, onde é utilizado apenas o centro do MBR para decidir se o objeto faz interseção com a célula. Este novo método verifica o quanto um objeto está sobreposto em uma célula, sendo armazenado no *bucket* o percentual dessa sobreposição. Por exemplo, se um objeto sobrepõe

30% de uma célula e 70% de outra, ao realizar a soma dos valores resulta no objeto total, ou seja, 100%. Logo, o problema da contagem múltipla é reduzido (OLIVEIRA, 2017).

Além disso, Oliveira (2017) também identificou por meio de um estudo que o tipo dos objetos contidos nos *datasets* influenciam na estimativa de seletividade. Com isso, propôs um novo modelo estatístico que apresentou uma melhoria dos resultados, o qual leva em consideração o tipo dos objetos.

O histograma IHWAF foi desenvolvido pensando no processamento distribuído e no problema de escalonamento de consultas de multijunção espacial. Portanto, dentre vários outros metadados, cada célula armazena a sua localização em relação ao *cluster*. Foi proposto também modelos capazes de controlar o *tradeoff* entre os custos de processamento e comunicação.

2.6 Processamento Distribuído de Consultas Espaciais

Um sistema distribuído é uma infraestrutura composta por múltiplos computadores independentes que, quando utilizados, são percebidos pelos usuários como sendo uma única entidade. Dentro dessa categoria, existe um subgrupo de sistemas voltados para tarefas de alto desempenho. Um exemplo notável é a computação em *cluster*, que consiste em um conjunto de nós (máquinas dedicadas ou computadores pessoais similares) interconectados por uma rede local de alta velocidade. Cada nó dentro do *cluster* executa o mesmo sistema operacional, o que facilita a coordenação e o compartilhamento de recursos (STEEN; TANENBAUM, 2017).

Um sistema de banco de dados centralizado opera concentrando o processamento de solicitações e posteriormente remetendo os resultados aos usuários. Essa abordagem oferece algumas vantagens, como facilidade de gerenciamento e proteção dos dados. No entanto, sua eficácia depende fortemente do desempenho do servidor, especialmente quando enfrenta um grande volume de solicitações. Para otimizar o desempenho e lidar com demandas escaláveis, a computação em *cluster* foi introduzida como solução (ÖZSU; VALDURIEZ, 2011).

Özsu e Valduriez (2011) definiu um sistema de banco de dados distribuído sendo uma coleção de vários bancos de dados logicamente interconectados e distribuídos através de uma rede de computadores. Neste contexto, um Sistema Gerenciador de Banco de Dados Distribuído (SGBDD) é responsável por facilitar a administração do banco de dados distribuído e ao mesmo tempo, assegurar que a distribuição do sistema permaneça transparente para os usuários.

O processamento de consultas de multijunção espacial frequentemente apresenta um alto custo, principalmente devido ao grande volume de dados envolvidos e a complexidade

dos objetos espaciais. Para melhorar o desempenho, uma solução é a utilização de um sistema distribuído. Diversos algoritmos foram propostos para o processamento de consultas de multijunção espacial em ambientes computacionais paralelo ou distribuídos. Grande parte desses algoritmos empregam uma estratégia de particionamento de dados, conhecida como *declustering*, que fragmenta os objetos contidos no *dataset* em grupos denominados de partições ou células (OLIVEIRA, 2017).

2.7 DGEO

Para conduzir os experimentos deste trabalho, faremos uso do DGEO, uma suíte de aplicações desenvolvida para realizar consultas espaciais. A mesma encontra-se disponível no seguinte repositório do GitHub: <https://github.com/thborges/dgeoijgis>.

O DGEO foi concebido durante um projeto de pesquisa na Universidade Federal de Goiás, sendo escrito nas linguagens de programação C e Go. Para maximizar a eficiência, foram incorporadas *threads* nativas que exploram múltiplos níveis de paralelismo, bem como um sistema de processamento distribuído com protocolos próprios, incluindo a serialização Gob sobre *sockets* TCP.

O destaque do DGEO reside na capacidade de processar junções espaciais de forma paralela e distribuída, conferindo uma notável eficiência quanto ao processamento de dados espaciais. Todos os métodos e algoritmos implementados utilizam as bibliotecas GDAL (*Geospatial Data Abstraction Library*), disponível sob uma licença de código aberto estilo MIT em <https://www.gdal.org/>, e a GEOS (*Geometry Engine*), disponível sob os termos da licença *GNU Lesser General Public License* (LGPL) em <https://libgeos.org> (OLIVEIRA et al., 2023).

3 Trabalhos relacionados

3.1 Introdução

Com o intuito de identificar publicações relacionados a este trabalho, foi realizado um levantamento por estudos que tratam da estimativa de seletividade em consultas espaciais com vários níveis de junção.

3.2 Critérios de busca

Os trabalhos apresentados foram selecionados através do motor de busca *Google Scholar*. Utilizamos a seguinte *string* de busca: "multiway spatial join" AND "selectivity estimation". Para garantir a inclusão de publicações relevantes que possam não ter sido identificadas pela busca, realizou-se uma análise manual das referências dos artigos previamente recuperados.

3.3 Metodologia de análise

Para garantir a seleção de trabalhos relevantes para este estudo, foram elaborados e aplicados os seguintes critérios:

3.3.1 Multijunção Espacial (C1)

O primeiro critério visa filtrar trabalhos que tratam da consulta de multijunção espacial, a qual é de particular relevância para este estudo, uma vez que envolve múltiplas etapas de processamento e gera resultados intermediários.

3.3.2 Estimativa de Seletividade (C2)

O segundo critério abrange os trabalhos que exploram a estimativa de seletividade, o que é fundamental, uma vez que este presente estudo se propõe a avaliar especificamente a precisão desse cálculo ao longo das etapas da multijunção espacial.

3.3.3 Histogramas Multidimensionais (C3)

O terceiro critério destaca os trabalhos que incorporam o uso de histogramas espaciais na estimativa de seletividade. Este critério é crucial, uma vez que o objetivo deste estudo é identificar padrões nos histogramas que contribuem para a deterioração da precisão na estimativa de seletividade.

3.3.4 Avalia a precisão os longo das etapas (C4)

O quarto e último critério, porém igualmente relevante, foca na seleção de estudos que examinam a precisão da estimativa de seletividade em cada fase da consulta de multijunção espacial, em vez de apenas considerar o erro global. Esta abordagem mais detalhada permite uma análise minuciosa das diferentes etapas do processo da consulta, proporcionando *insights* valiosos sobre a eficácia e os pontos críticos.

3.4 Trabalhos analisados

Com base nos critérios de busca estabelecidos, foram encontrados cerca de 36 (trinta e seis) trabalhos relevantes. Destes, 4 (quatro) foram selecionados de acordo com os critérios definidos na seção anterior e são discutidos a seguir.

3.4.1 Selectivity Estimation of Complex Spatial Queries (T1)

Mamoulis e Papadias (2001a) abordaram a otimização de consultas espaciais complexas, destacando a importância de estimar com precisão a seletividade para consultas que envolvem múltiplas junções e seleções espaciais. Como resultado da pesquisa, eles desenvolveram fórmulas para estimar a saída dessas consultas e conduziram experimentos para avaliar a precisão dessas fórmulas. A análise consistiu em comparar a seletividade estimada com os resultados reais das consultas, revelando erros relativos de 8% para junções binárias e 38% para consultas de quatro entradas.

Além disso, os autores expandiram seu método para lidar com dados distorcidos da vida real usando histogramas 2D, comparando a precisão do método baseado em histogramas com a aplicação direta de fórmulas que presumem uniformidade. Eles também ressaltaram a aplicabilidade dos modelos propostos na otimização de consultas que envolvem dados espaciais e não espaciais, assim como em conjuntos de dados com distribuição não uniforme.

3.4.2 Multiway Spatial Joins (T2)

Mamoulis e Papadias (2001b) conduziram um estudo abrangente sobre multijunção espacial. Neste trabalho, foi demonstrado como é possível combinar os algoritmos de junção simples para utilizar vários *datasets* de entrada. Propuseram métodos de particionamento de dados, além de técnicas para enumerar os planos de execução, estimar custos e selecionar um bom plano. Adicionalmente, apresentaram uma fórmula para calcular a estimativa de seletividade tanto da junção quanto da multijunção espacial.

Por fim, executaram experimentos para analisar a precisão desta estimativa. No caso da multijunção, foi considerado diferentes tipos de consultas, como a de cadeia, ciclo e clique, para dimensões variadas de grades. Entretanto, as consultas possuem apenas 3 etapas e não foi avaliado o erro da estimativa ao longo dessas etapas, somente o erro global.

3.4.3 Selectivity Estimation for Spatial Joins with Geometric Selections (T3)

O trabalho de Sun, Agrawal e Abbadi (2002) introduziu o conceito do Histograma de Euler como uma abordagem para calcular a estimativa de seletividade em consultas de junção espacial. Eles demonstraram que, quando os *datasets* tem suas grades alinhadas e os objetos também se alinham com suas respectivas grades, a seletividade pode ser calculada sem erro. No entanto, em banco de dados que lidam com objetos reais, é improvável que as grades dos histogramas se alinhem devido às diferentes dimensões espaciais dos *datasets*.

Para resolver esse problema, os autores propuseram a Suposição de Quantização, um método que realinha os objetos dos *datasets* por meio de uma transformação, embora isso possa aumentar o tempo de processamento devido ao tamanho dos objetos. Além disso, Sun, Agrawal e Abbadi (2002) apresentaram uma abordagem mais geral do Histograma de Euler, permitindo a manipulação de objetos e janelas de seleção que não estão alinhadas com uma grade específica. Essa generalização foi baseada na teoria dos grafos, alocando *buckets* para faces, arestas e vértices, evitando assim o problema da contagem múltipla de objetos.

3.4.4 Efficient Processing of Multiway Spatial Join Queries in Distributed Systems (T4)

Oliveira (2017) realizou uma análise detalhada das consultas de multijunção espacial, com ênfase em sistemas distribuídos. Ele desenvolveu e avaliou um novo modelo de custo

para estimar o consumo de recursos em consultas de multijunção distribuídas. No estudo, descreveu a construção de histogramas intermediários para consultas mais complexas e propôs fórmulas estatísticas para estimar a seletividade com mais precisão, considerando a natureza dos objetos envolvidos. Introduziu o método de sobreposição proporcional como uma abordagem para minimizar o erro da contagem múltipla, além do histograma IHWAF, que implementa esse método ao particionar o conjunto de dados em uma grade de tamanho variável.

O histograma IHWAF foi concebido com foco no processamento distribuído e na questão do escalonamento de consultas de multijunção espacial. Cada célula do histograma armazena, entre outros metadados, sua localização em relação ao *cluster*. Também foram propostos modelos para equilibrar os custos de processamento e comunicação. Experimentos forma conduzidos, com consultas de multijunção envolvendo até 7 *datasets* de entrada. No entanto, a precisão da estimativa de seletividade ao longo das etapas não foi avaliada, apenas a precisão da comunicação e execução em relação à vários planos.

3.5 Resumo Comparativo

Os trabalhos analisados destacam diversas abordagens na estimativa de seletividade em consultas espaciais complexas. Por exemplo, Mamoulis e Papadias (2001a) desenvolveram fórmulas para esta estimativa e conduziram experimentos para avaliar a suas respectivas precisões, inclusive considerando dados distorcidos da vida real. Além disso, algumas pesquisas abordam questões específicas, como o desenvolvimento de técnicas para reduzir o erro da contagem múltipla. Por exemplo, Sun, Agrawal e Abbadi (2002) propuseram o Histograma de Euler, enquanto Oliveira (2017) introduziu o histograma IHWAF e o método da sobreposição proporcional.

Enquanto alguns estudos se concentram mais na criação de métodos para estimar a seletividade com maior exatidão, outros direcionam seu foco para a eficiência do processamento, é o caso do trabalho de Mamoulis e Papadias (2001b) em que é apresentado um estudo abrangente sobre multijunção espacial, fornecendo métodos de particionamento de dados, além de técnicas para enumerar os planos de execução, estimar custos e selecionar um bom plano. E do trabalho de Oliveira (2017), que por sua vez concentra-se na eficiência do processamento de consultas de multijunção espacial em ambientes distribuídos, propondo modelos para equilibrar os custos de processamento e comunicação.

A Tabela 1 apresenta uma comparação entre os trabalhos analisados T1-T4, este referido trabalho TR e os critérios de seleção C1-C4. Observa-se que todos os trabalhos atendem os critérios C1 e C2, enquanto apenas o trabalho T3 não atende ao critério C3, e apenas o TR atende a todos os critérios.

Tabela 1 – Comparativo entre trabalhos.

	C1	C2	C3	C4
T1	Sim	Sim	Sim	Não
T2	Sim	Sim	Sim	Não
T3	Não	Sim	Sim	Não
T4	Sim	Sim	Sim	Não
TR	Sim	Sim	Sim	Sim

Dessa forma, os trabalhos elencados neste capítulo oferecem uma gama de técnicas que visam melhorar as consultas espaciais complexas. Desde o desenvolvimento de fórmulas mais precisas até a introdução de abordagens para diminuir o desperdício de recursos computacionais.

No entanto, é relevante ressaltar que, apesar dos avanços apresentados, nenhum dos trabalhos realiza uma análise específica sobre a precisão da estimativa de seletividade ao longo das diversas etapas envolvidas na multijunção espacial. Diante dessa lacuna na literatura, o presente trabalho teve como objetivo realizar esta análise detalhada, contribuindo assim para um entendimento completo e refinado.

4 Avaliação e Testes

4.1 Introdução

Neste capítulo, detalharemos a metodologia adotada para conduzir os experimentos da pesquisa. Na seção 4.2 descreveremos as configurações do ambiente de execução, os conjuntos de dados utilizados, as consultas espaciais realizadas, bem como os histogramas e métodos empregados. Em seguida, na seção 4.4 apresentaremos os resultados dos experimentos realizados, incluindo uma análise do comportamento da estimativa de seletividade.

4.2 Ambiente Experimental

Os experimentos foram executados no próprio computador pessoal do pesquisador, tendo suas especificações descritas na Tabela 2. Foi utilizado o software DGEO, emulando um *cluster* virtual constituído por dois nós escravos e um nó mestre, os quais se comunicam através da interface de *loopback*.

Tabela 2 – Especificações do computador utilizado para a realização dos experimentos.

Item	**Descrição**
Processador	Intel Core i7-10750H
Cache	12MB
Geração	10^a
Frequência Base	2.6GHz
Frequência Máxima	5.0GHz
Núcleos	6
Threads	12
Memória RAM	32GB DDR4 3200MHz
Memória SWAP	100GB SSD NVMe

4.2.1 Datasets

Para avaliar a proposta apresentada, optou-se por utilizar um conjunto de *datasets* reais e públicos, obtidos a partir da Infraestrutura Nacional de Dados Espaciais (INDE)[1], do

[1] <https://visualizador.inde.gov.br/>

Laboratório de Processamento de Imagens e Geoprocessamento (LAPIG)[2] e do GIS-Lab[3]. Os *datasets* utilizados, juntamente com suas respectivas características, encontram-se detalhados na Tabela 3. Todos os *datasets* estão no formato Shapefile (SHP) e são constituídos por representações bidimensionais de entidades na superfície terrestre.

É relevante salientar que, devido a natureza dos objetos pontuais, não foram utilizados *datasets* que contenham esse tipo de objeto, pois o mesmo não introduz erro na estimativa de seletividade. Isso se deve ao fato de que, ao ser composto por apenas um par de coordenadas, o predicado de uma consulta de junção é satisfeito ou não, diferentemente dos objetos do tipo linha e polígono que podem causar falsos positivos.

Tabela 3 – *Datasets* a serem utilizados nos experimentos.

Nome	**Sigla**	**Cardinalidade**	**Área**	**Tamanho SHP (MB)**
		Linhas		
Estradas	ES	563.492	46.666,2	105,9
Ferrovias	FE	194.261	45.281,6	28,7
Hidrografia	HI	226.914	1.835,4	64,5
Limites Políticos	LP	8.079	64.800,0	12,6
Linhas de Transmissão	LT	3.807	972,0	1,3
Pistas de Avião	PA	2.035	1.449,6	<1,0
Relevo	RE	703.574	62.117,2	572,5
Rios	RI	943.638	49.904,2	243,2
Rodovias	RO	51.593	1.937,3	15,2
Trilhas	TR	52.185	44.778,9	14,3
		Polígonos		
Águas Interiores	AI	338.860	55.617,5	136,7
Alertas de Desmatamento	AD	32.578	395,1	11,3
Áreas Urbanizadas	AU	128.459	1.579,4	57,7
Bacias	BC	4.319	1.525,5	125,2
Bancos de Areia	BA	3.181	740,5	1,7
Florestas	FL	167.394	45.835,7	101,5
Mineração	MI	240.527	3.612,1	323,3
Municípios	MU	5.570	1.761,2	97,1
Regiões Rurais	RR	104	1.762,1	18,9
Vegetação	VE	145.458	1.761,2	461,5

[2] <https://lapig.iesa.ufg.br/>
[3] <http://gis-lab.info/qa/vmap0-eng.html>

4.2.2 Consultas

Para que fosse possível avaliar o efeito de propagação do erro da estimativa para as próximas etapas da multijunção, foram realizadas consultas com dez *datasets* de entrada, sendo eles compostos por objetos do tipo linha e polígono. Propõe-se o uso das combinações de tipos e ordem dos tipos apresentados na Tabela 4, que permite uma investigação suficiente sobre como os tipos de dados interferem no erro de seletividade das multijunções espaciais.

Tabela 4 – Abordagem para avaliar a estimativa.

Característica	**Consulta**
Apenas linha	$L \bowtie L \bowtie L \bowtie L \bowtie L \bowtie L \bowtie L \bowtie L \bowtie L$
Apenas polígono	$P \bowtie P \bowtie P \bowtie P \bowtie P \bowtie P \bowtie P \bowtie P \bowtie P$
Início linha	$L \bowtie L \bowtie L \bowtie L \bowtie L \bowtie P \bowtie P \bowtie P \bowtie P \bowtie P$
Início polígono	$P \bowtie P \bowtie P \bowtie P \bowtie P \bowtie L \bowtie L \bowtie L \bowtie L \bowtie L$
Intercalado	$L \bowtie P \bowtie L \bowtie P \bowtie L \bowtie P \bowtie L \bowtie P \bowtie L \bowtie P$

A Tabela 5 apresenta as consultas de multijunção espacial que foram executadas a fim de obter a seletividade estimada e a real, considerando a abordagem da Tabela 4 e os *datasets* listados na Tabela 3. Estas consultas são exclusivamente do tipo cadeia, ou seja, na primeira etapa há a junção de dois *datasets*, e nas etapas subsequentes é realizada a junção entre o resultado intermediário da etapa anterior mais um novo *dataset*.

Adicionalmente, as estimativas para as etapas das multijunções espaciais foram calculadas utilizando os métodos IHWAF e MP, propostos por Oliveira (2017) e Mamoulis e Papadias (2001a) respectivamente.

Tabela 5 – Multijunções espaciais que foram executadas.

Nome	**Consulta**
q_1	$LP \bowtie ES \bowtie RI \bowtie FE \bowtie RE \bowtie LT \bowtie HI \bowtie RO \bowtie PA \bowtie TR$
q_2	$AI \bowtie VE \bowtie RR \bowtie BA \bowtie AU \bowtie MI \bowtie AD \bowtie MU \bowtie BC \bowtie FL$
q_3	$ES \bowtie LP \bowtie RI \bowtie FE \bowtie LT \bowtie RR \bowtie FL \bowtie VE \bowtie BC \bowtie MI$
q_4	$VE \bowtie MI \bowtie BC \bowtie RR \bowtie FL \bowtie LP \bowtie LT \bowtie FE \bowtie PA \bowtie ES$
q_5	$FE \bowtie MU \bowtie LT \bowtie RR \bowtie ES \bowtie MI \bowtie RI \bowtie FL \bowtie PA \bowtie VE$

4.3 Métricas

Considerando que o objetivo central não reside na avaliação do erro global da estimativa de seletividade para as consultas, mas sim na análise individual de cada etapa, durante a realização dos experimentos, foram registradas a quantidade estimada e real de objetos retornados em cada etapa das consultas. Esses dados foram posteriormente aplicados na Equação 4.1.

$$\lambda_{q_i,s_j} = \frac{|r_{q_i,s_j} - e_{q_i,s_j}|}{r_{q_i,s_j}} \times 100 \tag{4.1}$$

A Equação 4.1 foi proposta com o intuito de calcular o erro relativo em termos percentuais para cada etapa das consultas. Nesta equação, q_i representa uma consulta do conjunto Q e s_j representa uma etapa do conjunto S, sendo r a seletividade real e e a estimada.

A representação do erro em porcentagem é de suma importância, pois simplifica a comparação entre diferentes tipos de consultas e tamanhos variados de *datasets*. O resultado da fórmula é intuitivo e de fácil interpretação: um valor baixo indica que a estimativa está próxima do esperado, enquanto um valor alto revela uma grande discrepância entre os dois.

4.4 Análise dos Resultados Obtidos

A Tabela 6 lista os tamanhos dos histogramas gerados a partir de cada um dos *datasets* no momento em que foram carregados através da aplicação cliente do DGEO. Estes histogramas foram utilizados para estimar a seletividade de uma ou mais etapas das consultas. Sabe-se que na primeira etapa são utilizados dois histogramas de seus respectivos *datasets*. E nas etapas subsequentes, por outro lado, emprega-se o histograma intermediário, gerado a partir dos histogramas da etapa anterior, juntamente com o histograma de um novo *dataset*. Por exemplo, na consulta q_1 da Tabela 5, os histogramas dos *datasets* "Limites Políticos" e "Estradas" foram utilizadas na primeira junção. Já na segunda etapa, o histograma gerado a partir dos histogramas da etapa anterior foi avaliado com o histograma do *dataset* "Rios", para produzir outro histograma intermediário para a próxima etapa, e assim por adiante.

Após a finalização dos experimentos, tanto a seletividade real quanto a estimada foram coletadas para as consultas listadas na Tabela 5. Em seguida, esses dados foram aplicados à Equação 4.1, e os resultados estão ilustrados em gráficos para proporcionar uma melhor compreensão do desempenho da estimativa. Além disso, os dados brutos da experimentação podem ser encontrados em forma de tabelas no Apêndice A.

Tabela 6 – Tamanhos dos histogramas gerados para cada *dataset*.

Nome	Tamanho
Linhas	
Estradas	258 x 113
Ferrovias	231 x 126
Hidrografia	126 x 130
Limites Políticos	51 x 32
Linhas de Transmissão	15 x 24
Pistas de Avião	164 x 178
Relevo	181 x 161
Rios	233 x 125
Rodovias	77 x 75
Trilhas	219 x 133
Polígonos	
Águas Interiores	213 x 137
Alertas de Desmatamento	133 x 219
Áreas Urbanizadas	176 x 166
Bacias	17 x 17
Bancos de Areia	238 x 123
Florestas	235 x 114
Mineração	218 x 134
Municípios	14 x 13
Regiões Rurais	5 x 5
Vegetação	56 x 49

Os gráficos da Figura 10 representam o erro associado ao método IHWAF, enquanto os gráficos da Figura 11 mostram o erro relacionado ao método MP. Vale ressaltar que como a 9ª etapa (s_9) da 1ª consulta (q_1) retornou nenhum objeto, ao calcular o erro ocorre uma divisão por zero, portanto nos gráficos a consulta q_1 não possui a 9ª etapa.

Analisando os gráficos de forma geral, (Figura 10 e 11), percebe-se que na primeira etapa das consultas a estimativa é mais precisa em comparação com as outras etapas. Isso se deve ao fato de que na primeira etapa são utilizados histogramas de dois *datasets*, ao invés de histogramas de resultados intermediários, como é feito nas etapas posteriores.

A estimativa de seletividade para as consultas utilizando o método IHWAF (Figura 10) apresenta um comportamento variável, com exceção da consulta q_2, que por sua vez tende a um comportamento logarítmico. Sendo ela uma consulta formada apenas por *datasets* do tipo polígono, esperava-se encontrar um comportamento semelhante até a 4ª etapa da q_4, já que a mesma é composta por um conjunto de 5 *datasets* do tipo polígono e

5 do tipo linha. No entanto, há uma melhoria na estimativa, ao contrário da 4ª etapa da $q2$, que apresentou uma degradação em relação à etapa anterior. Além disso, ao adicionar o primeiro *dataset* do tipo linha na 5ª etapa da q_4, nota-se um ganho de acurácia na seletividade estimada.

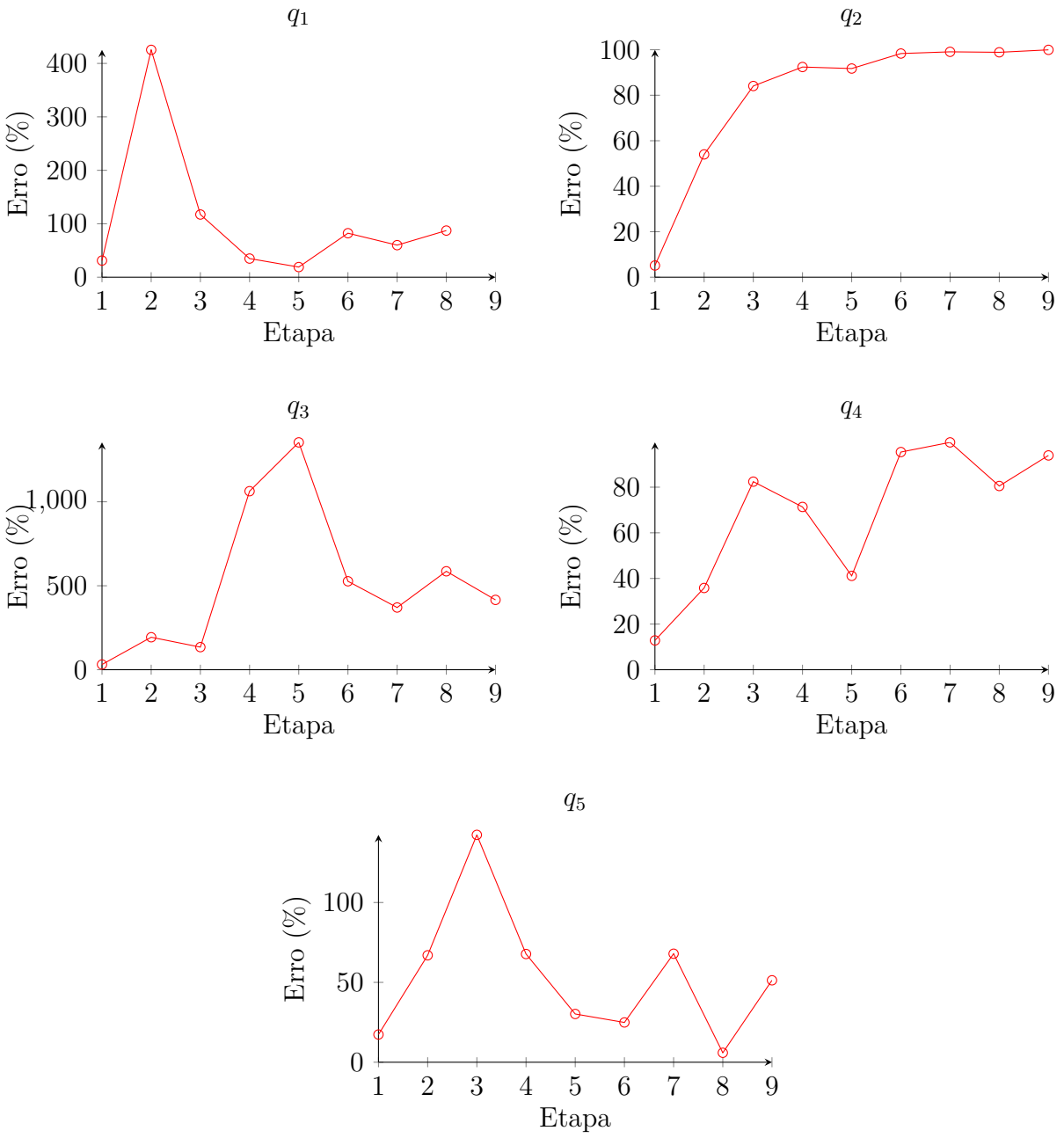


Figura 10 – Precisão da estimativa de seletividade utilizando o método IHWAF para cada uma das etapas das consultas.

Ao examinar os gráficos da Figura 11, que representam a precisão da estimativa realizada através do método MP, verifica-se que as consultas $\{q_1, q_2, q_3, q_5\}$ possuem um crescimento exponencial. Essa observação é fundamentada no fato de que os gráficos estão em uma escala logarítmica, onde uma reta sugere um crescimento exponencial dos dados.

Quanto a consulta q_4, seu comportamento é mais variável até a 7ª etapa, possuindo uma resistência de crescimento na 3ª e 5ª etapas. No entanto, a partir da 8ª etapa, ocorre um incremento no erro da estimativa.

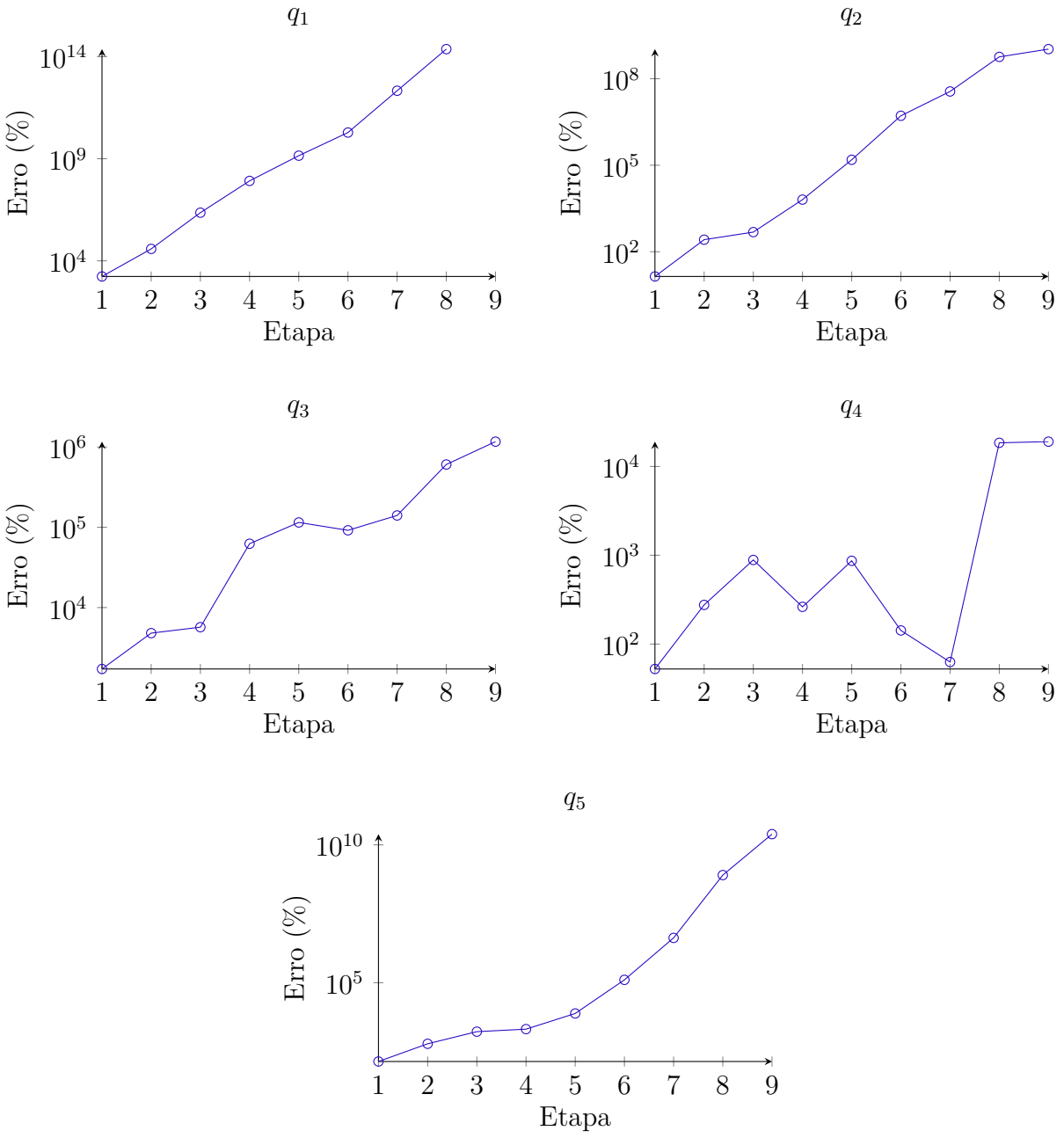


Figura 11 – Precisão da estimativa de seletividade utilizando o método MP para cada uma das etapas das consultas.

Por fim, foi calculada a média do erro para todas as etapas das consultas. A representação gráfica é apresentada na Figura 12. Vale ressaltar que o gráfico está em escala logarítmica devido os dados variarem em uma ordem de magnitude muito grande, o que dificulta a distinção de detalhes nos valores menores. A escala logarítmica comprime os valores maiores e expande os valores menores, tornando mais clara a compreensão de

tendências.

No gráfico da Figura 12, é perceptível que o método MP não consegue estimar com precisão a seletividade, já que apresenta um aumento significativo do erro em comparação com o IHWAF. Enquanto o método MP experimentou um crescimento exponencial do erro ao longo das etapas, o IHWAF manteve-se com uma boa assertividade, sem apresentar um aumento exagerado quando comparado ao MP.

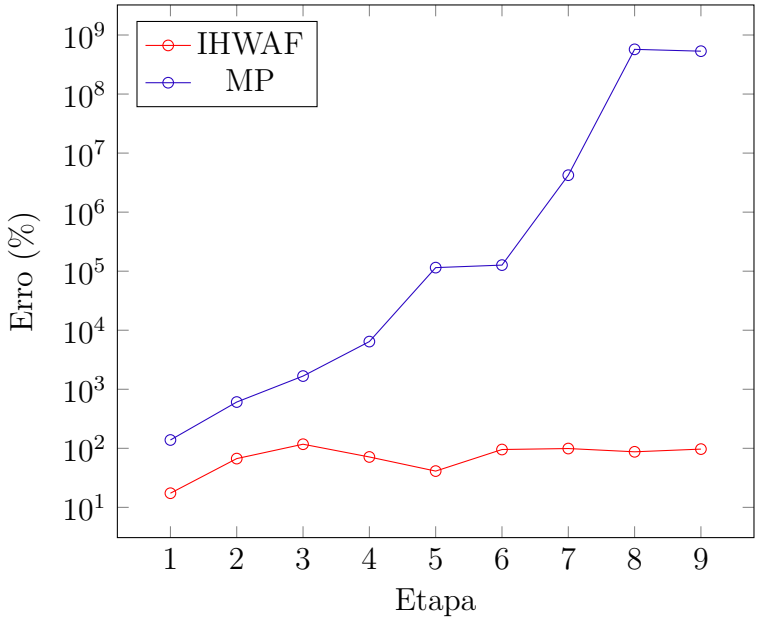


Figura 12 – Comparativo da precisão da estimativa de seletividade entre os métodos IHWAF e MP.

5 Conclusão e Trabalhos Futuros

Neste estudo, foi realizada uma avaliação da precisão da estimativa de seletividade ao longo das etapas que compõem a multijunção espacial. Para alcançar esse objetivo, conduzimos um experimento que consistiu na execução de cinco consultas, cada uma composta por dez *datasets* abrangendo objetos do mundo real. Foi calculado também as respectivas estimativas de seletividade com a utilização dos métodos IHWAF e MP. O primeiro tem como base a sobreposição proporcional, verificando o quanto um objeto está sobreposto em uma determinada célula, já o segundo por sua vez utiliza o centro do MBR para determinar se um objeto está contido em uma célula.

Os dados experimentais foram coletados e aplicados à métrica proposta, logo em seguida os resultados foram plotados em gráficos para facilitar a análise. Conclui-se então que seletividade estimada com o método IHWAF possui um comportamento variável, sem seguir uma tendência, tornando difícil identificar as causas dessa alternância na precisão. No entanto, uma das consultas (q_2) demostrou um padrão logarítmico único, notável por ser composta exclusivamente por conjuntos de dados do tipo polígono. Apesar disso, é impontante ressaltar que não há dados suficientes para afirmar que esse comportamento ocorre para todas as consultas envolvendo apenas *datasets* do tipo polígono.

Já a estimativa de seletividade empregando o método MP, revelou um comportamento semelhante para a maioria das consultas investigadas, manifestando uma degradação exponencial ao longo das etapas. Tal comportamento confirmou a hipótese central deste trabalho.

Por fim, como forma de permitir que outros pesquisadores reproduzam e validem os experimentos, disponibilizamos as consultas escritas na linguagem específica do DGEO, bem como os *datasets* utilizados em: <https://doi.org/10.6084/m9.figshare.25504963.v1>.

5.1 Trabalhos futuros

Os estudos conduzidos neste trabalho ofereceram uma visão inicial e inédita sobre o comportamento da estimativa de seletividade para consultas de multijunção espacial. Enquanto o método MP seguiu em direção da hipótese estabelecida, o IHWAF apresentou um padrão divergente, o que representa um avanço significativo na área de banco de dados espaciais.

Sugere-se, como trabalhos futuros, a elaboração de uma variedade maior de consultas, utilizando *datasets* reais e sintéticos, a fim de investigar as causas subjacentes à variação na precisão da seletividade estimada pelo método IHWAF. Essa investigação

permitiria entender e aprimorar o desempenho do mesmo.

Ao contrário dos tradicionais Sistemas de Banco de Dados Relacionais (SBDR), que já foram extensivamente estudados ao longo de décadas, os Sistemas de Banco de Dados Espaciais (SBDE) ainda enfrentam desafios notáveis e oferecem oportunidades para explorações futuras. Uma contribuição valiosa seria expandir a pesquisa para incluir diferentes tipos de multijunção além da cadeia, que foi o foco deste estudo. Isso possibilitaria uma compreensão mais abrangente sobre o comportamento das estimativas em cenários variados.

Além disso, caberia também comparar essas estimativas com as do PostGIS[1], que é uma extensão espacial de código aberto para o SGBD PostgresSQL[2], sendo um sistema consolidado e comumente utilizado pela comunidade.

Ao explorar os tópicos mencionados acima, podemos contribuir para aprimorar a eficiência e precisão dessas aplicações, beneficiando a sociedade como um todo, já que os dados espaciais podem ser utilizados em diversas áreas da atividade humana.

[1] <https://postgis.net/>
[2] <https://www.postgresql.org/>

Referências

ACHARYA, S.; POOSALA, V.; RAMASWAMY, S. Selectivity estimation in spatial databases. *SIGMOD Record*, v. 28, n. 2, p. 13–24, 1999. Citado 2 vezes nas páginas 22 e 26.

AJI, A.; WANG, F.; SALTZ, J. H. Towards building a high performance spatial query system for large scale medical imaging data. In: *Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. Redondo Beach, CA, USA: [s.n.], 2012. p. 309–318. Citado na página 15.

BOUROS, P.; MAMOULIS, N. Spatial joins: What's next? *SIGSPATIAL Special*, v. 11, n. 1, p. 13–21, 2019. Citado 2 vezes nas páginas 15 e 24.

BRINKHOFF, T.; KRIEGEL, H.-P.; SEEGER, B. Parallel processing of spatial joins using r-trees. In: *Proceedings of the Twelfth International Conference on Data Engineering*. New Orleans, LA, USA: [s.n.], 1996. p. 258–265. Citado 3 vezes nas páginas 15, 22 e 24.

CAMPBELL, J. E.; SHIN, M. *Geographic Information System Basics*. [S.l.]: The Saylor Foundation, 2012. Citado 5 vezes nas páginas 9, 20, 21, 22 e 23.

FITZ, P. R. *Geoprocessamento sem complicação*. [S.l.]: Oficina de textos, 2018. Citado 3 vezes nas páginas 19, 20 e 21.

FRANçA, A. G. *Precisão da Estimativa de Seletividade de Tarefas de Junção Espacial Distribuída usando Histogramas de Euler*. 63 p. Monografia — Universidade Federal de Goiás, Regional Jataí, Jataí, GO, Brasil, 2018. Citado 2 vezes nas páginas 9 e 17.

HUISMAN, O.; BY, R. A. de. *Principles of Geographic Information Systems*. [S.l.]: ITC, 2009. Citado na página 21.

JACOX, E. H.; SAMET, H. Spatial join techniques. *ACM Trans. Database Syst.*, Association for Computing Machinery, New York, NY, USA, v. 32, n. 1, p. 7–es, 2007. Citado 2 vezes nas páginas 9 e 24.

LONGLEY, P. A. et al. *Geographic Information Systems and Science*. 3. ed. [S.l.]: Wiley, 2010. Citado na página 21.

LU, H.; YIU, M. L.; XIE, X. Querying spatial data by dominators in neighborhood. *Information Systems*, v. 77, p. 71–85, 2018. Citado na página 19.

MAMOULIS, N.; PAPADIAS, D. *Advances in Spatial and Temporal Databases*. [S.l.]: Springer, 2001. v. 2121. 155–174 p. (Lecture Notes in Computer Science, v. 2121). Citado 7 vezes nas páginas 15, 16, 26, 27, 31, 33 e 37.

MAMOULIS, N.; PAPADIAS, D. Multiway spatial joins. *ACM Transactions on Database Systems (TODS)*, New York, NY, USA, v. 26, n. 4, p. 424–475, 2001. Citado 8 vezes nas páginas 15, 17, 22, 24, 26, 27, 32 e 33.

OLIVEIRA, T. B. de. *Efficient Processing of Multiway Spatial Join Queries in Distributed Systems*. 152 p. Tese (Doutorado) — Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brasil, 11 2017. Citado 12 vezes nas páginas 9, 15, 16, 24, 25, 26, 27, 28, 29, 32, 33 e 37.

OLIVEIRA, T. B. de et al. Scheduling distributed multiway spatial join queries: optimization models and algorithms. *International Journal of Geographical Information Science*, v. 37, n. 6, p. 1388–1419, 2023. Citado 3 vezes nas páginas 15, 25 e 29.

OLIVEIRA, T. B. de; COSTA, F. M.; RODRIGUES, V. J. S. Definição de planos de execução distribuídos para consultas de junção espacial usando histogramas multidimensionais. In: *Proceedings of the Brasilian Symposium on Databases*. Petrópolis, RJ, Brasil: [s.n.], 2015. p. 89–100. Citado 2 vezes nas páginas 15 e 16.

ÖZSU, M. T.; VALDURIEZ, P. *Principles of Distributed Database Systems*. 3. ed. [S.l.]: Springer, 2011. Citado na página 28.

SANTOS, M. C. dos; OLIVEIRA, T. B. de. Histograma intermediário de euler para estimativa de seletividade de multijunções espaciais. In: *Proceedings of XX Geoinfo*. São José dos Campos, SP, Brasil: [s.n.], 2019. p. 267–273. Citado na página 16.

STEEN, M. V.; TANENBAUM, A. S. *Distributed Systems*. 3. ed. [S.l.]: Maarten van Steen Leiden, The Netherlands, 2017. Citado na página 28.

SUN, C.; AGRAWAL, D.; ABBADI, A. E. Selectivity estimation for spatial joins with geometric selections. In: *Advances in Database Technology*. Berlin, Heidelberg: [s.n.], 2002. p. 609–626. Citado 2 vezes nas páginas 32 e 33.

Apêndices

APÊNDICE A – Dados Experimentais

Neste apêndice, foram registrados os dados brutos dos experimentos para referência futura, continuação da pesquisa ou aprofundamento no tema. As Tabelas 7, 8 e 9 exibem a seletividade real, a estimativa conforme o método IHWAF e a estimativa segundo o método MP, respectivamente. As Tabelas 10 e 11 detalham o erro associado a cada um dos métodos. Por fim, a Tabela 12 oferece uma comparação dos dados entre os métodos utilizados.

Tabela 7 – Seletividade real para cada uma das etapas das consultas.

$\mathbf{r_{q_i,s_j}}$	$\mathbf{q_1}$	$\mathbf{q_2}$	$\mathbf{q_3}$	$\mathbf{q_4}$	$\mathbf{q_5}$
$\mathbf{s_1}$	26.718	33.053	26.730	500.836	4.506
$\mathbf{s_2}$	241.370	24.880	26.044	661.508	1.947
$\mathbf{s_3}$	576.895	29.842	7.337	545.448	2.247
$\mathbf{s_4}$	11.759.688	125.080	30	979.560	4.796
$\mathbf{s_5}$	556.096	702.164	46	213.651	48.390
$\mathbf{s_6}$	37.421.420	202.350	30	1.241.102	88.928
$\mathbf{s_7}$	185.830.618	707.850	234	3.300.166	57.568
$\mathbf{s_8}$	32.657.014	1.214.100	300	1.774	312
$\mathbf{s_9}$	0	10.148.400	7.412	11.056	1.608

Tabela 8 – Estimativa de seletividade utilizando o método IHWAF para cada uma das etapas das consultas.

$\mathbf{e_{q_i,s_j}}$	$\mathbf{q_1}$	$\mathbf{q_2}$	$\mathbf{q_3}$	$\mathbf{q_4}$	$\mathbf{q_5}$
$\mathbf{s_1}$	35.034	34.751	35.034	565.139	3.724
$\mathbf{s_2}$	1.268.085	38.314	76.411	898.543	3.250
$\mathbf{s_3}$	1.253.422	4.756	17.211	995.055	5.445
$\mathbf{s_4}$	7.660.821	9.494	349	280.758	8.046
$\mathbf{s_5}$	450.616	58.008	669	125.772	63.007
$\mathbf{s_6}$	6.616.694	3.369	188	57.090	66.779
$\mathbf{s_7}$	74.377.400	6.443	1.101	11.984	18.456
$\mathbf{s_8}$	4.164.454	13.608	2.059	346	330
$\mathbf{s_9}$	36.496	3.786	38.233	672	2.433

Tabela 9 – Estimativa de seletividade utilizando o método MP para cada uma das etapas das consultas.

$\mathbf{e_{q_i,s_j}}$	$\mathbf{q_1}$	$\mathbf{q_2}$	$\mathbf{q_3}$	$\mathbf{q_4}$	$\mathbf{q_5}$
$\mathbf{s_1}$	$4{,}86 \times 10^5$	$3{,}76 \times 10^4$	$4{,}86 \times 10^5$	$7{,}63 \times 10^5$	$1{,}07 \times 10^4$
$\mathbf{s_2}$	$9{,}31 \times 10^7$	$9{,}01 \times 10^4$	$1{,}28 \times 10^6$	$2{,}49 \times 10^6$	$1{,}38 \times 10^4$
$\mathbf{s_3}$	$1{,}32 \times 10^{10}$	$1{,}73 \times 10^5$	$4{,}25 \times 10^5$	$5{,}38 \times 10^6$	$3{,}98 \times 10^4$
$\mathbf{s_4}$	$9{,}68 \times 10^{12}$	$8{,}16 \times 10^6$	$1{,}88 \times 10^4$	$3{,}55 \times 10^6$	$1{,}04 \times 10^5$
$\mathbf{s_5}$	$7{,}86 \times 10^{12}$	$1{,}09 \times 10^9$	$5{,}29 \times 10^4$	$2{,}06 \times 10^6$	$3{,}75 \times 10^6$
$\mathbf{s_6}$	$7{,}19 \times 10^{15}$	$1{,}05 \times 10^{10}$	$2{,}75 \times 10^4$	$3{,}01 \times 10^6$	$1{,}13 \times 10^8$
$\mathbf{s_7}$	$3{,}98 \times 10^{18}$	$2{,}56 \times 10^{11}$	$3{,}29 \times 10^5$	$1{,}23 \times 10^6$	$2{,}43 \times 10^9$
$\mathbf{s_8}$	$7{,}65 \times 10^{19}$	$6{,}94 \times 10^{12}$	$1{,}82 \times 10^6$	$3{,}29 \times 10^5$	$2{,}50 \times 10^9$
$\mathbf{s_9}$	$2{,}15 \times 10^{18}$	$1{,}08 \times 10^{14}$	$8{,}65 \times 10^7$	$2{,}12 \times 10^6$	$3{,}94 \times 10^{11}$

Tabela 10 – Precisão da estimativa de seletividade utilizando o método IHWAF para cada uma das etapas das consultas.

$\mathbf{\Delta_{q_i,s_j}}$	$\mathbf{q_1}$	$\mathbf{q_2}$	$\mathbf{q_3}$	$\mathbf{q_4}$	$\mathbf{q_5}$
$\mathbf{s_1}$	31,12	5,14	31,07	12,84	17,35
$\mathbf{s_2}$	425,37	53,99	193,39	35,83	66,95
$\mathbf{s_3}$	117,27	84,06	134,57	82,43	142,34
$\mathbf{s_4}$	34,86	92,41	1.063,17	71,34	67,76
$\mathbf{s_5}$	18,97	91,74	1.353,97	41,13	30,21
$\mathbf{s_6}$	82,32	98,34	526,75	95,40	24,91
$\mathbf{s_7}$	59,98	99,09	370,46	99,64	67,94
$\mathbf{s_8}$	87,25	98,88	586,17	80,52	5,92
$\mathbf{s_9}$	∞	99,96	415,83	93,92	51,30

Tabela 11 – Precisão da estimativa de seletividade utilizando o método MP para cada uma das etapas das consultas.

$\mathbf{\Delta_{q_i,s_j}}$	$\mathbf{q_1}$	$\mathbf{q_2}$	$\mathbf{q_3}$	$\mathbf{q_4}$	$\mathbf{q_5}$
$\mathbf{s_1}$	$1{,}72 \times 10^3$	$1{,}39 \times 10^1$	$1{,}72 \times 10^3$	$5{,}24 \times 10^1$	$1{,}38 \times 10^2$
$\mathbf{s_2}$	$3{,}85 \times 10^4$	$2{,}62 \times 10^2$	$4{,}81 \times 10^3$	$2{,}76 \times 10^2$	$6{,}08 \times 10^2$
$\mathbf{s_3}$	$2{,}29 \times 10^6$	$4{,}80 \times 10^2$	$5{,}70 \times 10^3$	$8{,}85 \times 10^2$	$1{,}67 \times 10^3$
$\mathbf{s_4}$	$8{,}23 \times 10^7$	$6{,}42 \times 10^3$	$6{,}24 \times 10^4$	$2{,}62 \times 10^2$	$2{,}06 \times 10^3$
$\mathbf{s_5}$	$1{,}41 \times 10^9$	$1{,}55 \times 10^5$	$1{,}15 \times 10^5$	$8{,}66 \times 10^2$	$7{,}64 \times 10^3$
$\mathbf{s_6}$	$1{,}92 \times 10^{10}$	$5{,}19 \times 10^6$	$9{,}17 \times 10^4$	$1{,}42 \times 10^2$	$1{,}27 \times 10^5$
$\mathbf{s_7}$	$2{,}14 \times 10^{12}$	$3{,}62 \times 10^7$	$1{,}40 \times 10^5$	$6{,}27 \times 10^1$	$4{,}22 \times 10^6$
$\mathbf{s_8}$	$2{,}34 \times 10^{14}$	$5{,}72 \times 10^8$	$6{,}06 \times 10^5$	$1{,}85 \times 10^4$	$8{,}01 \times 10^8$
$\mathbf{s_9}$	∞	$1{,}06 \times 10^9$	$1{,}17 \times 10^6$	$1{,}90 \times 10^4$	$2{,}45 \times 10^{10}$

Tabela 12 – Comparativo da precisão da estimativa de seletividade entre os métodos IHWAF e MP.

Etapa/Método	**IHWAF**	**MP**
$\mathbf{s_1}$	17,35	138,25
$\mathbf{s_2}$	66,95	607,54
$\mathbf{s_3}$	117,27	1.672,99
$\mathbf{s_4}$	71,34	6.420,89
$\mathbf{s_5}$	41,13	114.882,11
$\mathbf{s_6}$	95,40	126.729,90
$\mathbf{s_7}$	99,09	4.222.566,60
$\mathbf{s_8}$	87,25	571.760.703,14
$\mathbf{s_9}$	96,94	532.258.673,62