Murilo Cunha dos Santos

Estimativa de Custo de Multijunções Espaciais usando Histogramas Intermediários de Euler para *Datasets* de Linhas e Polígonos

Jataí-Goiás

2019

Murilo Cunha dos Santos

Estimativa de Custo de Multijunções Espaciais usando Histogramas Intermediários de Euler para *Datasets* de Linhas e Polígonos

Universidade Federal de Goiás - Regional Jataí - UFG-REJ

Instituto de Ciências Exatas e Tecnológicas (ICET)

Bacharelado em Ciências da Computação

Orientador: Dr. Thiago Borges de Oliveira

Jataí-Goiás

2019

Murilo Cunha dos Santos

Estimativa de Custo de Multijunções Espaciais usando Histogramas Intermediários de Euler para *Datasets* de Linhas e Polígonos/ Murilo Cunha dos Santos. – Jataí-Goiás, 2019-

61 p. : il. (algumas color.) ; 30 cm.

Orientador: Dr. Thiago Borges de Oliveira

Monografia (Graduação) – Universidade Federal de Goiás - Regional Jataí - UFG-REJ

Instituto de Ciências Exatas e Tecnológicas (ICET)

Bacharelado em Ciências da Computação, 2019.

1. Dados Espaciais 2. Multijunção Espacial 3. Histograma de Euler 4. Estimativa de Custo

Murilo Cunha dos Santos

Estimativa de Custo de Multijunções Espaciais usando Histogramas Intermediários de Euler para *Datasets* de Linhas e Polígonos

Trabalho aprovado. Jataí-Goiás, data da defesa: 12/12/2019.

Prof. Dr. Thiago Borges de Oliveira
Orientador

Profa. Mestra Franciny Medeiros Barreto
Avaliadora

Prof. Mestre Bruno Moraes Rocha
Avaliador

Jataí-Goiás
2019

Este trabalho é dedicado à minha família, a minha namorada e aos meus amigos que sempre me apoiaram.

AGRADECIMENTOS

Ao longo destes anos, foi um desafio cursar Ciências da Computação em outra localidade da minha família. Durante esta jornada diversas experiências serão levadas para minha vida toda, desde o conhecimento adquirido na universidade como também aos amigos que conheci durante minha graduação.

Por isso, gostaria de agradecer aos meus amigos que estiveram comigo durante esta caminhada, desde os momentos de diversão como os aprendizados. Gostaria de agradecer a minha namorada, Tainara Pascoaleto, que esteve comigo sempre, me apoiando, me incentivando.

Gostaria de agradecer minha família, que me deram suporte para esta jornada, mas queria agradecer com mais ênfase a minha mãe, Cristina Cunha, que desde o momento que eu decidi sair de casa para buscar oportunidades de vida até este instante, sempre esteve comigo, me apoiando, me dando o suporte necessário para que eu conseguisse alcançar tudo o que eu desejei.

Meus agradecimentos ao meu orientador, Thiago Borges, que compartilhou seus conhecimentos e seus aprendizados durante o decorrer do curso, a paciência que teve durante nossos projetos.

"As oportunidades multiplicam-se à medida que são agarradas."
(Sun Tzu)

RESUMO

As consultas de junção são essenciais para o processamento de dados espaciais. O processamento desse tipo de consulta é intensivo em recursos de computação, principalmente ao considerar multijunções espaciais que podem ser executadas de várias maneiras distintas, chamadas de planos de execução. Um plano mal escolhido aumenta o tempo de processamento e o uso de recursos computacionais e, portanto, precisamos de métodos eficazes para estimar o custo de consultas, como o histograma espacial. Estudos recentes identificaram que o tipo de objeto espacial em conjuntos de dados (seja do tipo linha ou polígono) desempenha um papel significativo na assertividade da estimativa baseada em histograma de grade. No entanto, fórmulas de estimativa em histogramas mais sofisticados e recentemente propostos, como o Histograma de Euler, não receberam esse novo tratamento. Este trabalho propõe um novo histograma intermediário de Euler para estimar a cardinalidade de consultas de multijunção espacial, e adapta as fórmulas de estimativa para empregar o tipo de objetos nas estimativas, além de considerar conjuntos de dados cuja extensão espacial não se alinha. Acreditamos que este é um cenário mais realista para aplicar os métodos a um banco de dados espacial. Nossa avaliação mostra que o modelo proposto estima assertivamente as cardinalidades das consultas de junção espacial e que a estimativa baseada nos tipos de objeto do conjunto de dados melhora significativamente a assertividade dos histogramas de Euler.

Palavras-chaves: *Dados Espaciais; Multijunção Espacial; Histograma de Euler; Estimativa de Custo.*

ABSTRACT

Spatial join queries are essential to spatial data processing and also very compute-resource intensive, particularly when considering multiway spatial joins, which have many distinct ways of computing called execution plans. A poorly chosen plan increases the processing time and usage of computational resources and, consequently, we demand very effective methods for estimating the cost of queries such as spatial histograms. Recently studies identified that the type of spatial object in datasets (whether of line or polygon type) plays a significant role in the assertiveness of grid histogram-based estimation. However, estimation formulae in more sophisticated and recently proposed histograms, such as the Euler Histogram, did not receive this particular treatment. This work proposes a novel Euler Intermediate Histogram to estimate the cardinality of multiway spatial join queries, adapt the formulae of estimation to employ the type of objects in estimates, and consider datasets whose spatial extension does not align. We believe this is a more realistic scenario towards applying the methods to a spatial database. Our evaluation shows that the proposed model assertively compute cardinalities of spatial join queries and that the estimate based on the dataset object types significantly improves the assertiveness for Euler Histograms.

Keywords: *Spatial Data; Multiway Spatial Join; Euler Histogram; Query Cost Estimation.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MBR	*Minimum Bounding Rectangle*
SGBDE	Sistemas Gerenciadores de Banco de Dados Espaciais
SIG	Sistema de Informações Geográficas
GIS	*Geographical Information System*
ERM	Erro Relativo Médio
PCB	*Printed circuit board*
SRC	Sistemas de Referência de Coordenadas
IHWAF	*Intermediate Histogram With Average Length Fix*
HIG	Histograma Intermediário de Grade
HIE	Histograma Intermediário de Euler
HIEA	Histograma Intermediário de Euler Avançado
I/O	*Input/Output*
CPU	*Central Processing Unit*
IBGE	Instituto Brasileiro de Geografia e Estatística
LAPIG	Laboratório LAPIG do Instituto de Estudos Sócio-Ambientais da Universidade Federal de Goiás

LISTA DE SÍMBOLOS

$\in$	Pertence
θ	Predicado da Junção espacial
$\bowtie$	Junção espacial
$\sum$	Somatório
$M(i, j)$	Matriz com i linhas e j colunas
$G(V, E)$	Grafo com vértices e arestas
V	Vértices
E	Arestas
F	Faces
c	Cardinalidade de uma célula de um histograma
λ	Erro Relativo Médio
σ^2	Desvio Padrão
$J_{l,l}$	Junção espacial com objetos do tipo linha
$J_{l,p}$	Junção espacial com objetos do tipo linha e polígono
$J_{p,p}$	Junção espacial com objetos do tipo polígono
ω	Seletividade de uma consulta
$O^{\bar{w}}$	Cardinalidade de consulta de janela $\bar{w}$
O^{j}	Cardinalidade da junção espacial
d	Total de dimensões
$\bar{a}$	Cardinalidade do *dataset*
l_{ak}	Comprimento médio dos objetos de um histograma a na dimensão k
l_{uk}	Comprimento do histograma a na dimensão k
$l_{a\bar{w}}$	Comprimento médio dos objetos de um histograma $\bar{w}$ na dimensão k

$\bar{w}$	Consulta de janela
m_j	Comprimento da célula na dimensão j
l	Vetor de comprimento médio
γ	Indica o fator que aumenta o comprimento médio dos polígonos em cada dimensão
ρ	Densidade de polígonos
b_a	Área do histograma b
l_{abk}	Comprimento da interseção entre os histogramas a e b na dimensão k
η	Coeficiente da interseção da linha
avg_x	Comprimento médio do objeto na dimensão x
avg_y	Comprimento médio do objeto na dimensão y
c_f	Cardinalidade das faces
c_a	Cardinalidade das arestas
c_v	Cardinalidade dos vértices

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

Uma técnica implementada no Sistema de Informações Geográficas (SIG) para a obtenção de novas informações sobre os objetos contidos nos *datasets* é a junção espacial (*Spatial Join*). A junção espacial é um tipo de consulta que encontra elementos correlacionados entre dois *datasets*, de acordo com um predicado espacial θ, como interseção ou proximidade (BRINKHOFF; KRIEGEL; SEEGER, 1996). Quando o processamento ocorre com mais de duas entradas ou em etapas, processando dois *datasets* de cada vez e produzindo resultados intermediários, é definido como multijunção espacial (MAMOULIS; PAPADIAS, 2001b).

O processamento das consultas espaciais podem ser realizados de várias formas, combinando dois ou mais *datasets* ao mesmo tempo, chamados de planos de execução. Independente do plano escolhido o resultado final da consulta será o mesmo, porém o tempo de execução de cada plano é distinto. Dependendo do plano escolhido, o tempo necessário para realizar a consulta acaba sendo demorado e o custo de processamento muito elevado. Com isso, para não haver um desperdício de tempo e de recursos computacionais é utilizado um otimizador de consultas para selecionar métodos de estimar a seletividade de consultas espaciais (MAMOULIS; PAPADIAS, 2001a).

Uma técnica usada dentro do otimizador de consultas para estimar o custo das operações é o histograma espacial. Histogramas são estruturas de dados com a função de simplificar os *datasets*, assim dividindo o espaço do *dataset* em uma grade que contenha diversas células (*buckets*). Estes *buckets* podem possuir tamanhos fixos ou variados, dependendo da estratégia adotada na estrutura de dados. Para cada célula ou *bucket*, são armazenadas metadados a respeito dos objetos espaciais contidos espacialmente, como a quantidade de objetos (cardinalidade) e o tamanho dos objetos (quantidade de pontos) (OLIVEIRA; COSTA; RODRIGUES, 2015).

Na construção dos histogramas, durante a contagem ou *hashing* dos objetos nos *buckets*, utiliza-se uma aproximação conhecida como *Minimum Bounding Rectangle* ou Mínimo Retângulo Envolvente (MBR) para o objeto. Um MBR possui a função de delimitar o objeto, envolvendo-o de forma retangular e é definido por dois, um ponto no canto extremo inferior esquerdo e outro no limite superior à direita. O MBR é muito utilizado pelo seu pequeno custo de armazenamento, porém causa erro de contagem dos objetos no histograma devido a contagem da área morta, ou seja, área que não está sendo ocupada de fato pelo objeto, que em geral não é retangular (LIU; YUAN; LIN, 2003).

No Histograma de Grade (MAMOULIS; PAPADIAS, 2001a), um conjunto de células é formado dividindo-se a extensão espacial do *datasets* e os objetos são contados em cada célula do histograma que sobrepõem. Objetos que ocupam ou sobrepõem mais de uma célula são contados múltiplas vezes e isso provoca erros na estimativa de seletividade das consultas

Um método foi proposto na literatura que utiliza uma estimativa de custo avançada para o Histograma de Grade, o *Intermediate Histogram With Average Length Fix* ou Histograma Intermediário com a Comprimento Médio Fixo (IHWAF). Assim este novo histograma contém a implementação de fórmulas propostas por Oliveira (2017) para os tipos de *datasets* e utiliza um método de sobreposição proporcional para o enquadramento dos objetos.

O Histograma de Euler (SUN; AGRAWAL; ABBADI, 2002b), ao contrário do Histograma de Grade, adota métodos em sua estrutura que procuram evitar a contagem múltipla dos objetos alocando *buckets* para identificar a face da célula, as suas laterais ou arestas, e para os cantos da célula, ou vértices. O objeto é contado na estrutura do histograma tanto na face quanto nas arestas e vértices que sobrepõe. Essas contagens adicionais proporcionam uma forma de evitar a contagem múltipla do objeto durante as estimativas das consultas, tornando a estimativa do custo computacional mais assertiva.

Recentemente, identificou-se que considerar o tipo dos objetos presentes nos *datasets* (se do tipo linha ou polígono) melhora o resultado da estimativa baseadas em Histogramas de Grade (OLIVEIRA, 2017). No entanto, este tipo de tratamento não foi adaptado para o Histograma de Euler. Como o Histograma de Euler trata do problema da contagem múltipla de forma diferente e foi mostrado como superior nos experimentos de (SUN; AGRAWAL; ABBADI, 2002b), justifica-se avaliar o seu comportamento incluindo as novas fórmulas de estimativa que consideram o tipo de objeto dos *datasets*.

1.2 Objetivo do Trabalho

Esta pesquisa teve como objetivo implementar o Histograma Intermediário de Euler para ser utilizado em consultas de multijunções espaciais e adaptar as várias fórmulas propostas em Oliveira (2017) definidas como estimativa de custo avançadas para o mesmo. Desta forma, será avaliado o histograma construído quanto a sua assertividade da estimativa de seletividade de consultas de multijunções. Assim os objetivos específicos se definem a seguir:

- Implementar o Histograma Intermediário de Euler.
- Inserir as fórmulas de estimativa de custo para o Histograma Intermediário de Euler.

- Avaliar a construção de cada estrutura do Histograma Intermediário de Euler, como o cálculo para as faces, arestas e vértices.
- Avaliar a assertividade do Histograma Intermediário de Euler com a estimativa de custo avançada.

1.3 Contribuição do Trabalho

As principais contribuições desde trabalho são descritas a seguir:

- Implementação e construção de um Histograma Intermediário baseado na teoria de Euler e a inclusão neste histograma da estimativa de custo avançada que considera o tipo dos objetos dos *datasets* quando estimar uma junção.
- Avaliação da construção das fórmulas do histograma porém analisando cada estrutura do Histograma de Euler, como o cálculo para as faces, arestas e vértices.
- Avaliação da assertividade da seletividade com o Histograma Intermediário de Euler Avançado e uma comparação com outros histogramas existentes.

1.4 Organização da Monografia

A organização do texto está dividida em seis capítulos e é dada a seguir: o Capítulo 2 apresenta os conceitos que compõem está pesquisa como descrição sobre dados espaciais, histogramas e estimativa de custo. O Capítulo 3 abrange alguns trabalhos cujos conteúdos apresentam maior relevância para este trabalho. O Capítulo 4 engloba a construção e implementação dos algoritmos para a estimativa de custo avançada e também apresenta a implementação do Histograma Intermediário de Euler. O Capítulo 5 apresenta a classificação da pesquisa, assim como a metodologia utilizada para os experimentos e os *datasets* utilizados. Apresenta também os resultados sobre a construção do Histograma Intermediário analisando sua estrutura e também a cardinalidade total de cada junção verificando sua assertividade do histograma. E por fim, o Capítulo 6 apresenta uma conclusão do trabalho e os trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo tem como objetivo detalhar os assuntos abordados na pesquisa, dando uma fundamentação do assunto e das técnicas existentes ao leitor. A seção 2.1 descreve as propriedades e técnicas para o processamento dos dados espaciais. Enquanto a seção 2.2 apresenta as definições sobre as consultas espaciais e seus tipos. Por fim, a seção 2.3 demonstra um otimizador de baixo custo.

2.1 Dados Espaciais

Dados espaciais são dados encontrados no espaço que são obtidos no nosso mundo. Podem ser dados geográficos reais como rios, vegetação, municípios, cidades e suas devidas coordenadas. Estes dados possuem várias características que podem ser descritas por nome, tipo, dentre outros (RIGAUX; SCHOLL; VOISARD, 2001).

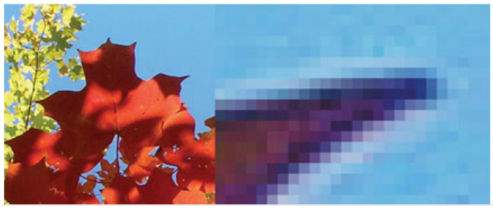
Como qualquer outro dado, os dados espaciais também possuem a necessidade de armazenamento, tratamento e recuperação dos dados em um banco de dados para poder acessar as informações, os SGBDE. Com os SGBDEs é possível realizar análises sobre o tipo de dado observado, como exemplo poder conseguir extrair informações de uma localidade que pode haver uma possível chance de aumento de queimadas. A forma como as informações são processados nos SGBDEs é através de sistemas distribuídos com objetivo de diminuir o tempo para processar as consultas (CAMPBELL, 2015).

Para armazenar os dados espaciais são através de imagens, mapas, dentre outras formas existentes. Para ser representado computacionalmente foram classificados como sendo diferenciados por sua estrutura como matriciais ou vetoriais (FITZ, 2018), e são apresentados pela Figura 1.

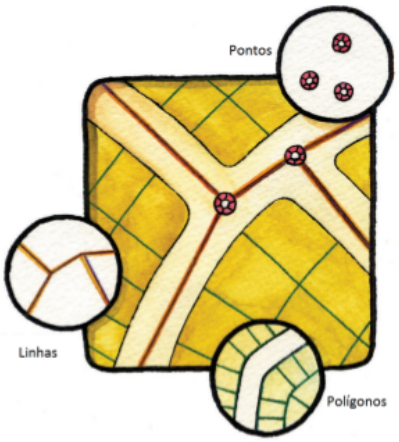
Pela Figura 1a pode-se observar como é o armazenamento da estrutura matricial (*raster structure*), representada por um conjunto de células, composta por uma matriz M com i linhas e j colunas. Cada célula $\mathrm{M}(i, j)$, para $i \in \{1, 2, ..., m\}$ e $j \in \{1, 2, ..., n\}$, é denominada por pixel, que pode conter um valor que representa, por exemplo, uma tonalidade de cinza ou qualquer outra cor (FITZ, 2018).

Para a demonstração do armazenamento da estrutura vetorial, observa-se a Figura 1b. A estrutura vetorial para traduzir a imagem captada dos sensores são composta por um conjunto de três tipos de objetos, sendo elas: pontos, linhas e polígonos. Este conjunto é construído a partir de pares de coordenadas (x, y), com a necessidade de apenas uma coordenada para a representação de um ponto. Para representar uma linha utiliza-se mais de uma coordenada, porém a primeira coordenada (x_1, y_1) necessitam ser diferente da

última (x_n, y_n), então $(x_1, y_1) \neq (x_n, y_n)$. E para representar um polígono também utiliza-se mais de uma coordenada mas a primeira coordenada (x_1, y_1) e a última coordenada (x_n, y_n) são as mesmas, obtemos então $(x_1, y_1) = (x_n, y_n)$ (CAMPBELL, 2015).



(a) Imagem em formato matricial



(b) Imagem em formato vetorial

Figura 1 – Modelos de representações espaciais. Fonte: (CAMPBELL, 2015).

Para Huisman e By (2009), dados espaciais não se restringem apenas aos dados geoespaciais, mas também com componentes eletrônicos em uma placa de circuito ou dados sobre o corpo humano capturado por imagens médicas. Para isso utiliza Sistemas de Referência das Coordenadas (SRC) que criam uma projeção específica para nivelar os dados em uma representação planar.

A partir dos dados espaciais coletados é necessário que haja um tratamento para que sejam processados, armazenados, corrigidos, dentre outros para a obtenção das informações destes dados. Para isso utiliza-se uma ferramenta denominada SIG (Sistema de Informações Geográficas ou do inglês GIS - *Geographical Information System*) (FITZ, 2018).

Para Rigaux, Scholl e Voisard (2001), um SIG é mais do que apenas uma ferramenta cartográfica para produzir mapas. Sua capacidade se expande para outras tarefas, como: armazenar, recuperar e combinar dados geográficos, a fim de criar representações de um espaço geográfico.

Antes de serem adicionados ao SIG, os dados precisam passar por um tratamento. O tratamento é utilizado para permitir uma melhor compreensão e obtenção dos dados, que são definidos em forma de *Layers* (camadas). Ou seja, existe um *dataset* conforme a Figura 2, que pode exemplificar o *dataset* completo como o conjunto de todos os dados que na imagem definimos como o mundo real. Porém, a imagem real é segmentada em várias partes, cada uma é um *dataset* ou *layer* com sua propriedade, como terreno, divisões, ruas, etc. Esses *datasets* menores são camadas específicas, que além de serem menores em função do custo de armazenamento são também mais rápidas na hora de combinar informações, evitando assim um custo maior de processamento (CAMPBELL, 2015).

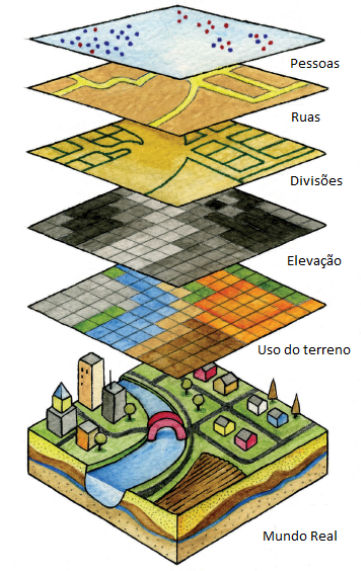


Figura 2 – Exemplo de um *dataset* e suas respectivas *layers*.

2.2 Consultas Espaciais

Um método de análise para se extrair as informações de um *dataset* do banco de dados são as consultas espaciais. Com as consultas podemos obter informações extras quando comparados os *datasets*, por exemplo um *dataset* de vegetação e um *dataset* sobre queimadas. Neste cenário, pode-se descobrir quais vegetações estão sendo mais prejudicadas e quais podem ser as futuras vegetações afetadas pelas queimadas. Na consulta citada anteriormente, observa que a consulta foi realizada com características em comuns nos *datasets*.

Para realizar uma consulta espacial necessita de um predicado espacial θ, que verifica alguma característica entre dois objetos de *datasets* distintos. Para exemplificar alguns predicados são listados a seguir (BRINKHOFF; KRIEGEL; SEEGER, 1996):

- Disjunção – Nenhum objeto entre os *datasets* possuem características em comum.
- Interseção – Ao contrário da disjunção, existem objetos entre os *datasets* que possuem relação de igualdade.

2.2.1 Tipos de Consultas Espaciais

Alguns tipos de consultas espaciais para encontrar as relações entres os *datasets* são o sistema de consulta por janela (*Windows Query*), a junção espacial e a multijunção espacial (MAMOULIS; PAPADIAS, 2001b).

Para a consulta de janela, pode-se observar pela Figura 3 que dentro de um *dataset* é criada uma área sombreada de formato retangular, chamada de janela (*window*). A janela

de consulta pode variar em seu tamanho e também na posição que ocupa no *dataset*. Com isso consegue percorrer todo o *dataset* e obter várias informações. O retorno da consulta é composto por todos os objetos que se intersectam do *dataset* com a janela de consulta.

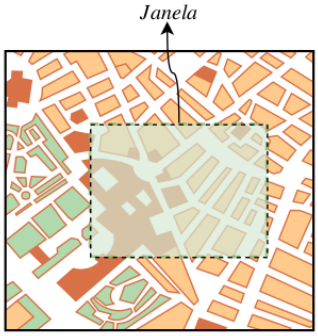


Figura 3 – Exemplo de uma consulta de janela. Fonte: (FRANÇA, 2018).

A operação de junção espacial ocorre com a combinação dos objetos de dois *datasets* que satisfazem um predicado espacial θ, podendo ser uma interseção. Para definir formalmente uma junção espacial, é definida pela definição 2.1 (BRINKHOFF; KRIEGEL; SEEGER, 1996).

Definição 2.1 (**θ-junção**). *Sejam $A = \{a_1, a_2, ..., a_n\}$ e $B = \{b_1, b_2, ..., b_n\}$ 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 e B. Formalmente uma junção espacial pode ser definida como uma função:*

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

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

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

Para ilustrar uma junção espacial é apresentada a Figura 4. Pela imagem encontra-se alguns objetos do *dataset* R que fazem interseção com os objetos do *dataset* S, já para R, temos os objetos r1, r2 e r3; para S temos os objetos s1, s2 e s3. O único objeto que não faz interseção nenhuma é o s1 (JACOX; SAMET, 2007).

2.2.2 Multijunção espacial

Diferente da junção espacial, que ocorre com apenas duas entradas, outro tipo de consulta que aceita um número de entrada maior que dois *datasets* é a Multijunção Espacial (MAMOULIS; PAPADIAS, 2001b). Exemplo dessas consultas: “Encontrar todas as espécies de animais que moram em áreas de preservação nas quais há perigo de fogo do

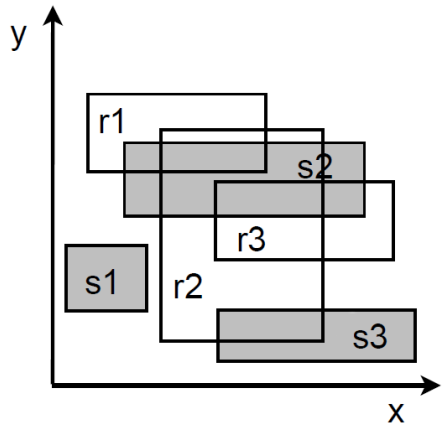


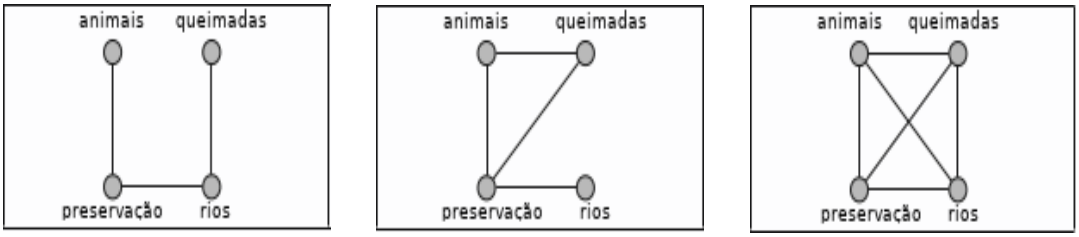
Figura 4 – *Datasets* R e S em um plano cartesiano. Fonte: (PITA, 2016).

lado do rio." Nesse caso, quatro *datasets* espaciais serão utilizados, que são: animais, áreas de preservação, propagação pelo fogo e rios, que precisam ser combinados para resultar na consulta (OLIVEIRA, 2017).

A multijunção espacial pode ser representada como um grafo $G = (V, E)$, sendo G o grafo, V vértice e E aresta. Cada vértice V representa um *dataset* distinto e as arestas entre cada vértice são o predicado da junção entre eles. A Figura 5 ilustra um grafo G de uma multijunção espacial. Cada item na figura (a,b,c) representa um tipo diferente de consulta, classificando de acordo com as características do grafo. A Figura 5a mostra uma consulta em árvore ou consulta em cadeia (*tree ou chain*), um tipo comum de consulta na relação do processamento dos dados espaciais que possui todos os *datasets* combinados em pares, sem repetição. A Figura 5b representa um grafo que contém um ciclo (*cycle*), portanto, uma consulta de ciclo. Finalmente, a Figura 5c mostra um caso que todos *datasets* precisam ser verificados entre si, para verificar a existência do predicado da multijunção classificado como Clique (MAMOULIS; PAPADIAS, 2001b).

Cada consulta da Figura 5 pode ser dividida em passos e processadas em ordens diferentes. Houve uma investigação sobre o número de maneiras de consultas para serem processadas com processamento serial (não-paralelo, não distribuído) (PAPADIAS; MAMOULIS; THEODORIDIS, 1999). Pela investigação foi demonstrado que em uma função do tipo de consulta, o número de entradas de *datasets* e o número de diferentes algoritmos de junções podem ser usados em cada passo. Por exemplo, existem aproximadamente cem combinações para cinco entradas de *datasets* com o grafo de consulta Clique, considerando três algoritmos de junção (MAMOULIS; PAPADIAS, 2001b).

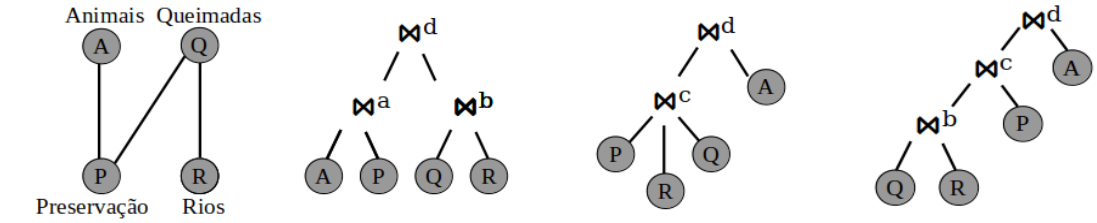
Uma consulta de multijunção pode ser processada de várias formas diferentes, chamadas planos de execução. A Figura 6 ilustra uma consulta de grafo e três formas diferentes para a consulta de cadeia (*chain*). A Figura 6a ilustra o grafo, a Figura 6b, dois pares de *datasets* são produzidos em um passo, produzindo dois resultados intermediários



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

Figura 5 – Tipos de consultas de multijunção.

($\bowtie^a, \bowtie^b$), depois os resultados intermediário geram o resultado intermediário ($\bowtie^d$). Na Figura 6c três *datasets* foram combinados no primeiro passo, gerando um resultado intermediário ($\bowtie^c$), para que no segundo passo fosse combinado com o outro *dataset* e assim gerasse o resultado intermediário ($\bowtie^d$). Na Figura 6d, vemos que a cada passo, dois *datasets* são combinados, gerando um novo intermediário ($\bowtie^b, \bowtie^c$). E cada qual intermediário combinado com outro *dataset*, sempre combinando em pares de forma recursiva que no final geram o resultado intermediário ($\bowtie^d$) (OLIVEIRA; COSTA; RODRIGUES, 2015).



(a) Consulta do grafo (b) Plano 1 (c) Plano 2 (d) Plano 3

Figura 6 – Planos de execução de multijunção.

Todos os diferentes planos de execução para as consultas preservam a semântica dos *datasets*, porém cada processo é diferente, consequentemente o tempo de custo para a execução será variado no resultado final. Para otimizar a consulta ou plano de seleção do algoritmo, geralmente utiliza como entrada o grafo que representa a consulta, e depois de enumerar os possíveis planos, seleciona um para verificar sua performance, através de um otimizador. O otimizador considera alguns aspectos dos *datasets* e associa o custo dos algoritmos para selecionar qual o plano de execução e determina: i) quais *datasets* serão combinados, ii) a ordem do processo para os *datasets*, e iii) qual algoritmo que será usado em cada passo (OLIVEIRA, 2017).

2.2.3 Estimando o Custo para Seleção ou Escolha de Planos

Para a execução dos planos de dados espaciais e medir a complexidade da estimação, foram propostos algumas expressões que combinavam métodos de estimação de custo das consultas para multijunção espacial (FORNARI; COMBA; IOCHPE, 2006; AN; YANG;

SIVASUBRAMANIAM, 2001). Estes métodos e um conjuntos de equações são chamados ou conhecidos como modelo de custo (OLIVEIRA, 2017).

Algumas dessas equações predizem um custo para a execução de planos por calcular o custo de I/O (*Input/Output* – Entrada/Saída) e da CPU (*Central Processing Unit*) por algoritmos de junção, assumindo que os objetos espaciais preenchessem a extensão espacial uniformemente. Estudos realizados por Mamoulis e Papadias (2001b) encontraram uma dificuldade de usar *datasets* reais, quando utilizados *datasets* espaciais reais, a equação pode induzir a escolha de planos de execuções ruins (MAMOULIS; PAPADIAS, 2001b).

Além deste estudo, houve também outro trabalho sobre a aplicabilidade das equações em regiões pequenas dos *datasets*, propondo o uso de histogramas uniformes que dividem a extensão espacial do *dataset* em um mapa com tamanho fixo de células (MAMOULIS; PAPADIAS, 2001a). A vantagem principal dos histogramas uniformes são i) construção simplificada, ii) tempo eficiente para estimar as consultas e iii) manutenção incremental para *datasets* não estáticos (CORMODE et al., 2011).

Para a construção de um histograma multidimensional o maior desafio se encontra na contagem dos objetos espaciais nas células do histograma. Para a contagem dos objetos é baseado no número de objetos presentes dentro dos limites da célula, porém os objetos usam seu retângulo mínimo envolvente (MBR). Contudo, o MBR utiliza células com tamanhos menores, aumentando assim a quantidade de células, o número do erro aumenta devido ter um aumento da contagem múltipla de objetos em mais de uma célula (MAMOULIS; PAPADIAS, 2001b).

Uma técnica para melhorar a estimativa de custo é a utilização do comprimento médio dos objetos como metadados adicionais para cada célula do histograma. Este comprimento médio utilizado toma como base o MBR da célula de todos os objetos nos histogramas. Pela Equação 2.3 é demonstrado o uso do comprimento médio dos objetos gerando a estimativa de cardinalidade de uma consulta $O^{\bar{w}}$ e a consulta da junção O^{j} ambas descritas detalhadamente abaixo (MAMOULIS; PAPADIAS, 2001a).

Para a saída da cardinalidade de $O^{\bar{w}}$ da consulta de janela $\bar{w}$, demonstrado pela Equação 2.3 é definido $\bar{a}$ como a cardinalidade do *dataset* a, d é o número total de dimensões e k percorre todas as dimensões. Para representar comprimento médio dos objetos de a na dimensão k temos então l_{ak}, para o comprimento de $\bar{w}$ é representado por $l_{a\bar{w}}$ e l_{uk} para representar o comprimento de a dimensão k, sendo $l_{uk} \neq 0$. O resultado da equação é a quantidade de objetos estimados do predicado espacial adotado a interseção, considerando a consulta de $\bar{w}$ com os objetos de a.

$$O^{\bar{w}}\left(a, \bar{w}\right) = \bar{a} * \prod_{k=1}^{d} min\left(1, \frac{l_{ak} + l_{a\bar{w}}}{l_{uk}}\right) \tag{2.3}$$

As junções espacias utilizam um par de *datasets*, sendo estes $\{a, b\}$. Cada *dataset* contém seu próprio histograma e seus respectivos resultados de cardinalidade de suas junções. Demonstrado pela Equação 2.4, obtemos o resultado da estimativa de seletividade, buscando no histograma ou nas células individualmente em $O^j\,(a, b)$, as cardinalidades das consultas dos *datasets* a e b são demonstrados por $O^{\bar{w}}\,(a, i)$ e $O^{\bar{w}}\,(b, i)$ respectivamente. A interseção dos MBRs dos objetos de a e b é representada por i, l_{ik} é o comprimento de i na dimensão k e l_{ak} e l_{bk} são os comprimentos médio dos objetos de a e b, respectivamente, e todos na dimensão k.

$$O^j\,(a, b) = O^{\bar{w}}\,(a, i) * O^{\bar{w}}\,(b, i) * \prod_{k=1}^{d} min\left(1, \frac{l_{ak} + l_{bk}}{l_{ik}}\right) \tag{2.4}$$

É demostrado a aplicação da Equação 2.4 para cada par de célula obedecendo um predicado no plano de execução para melhorar o resultado estimado da cardinalidade das consultas de junção. Além disso, percorrendo os pares das células com intersecção dos histogramas gerados H_a e H_b, e aplicando a Equação 2.4 é possível então construir um histograma intermediário com o resultado de cada par. Seguindo em frente, utilizou esse histograma intermediário com os planos de execução sendo uma entrada de uma etapa do plano de execução (MAMOULIS; PAPADIAS, 2001a).

2.2.4 Seletividade das Consultas Espaciais

Estimado que os tempos de consultas não são uniformes, podendo algumas vezes até ser custoso demais computacionalmente, um otimizador é selecionado para que escolha o melhor plano. Um otimizador apropriado é a seletividade de consultas espacias que tentam trazer informações relevantes como a quantidade aproximada de objetos de um *dataset* (cardinalidade), assim economizando tempo de processamento (IOANNIDIS; POOSALA, 1999).

A seletividade pode ser calculada por um fragmento do *dataset* como, por exemplo, o cálculo em apenas uma partição. Portanto, a seletividade é utilizada como parâmetro para a distribuição de tarefas de um *cluster*, posto que partições com seletividade alta seriam mais dispendiosas em termos de tempo de processamento (OLIVEIRA, 2017).

2.3 Histogramas Espaciais

Uma estrutura para dados espaciais que apoia o cálculo da estimativa de seletividade das consultas espaciais de multijunção é o histograma espacial. Um histograma espacial simplifica o *dataset* real dividindo a extensão espacial do *dataset* em uma quantidade de células ou *buckets*. Cada uma das células armazenam metadados sobre os objetos dos

datasets, como a quantidade (cardinalidade) e o tamanho dos objetos (OLIVEIRA; COSTA; RODRIGUES, 2015).

Um exemplo de histograma é o Histograma de Grade. O Histograma de Grade, exemplificado pela Figura 7, mostra o *dataset* de municípios do Brasil dividido em células de mesmo tamanho. Cada cada célula armazenará o valor da cardinalidade da quantidade de objetos presentes (AN; YANG; SIVASUBRAMANIAM, 2001).

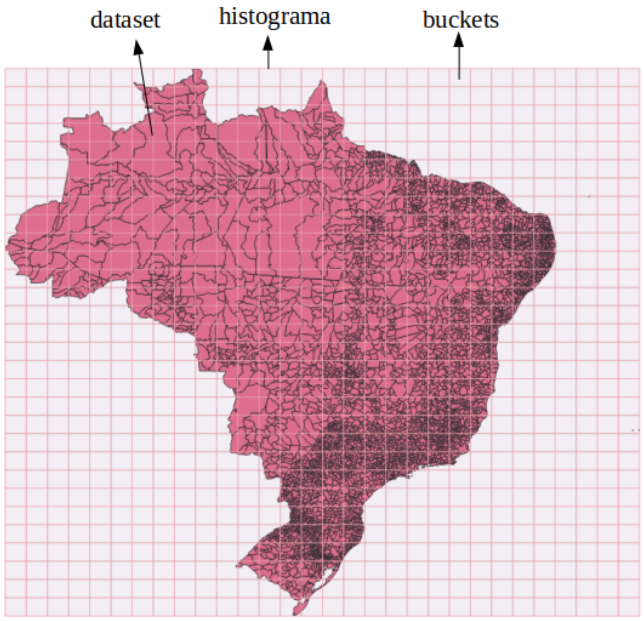


Figura 7 – Exemplo do Histograma de Grade para o *dataset* de Municípios do Brasil.

Um dos erros de aproximação existentes nos histogramas espaciais, e que impactam no custo das consultas, é o problema da contagem múltipla. A contagem múltipla ocorre quando um objeto espacial se sobrepõe, devido sua extensão espacial ou devido a área morta de seu MBR, em mais de uma célula do histograma. Um exemplo de demostração da área morta pode ser ilustrado pela Figura 8 . Este problema foi estudado por Sun, Agrawal e Abbadi (2002b) e por Oliveira (2017).

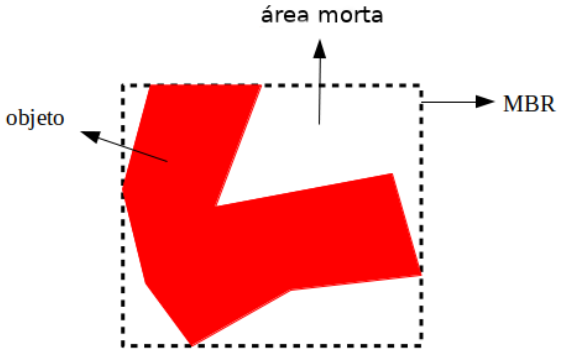


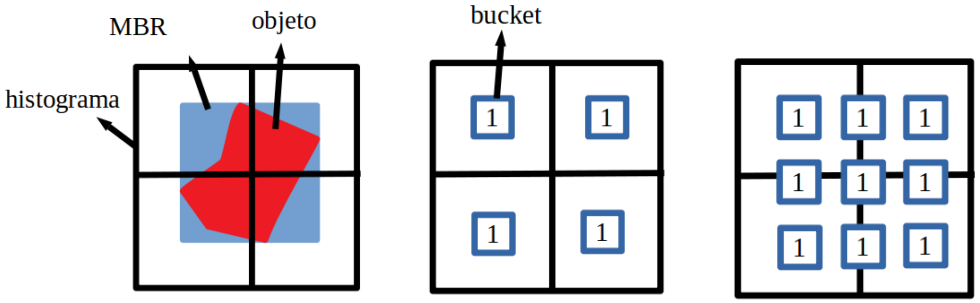
Figura 8 – Exemplo MBR com objeto e área morta.

Oliveira (2017) propôs novas técnicas para a construção do Histograma de Grade. Uma delas, para lidar com o problema da contagem múltipla, chamada de Sobreposição Proporcional, conta o objeto das células de forma proporcional.

Uma abordagem distinta, fundamentada na teoria dos grafos, mas também para o problema da contagem múltipla, é o Histograma de Euler, apresentado a seguir (SUN; AGRAWAL; ABBADI, 2002a).

2.3.1 Histograma de Euler

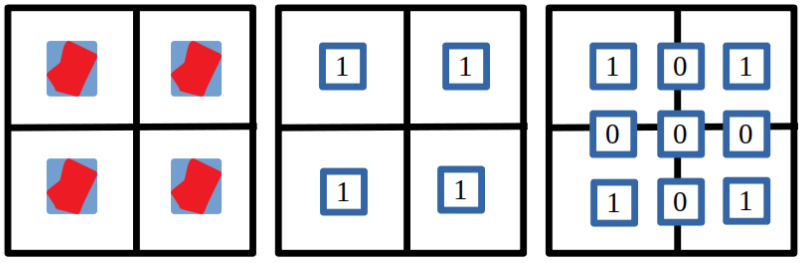
Um histograma elaborado busca não conter o problema de contagem múltipla nos Histogramas de Grade, é denominado de Histograma de Euler. Sua estrutura se diferencia do Histograma de Grade pois o Histograma de Euler aloca os *buckets* não apenas para as faces das células como também para os cantos e os contornos das células. Uma comparação entre o Histograma de Grade com o Histograma de Euler é ilustrado pela Figura 9 (SUN; AGRAWAL; ABBADI, 2002b).



(a) Objeto contido em um histograma

(b) Alocação de *buckets* no Histograma de Grade

(c) Alocação de *buckets* no Histograma de Euler



(d) Outro exemplo com quatro objetos em ambos os histogramas

Figura 9 – Comparação entre o Histograma de Grade e Histograma de Euler.

Conforme mostra a Figura 9a podemos ver um objeto e seu MBR presente no centro de um histograma com quatro células. Para atribuirmos a posição do objeto em qual célula está presente, adicionamos o valor de um. Figura 9b observa-se o Histograma Grade do *dataset*. Figura 9c temos o objeto em um Histograma de Euler. Nota-se a diferença na quantidade de *buckets* que o Histograma de Euler possui a mais que o de Grade. Na Figura 9d podemos visualizar outro exemplo, porém neste exemplo é apresentado quatro

objetos em um Histograma e ao lado dela apresenta a contagem do objeto nos Histogramas de Grade e de Euler.

E utilizando um exemplo do problema de contagem múltipla do Histograma de Grade pela Figura 9a, o histograma não consegue distinguir o tamanho do objeto sendo este grande ou pequeno, estando o objeto presente em todas faces. No caso do Histograma de Euler é possível notar que o objeto só está contido no meio das faces, mas não está contido em nenhuma aresta, conseguindo distinguir o tamanho do objeto e não contando múltiplas vezes o objeto.

2.3.2 Teoria dos Grafos

Teoria dos grafos é um exploração de técnicas de prova em matemática discreta. Um grafo possui uma estrutura básica que são $G = (V, E)$, sendo G o próprio grafo, V um conjunto de vértices e E as arestas presentes no grafo. E a teoria dos grafos pode ser aplicado em diversos tipos de problemas (WEST et al., 1996).

Por exemplo, um problema que envolve teoria dos grafos é quando busca-se encontrar o menor trajeto em uma viagem. Assim, ilustrado pela Figura 10 um conjunto de sete cidades e seus respectivos trajetos para outras cidades. Então, para representar um grafo temos os vértices como sendo as cidades, as arestas sendo os caminhos entre as cidades. Para este problema, tem-se que a cidade de origem é a "cidade a"e o destino é a "cidade e". Entretanto, existe dois caminhos a percorrer até chegar no destino, consequentemente a escolha até o destino será escolhida com base nos valores que as arestas podem assumir, neste caso como quilômetros entre as cidades, e assim escolher o menor custo de viagem.

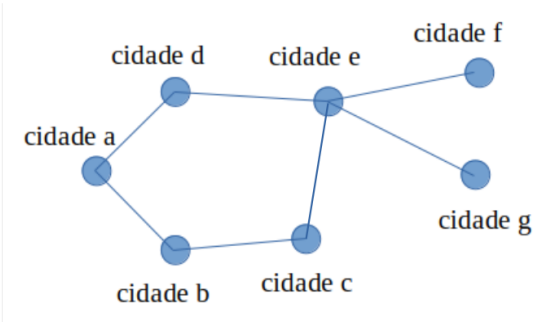


Figura 10 – Exemplo de grafo de trajeto entre cidades.

Um grafo pode ser completo onde as arestas cujos pontos de extremidade são iguais, assim todas as arestas se ligam com todos os vértices. Um grafo simples é um grafo que não contém mais do que um vértice. A Figura 11 ilustra um grafo completo e simples (WEST et al., 1996).

Um caminho de um grafo é uma dupla alternância de vértices, em que cada aresta é incidente com os dois vértices. Quando o vértice inicial é igual o final é dito como caminho

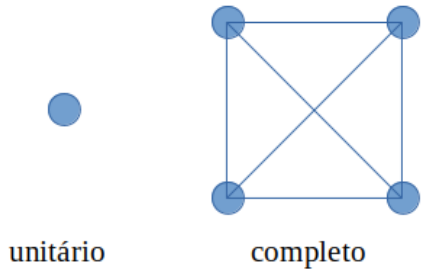


Figura 11 – Exemplo de grafo unitário e grafo completo.

fechado. É um circuito se todas as arestas forem distintas, fechado e desde que seus n pontos sejam distintos e $n > 3$. É ilustrado pela Figura 12 um circuito mínimo com três vértices e um circuito com quatro vértices cujo interior é definido como face (HARARY, 1969).

Figura 12 – Exemplo de grafo com circuito mínimo e com face.

Proposto por Harary (1969) a fórmula de Euler pode ser analisada pela Equação 2.5 abaixo e ilustrada sua utilização pela Figura 13a:

Teorema 2.1. *Seja G um grafo qualquer com V vértices, E arestas e F faces,*

$$V - E + F = 2 \tag{2.5}$$

Beigel e Tanin (1998) provaram o corolário da fórmula de Euler e utilizaram-se em seu artigo apenas a versão bi-dimensional do corolário, abaixo encontra o corolário proposto e ilustrado pela Figura 13b:

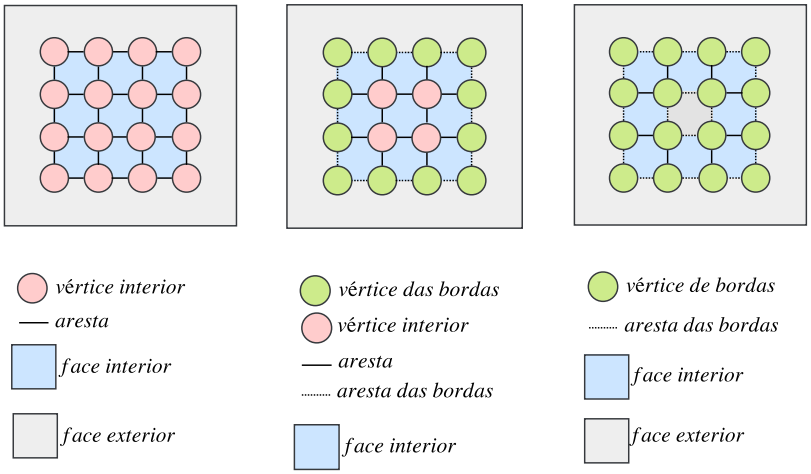
Corolário 2.1. *Seja G um grafo qualquer, sendo V_i, E_i e F_i o número de vértices, arestas e faces interiores de G respectivamente.*

$$V_i - E_i + F_i = 1 \tag{2.6}$$

Em Sun, Agrawal e Abbadi (2002a) estendeu-se a fórmula proposta de grafos proposta no Equação 2.5 sendo esta descrita no teorema a seguir e ilustrado pela Figura 13c:

Corolário 2.2. *Seja G um grafo conexo e planar com k faces externas e não havendo duas faces externas compartilhando o mesmo limite (faces externas vizinhas). Considerando V_i, E_i e F_i sendo o número de vértices, arestas e faces interiores, então*

$$V_i - E_i + F_i = 2 - k \tag{2.7}$$



(a) $V_i - E_i + F_i = 16 - 24 + 10 = 2$

(b) $V_i - E_i + F_i = 4 - 12 + 9 = 1$

(c) $V_i - E_i + F_i = 0 - 8 + 8 = 0$

Figura 13 – Exemplos da Equação 2.5, da Equação 2.6 e da Equação 2.7. Fonte: (FRANÇA, 2018)

É possível observar pela Figura 13a uma grade de 3×3 onde o teorema proposto é exemplificado contendo as faces exteriores, as faces internas, os vértices internos e arestas, contendo $1, 10, 16$ e 24 respectivamente. A Equação 2.6 utiliza a remoção da face exterior, possuindo agora um total de 4 vértices interiores, 12 arestas interiores e 9 faces interiores, consequentemente as arestas interiores e vértices interiores se tornaram exteriores, ilustrados pela Figura 13b. Para a proposta da Equação 2.7 é utilizado para mais de uma face exterior, removendo uma face central e tornando-a exterior, ilustrado pela Figura 13c. Desta maneira a grade 3×3 que antes possuía apenas uma face externa passa a conter duas e consequentemente passa a conter 0 vértices internos, 8 arestas internas e 8 faces internas.

2.3.3 Construção do Histograma

Os histogramas são divididos em um conjunto de grades, cujo estas contém *buckets* que correspondem a uma célula da grade (AN; YANG; SIVASUBRAMANIAM, 2001). Então, quando o histograma é utilizado nos *datasets* passa a incrementar o valor dos *buckets* em 1 para cada objeto em que se intersecta com a célula. Entretanto, este tipo de histograma contém um problema que não consegue distinguir o tamanho dos objetos. Sun,

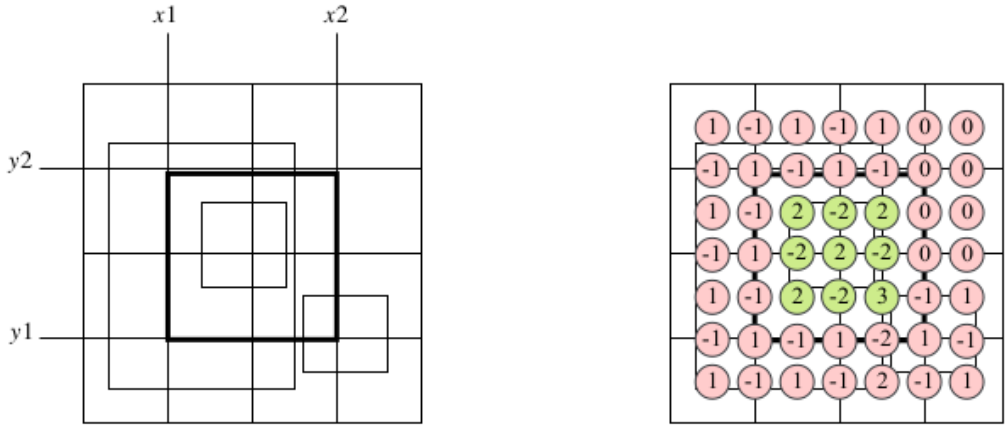
Agrawal e Abbadi (2002a) para melhorar o modelo já existente propuseram um novo tipo de histograma que pode ser construído dado as condições abaixo da seguinte forma:

- Dado uma grade $n_1 \times n_2$ no R^2, alocar $(2n_1 - 1)(2n_2 - 1)$ *buckets* para o histograma H. Um *bucket* de H corresponde a um vértice, aresta ou face da grade.
- Varrer o *dataset*. Para cada objeto, se um vértice, aresta ou face da grade intersecta seu interior, incrementar o *bucket* correspondente em 1.
- Uma vez que o *dataset* inteiro seja processado, inverter o sinal dos valores nos *buckets* que correspondem a arestas.

Dadas as condições para a criação do histograma pode ser computado a seletividade de uma consulta como no exemplo da consulta de janela. Dessa forma temos uma consulta de janela S e a seletividade (ω) de S pode ser calculada com o Histograma de Euler, seguindo a Equação 2.8 (SUN; AGRAWAL; ABBADI, 2002b):

$$\omega(S) = \sum_{0 \leq k \leq d} (-1)^k F_k(S) \tag{2.8}$$

e $F_k(S)$ é k-dimensão dentro de S. As dimensões seguem um esquema cujo a 0-dimensão como um vértice, a 1-dimensão como uma aresta e 2-dimensão como a célula.



(a) *Dataset* com janela de seleção (b) Histograma de Euler

Figura 14 – Histograma de Euler na consulta de janela. Fonte: (FRANÇA, 2018)

Um exemplo do cálculo descrito anteriormente é demostrado pela Figura 14. A Figura 14a demostra três objetos retangulares e uma janela de seleção em (x_1, x_2, y_1, y_2). A Figura 14b mostra o Histograma de Euler para os *datasets*. Utilizando a Equação 2.8, a

seletividade de S é dado como:

$$\overbrace{(-1)^0 \times 2}^{\text{vértices}} + \underbrace{(-1)^1 \times (2+2+2+2)}_{\text{arestas}} + \overbrace{(-1)^2 \times (2+2+2+3)}^{\text{faces}} = 3$$

Duas importantes propriedades para a Equação 2.8 foram estabelecidas (BEIGEL; TANIN, 1998):

- Se a janela de seleção alinha com a grade, a estimativa da equação ocorre sem erros.
- Equação 2.8 serve para qualquer d-dimensão no espaço, no qual $d \geq 1$.

2.3.4 Técnicas Avançadas de Construção do Histograma

Dentre as melhorias propostas por Oliveira (2017), em relação a estimativa de seletividade usando Histogramas de Grade, destacamos a seguir duas delas que pretende-se incorporar no Histograma de Euler neste projeto.

A primeira melhoria diz respeito à prevenção de interseções entres objetos espaciais, com o uso de estimativas avançadas, seguindo a Equação 2.9. Estas melhorias endereçam a imprecisão que os MBR induzem no campo de comprimento médio sobre objetos espaciais complexos (linhas e polígonos).

Um exemplo do erro do MBR é ilustrado pela Figura 15. Então têm-se o objeto como *o*, e o seu MBR. O comprimento médio de *o* de ambas as dimensões (x e y) é (avg_x e avg_y), respectivamente. Pode-se observar que os comprimentos médios é o mesmo de um quadrado ocupando a célula inteira. Entretanto, a probabilidade de interseção do objeto é de 0.5 e não de 1. Este problema acontece também com objetos do tipo linha.

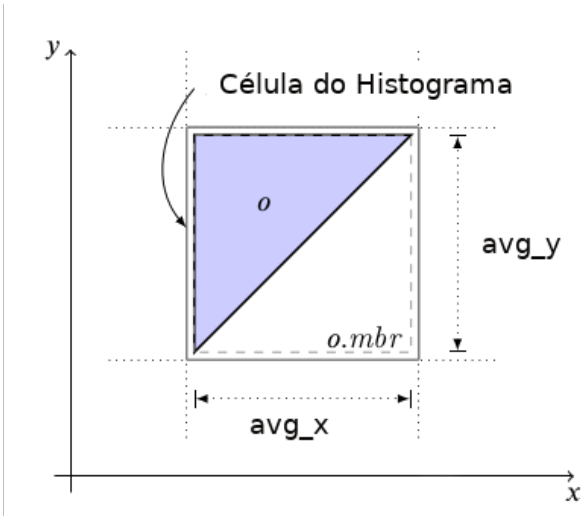


Figura 15 – Ilustração do erro introduzido pelo MBR. Fonte: (OLIVEIRA, 2017)

De fato, objetos espaciais do mesmo *dataset* frequentemente representam o mesmo fenômeno natural e não se sobrepõem. Deste modo, a soma do comprimento de todos

os objetos é menor ou igual ao comprimento da célula. No entanto, a aproximação da MBR pode levar a um comprimento médio que distorce esse comportamento, ou seja, o comprimento médio multiplicado pela cardinalidade da célula é maior que o comprimento da célula, causando uma maior probabilidade de interseção. Então foi proposto que o maior comprimento médio para todas as dimensões seja reduzido proporcionalmente, considerando o segundo maior comprimento médio

Assim, a Equação 2.9 foi proposta que o maior tempo médio para todas as dimensões deveriam ser proporcionalmente reduzidas considerando o segundo maior comprimento médio. Então, o novo comprimento médio é dado em μ que recebe h que é uma célula de histograma e i e j são os índices das dimensões com maior e segunda maior comprimento médio, m_i é o comprimento da célula h na dimensão i, c é a cardinalidade da célula, m_j é o comprimento da célula h na dimensão j, l é o vetor de comprimento médio.

$$\mu(h,i,j) = min\left(l_i, \frac{m_i}{(l_j \cdot c)/m_j}\right) \tag{2.9}$$

A segunda melhoria, proposta por Oliveira (2017), é um conjunto de fórmulas de seletividade, adequadas a objetos do tipo linha e polígonos.

O conjunto das equações seguem de (2.10) até (2.12) e são utilizadas para estimarem a interseção entre dois *datasets* com objetos do tipo linha e outro com objetos do tipo polígono. Essas equações determinam a cardinalidade de ($J_{l,p}$) da junção entre os objetos das duas células dos histogramas (a e b). O c é o comprimento da célula na dimensão k, l_{ak} e l_{bk} são as médias dos comprimentos de a e b na dimensão k, e para as cardinalidades das células a e b são representadas por $\bar{\bar{a}}$ e $\bar{\bar{b}}$, respectivamente. γ indica o fator que aumenta o comprimento médio dos polígonos em cada dimensão para completar o espaço da célula inteiro. A fórmula tem o objetivo de encontrar quantas vezes a linha do comprimento l_{ak} intersecta o polígono escalado de comprimento $l_{bk} * \gamma$. É utilizado para reduzir o valor retornado proporcionando a densidade ρ de polígonos em b, calculado usando a área b_a.

$$\rho = \frac{b_a}{\sum_{k=1}^{d} b_c} \tag{2.10}$$

$$\gamma = \sqrt[d]{\frac{\sum_{k=1}^{d} b_c}{\bar{\bar{b}} \sum_{k=1}^{d} l_{bk}}} \tag{2.11}$$

$$J_{l,p}(a,b) = \bar{\bar{a}} \cdot \rho \cdot \prod_{k=1}^{d} max\left(1.0, \frac{l_{ak}}{l_{bk} \cdot \gamma}\right) \tag{2.12}$$

Finalmente, as equações de (2.13) a (2.15) foram propostas para se estimar o número da interseção entre dois *datasets* apenas com objetos do tipo linha. A cardinalidade da junção é dada por ($J_{l,l}$) entre duas células de histogramas (a e b), h_x, é o comprimento

da linha da célula x, baseado no comprimento médio, assumindo que é uma linha em linha reta, e l_{abk} é o comprimento da interseção entre a e b na dimensão k. Outras variáveis são definidas abaixo, $0 \leq \eta \leq 1$ é o coeficiente da interseção da linha usada para reduzir o número de interseções, observando a probabilidade miníma que dois segmentos de linhas arbitrarias intersectam cada outro e o raio entre a comprimento médio das linhas para as duas células.

$$h_x = \sqrt{\sum_{k=1}^{d} (l_{xk})^2} \tag{2.13}$$

$$\eta = min\left(\frac{133}{432}, \frac{min\,(h_a, h_b)}{max\,(h_a, h_b)}\right) \tag{2.14}$$

$$J_{l,l}\,(a, b) = \bar{\bar{a}} \cdot \bar{\bar{b}} \cdot \eta \cdot \prod_{k=1}^{d} min\left(1.0, \frac{l_{ak} + l_{bk}}{l_{abk}}\right) \tag{2.15}$$

Para dois *datasets* com objetos do tipo polígono, é utilizado a Equação 2.4, mas corrigindo o vetor comprimento médio pela Equação 2.9. Então é referido como $J_{p,p}$.

3 TRABALHOS RELACIONADOS

Neste capítulo, é apresentado um levantamento e descrição de trabalhos relacionados que apresentavam junções e multijunções espaciais, métodos de estimativas de custos das junções, os tipos de histogramas existentes e trabalhos que utilizavam o Histograma de Euler. Assim, poder comparar os métodos existentes com suas particularidades com este trabalho. Logo, este capítulo está divido na seção 3.1 onde será apresentado os critérios de análise, o seção 3.2 apresenta os tipos de histogramas existentes. E por fim, a seção 3.3 apresenta uma tabela que ilustrada os tipos dos histogramas com cada critério.

3.1 Metodologia de análise

A base de trabalhos foram encontrados no mecanismo de busca Google Scholar[1] que indexa artigos do IEEE Explore, a ACM Digital Library e ACM Computing Reviews. O método de busca dos trabalhos relacionados foi com a utilização das palavras chaves deste trabalho, portanto, as palavras chaves consistiram-se em critérios que foram adotados para comparação entres os trabalhos. Assim, os trabalhos encontrados contêm temas relacionados a esta pesquisa ou técnicas para o cálculo da estimativa da seletividade de consultas espaciais. Os critérios estão descritos a seguir:

- **Junção Espacial (C1)** - O primeiro critério de busca de trabalhos com maior relevância a este foi trabalhos que apresentavam estudos e algoritmos que utilizam junções espaciais, pois junção espacial é um tipo de consulta que retorna informações importantes da interseção entre *datasets*.
- **Multijunção Espacial (C2)** - O segundo critério adotou um tipo de consulta espacial mais complexa que a junção espacial que é processada em etapas produzindo resultados intermediários, a multijunção espacial.
- **Estimativa de Custo (C3)** - O terceiro critério aborda as técnicas de estimativas de custo que consideram o tipo de objeto sendo linha ou polígono para o cálculo da seletividade de consultas espaciais.
- **Grades alinhadas (C4)** - No que se refere ao quarto critério foi adotado histogramas que apresentavam apenas junções com grades alinhadas, visto que grades alinhadas o retorno da estimativa de seletividade era calculado sem erro.

[1] https://scholar.google.com.br/.

- **Grades não alinhadas (C5)** - O quinto critério se refere aos histogramas que apresentavam *datasets* cujo a extensão espacial não se alinhavam, dado que podem haver erros no cálculo da seletividade quando as grades não se alinham sendo um cenário real para banco de dados espaciais.

3.2 Trabalhos analisados

A busca pelos trabalhos consistiu em encontrar trabalhos que continham uma estrutura de dados que apoia o cálculo da estimativa de seletividade das consultas espaciais. Três trabalhos foram selecionados e apresentam tipos de histogramas existentes. Os mesmo são descritos a seguir.

3.2.1 Histograma de Grade (T1)

Mamoulis e Papadias (2001a) definiram um histograma que divide o espaço do *dataset* em grades uniformes sendo C × C a grade e para cada célula (*bucket*) é armazenado o número de objetos e o comprimento total de sua MBR por eixo. Também definiram que o valor de cardinalidade de cada célula com base no número de objetos têm o centro de seu MBR dentro dos limites da célula.

Propuseram o uso do comprimento médio dos objetos como metadados adicionais em cada célula do histograma. Esses metadados foram definidos para cada dimensão com base na média de todos os objetos em uma determinada célula de histograma. Para objetos que se sobrepõem mais de uma célula do histograma, apenas a área dentro dos limites da célula é considerada.

3.2.2 Histograma IHWAF (T2)

Oliveira (2017) em um estudo realizado identificou que características sobre os *datasets* que o tipo dos objetos contidos nos *datasets* influenciavam na seletividade das junções. Então, um novo modelo estatístico foi proposto e avaliado para estimar o custo de consultas de multijunção espacial distribuídas e o modelo apresentou melhores resultados nos experimentos. Também propôs um método chamado Split que define o número de células em cada dimensão do histograma, baseado nos metadados dos *datasets*. Algumas razões foram citadas como a divisão da extensão espacial do conjunto de dados pelo comprimento médio dos objetos em cada dimensão produz um número de células que podem melhorar a precisão da estimativa de custos e um pequeno número de células também deve ser evitado, pois isso geraria partições de tamanho distorcidas, devido à natureza distorcida dos dados espaciais.

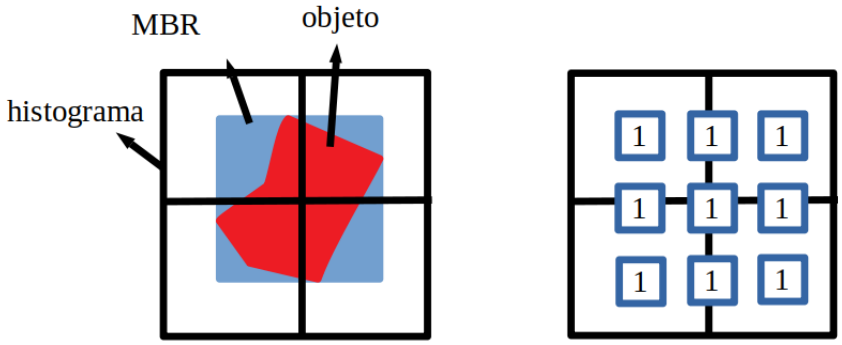
Além do modelo que considera o tipo dos objetos na estimativa o trabalho também apresentou um método estatístico de sobreposição proporcional para enquadrar os objetos, assim particiona o *dataset* em grades que não possuíam tamanhos fixos. Implementou-se um histograma denominado Histograma IHWAF que utiliza o modelo de sobreposição proporcional e o novo modelo estatístico.

3.2.3 Histograma de Euler (T3)

Sun, Agrawal e Abbadi (2002b) apresentaram uma nova abordagem para estimar a seletividade de junções espaciais que se consiste no Histograma de Euler. Porém, generalizaram a abordagem do Histograma de Euler para manipular objetos e janelas de seleção que não estão alinhadas com uma determinada grade.

Demonstraram a construção do Histograma de Euler que se baseia na teoria dos grafos, que aloca *buckets* dos objetos para as faces, arestas e vértices. Foi proposto com o propósito de evitar um problema presente nos histogramas que era a contagem múltipla de objetos. Apresentaram também que a construção do histograma generalizado foi validado a efetividade da estrutura.

Uma ilustração do funcionamento do Histograma de Euler pode ser visto na Figura 16. Logo podemos observar um *dataset* com um objeto e o Histograma de Euler alocando *buckets* para cada estrutura .



(a) *Dataset* com janela de seleção (b) Histograma de Euler

Figura 16 – Exemplo Histograma de Euler.

3.3 Resumo Comparativo

Nos trabalhos apresentados foi observado que existem trabalhos que não abordam algumas propriedades. Por exemplo, o trabalho de Mamoulis e Papadias (2001a) apresenta apenas Histograma para Grade que possuem as grades alinhadas e aborda apenas a junção espacial. O trabalho de Oliveira (2017) apresenta um Histograma de Grade porém com um método de sobreposição proporcional cuja construção também possui a capacidade

de realizar multijunções e produzir resultados intermediários e também apresentou uma estimativa de custo que considera o tipo dos objetos e também realiza junções espaciais para grades que não se alinham. O trabalho de Sun, Agrawal e Abbadi (2002b) elaborou um histograma baseado na teoria de Euler que aloca *buckets* para as faces, arestas e vértices como técnica para não contar múltiplas vezes os objetos e possui abordagem que permite uma junção para grades que não se alinham.

A Tabela 1 ilustra os trabalhos relacionados com os critérios apresentados. A tabela está dividida que do lado esquerdo apresenta os trabalhos relacionados e este trabalho[2], também apresenta o tipo de histograma que cada trabalho utilizou e os critérios que estão apresentados na barra superior.

Tabela 1 – Comparação entre os trabalhos relacionados com os critérios. A legenda para os símbolos: • indica a presença do critério em cada histograma e ○ indica a ausência.

Trabalho	**Tipo Histograma**	**Junção Espacial**	**Multijunção Espacial**	**Estimativa de custo**	**Grades alinhadas**	**Grades não alinhadas**
T1	Grade	•	○	○	•	○
T2	Grade	•	•	•	•	•
T3	Euler	•	○	○	•	•
Este trabalho	Euler	•	•	•	•	•

[2] Este trabalho é uma continuação de pesquisa dos trabalhos [(SANTOS; OLIVEIRA, 2019a; SANTOS; OLIVEIRA, 2019b)]. Estes trabalhos apresentam a elaboração do Histograma Intermediário de Euler.

4 IMPLEMENTAÇÃO E CONSTRUÇÃO DOS ALGORITMOS

Este capítulo apresenta a criação e desenvolvimento do Histograma Intermediário de Euler com a estimativa de custo que considera o tipo dos objetos, definido como Histograma Intermediário de Euler Avançado. Com isso, este capítulo apresenta os algoritmos e estruturas de dados utilizados para a criação do histograma e está organizado da seguinte forma: a seção 4.1 apresenta a elaboração do Histograma de Intermediário de Euler Avançado e suas propriedades, assim como os algoritmos utilizados. A seção 4.2 descreve a implementação das equações descritas em subseção 2.3.4. E finalmente, a seção 4.3 apresenta as considerações finais.

4.1 Implementação do Histograma Intermediário de Euler Avançado

Um histograma intermediário é construído a partir dos histogramas originais a partir de dois *datasets*. O Histograma Intermediário fará parte no processo de estimativa do custo dos planos de execução para consultas de multijunção espacial, como uma das entradas para as etapas intermediárias.

Assim, a função Construcao_Histograma_Intermediario_de_Euler_Avancado definida pelo Algoritmo 1 apresenta o processo de construção do histograma intermediário. A função recebe como parâmetros dois histogramas H_A e H_B. Entre as linhas 2 a 4 é definido a alocação de memória do histograma intermediário e também em qual histograma, H_A e H_B, o intermediário irá se baseará (de acordo com o predicado da próxima etapa). Assim, definido o histograma, as linhas seguintes apresentam a inserção dos valores para algumas propriedades que são quantidade de grades nos eixos x e y (xqtd e yqtd), o comprimento das grades (xtam e ytam) e vetores para as grades (xtics e ytics).

De maneira que a estrutura do Histograma de Euler é definida a alocar *buckets* para as faces, arestas e vértices é necessário então preencher as larguras, comprimentos máximos e mínimos de cada estrutura. Consequentemente, elaboramos nas linhas 13 a 21 uma estrutura que irá percorrer o histograma todo e atribuir os comprimentos máximos e mínimos. Além disso, a estrutura da face possui características extras que são os comprimentos médios dos objetos nos eixos x e y (avgWidth e avgHeight) e a média da área (avgArea).

O Algoritmo 2 apresenta um complemento do Algoritmo 1 e apresenta a inserção

Algoritmo 1 Estruturas de dados e inserção dos dados no Histograma Intermediário

```
1: função CONSTRUCAO_HISTOGRAMA_INTERMEDIARIO_DE_EULER_AVANCADO( H_A, H_B )
2:     H_I = novo Histograma Euler
3:     H_I.mbr = MBR-Intersecao( H_A.mbr, H_B.mbr )
4:
5:     H_I.xqtd = G.xqtd
6:     H_I.yqtd = G.yqtd
7:     H_I.xtam = G.xtam
8:     H_I.ytam = G.ytam
9:     H_I.xtics = G.xtics
10:    H_I.ytics = G.ytics
11:
12:    para cada g ∈ G.ehist | ehist ∈ (faces, arestas, vertices) faça
13:        H_I.faces = g.faces
14:        H_I.arestas = g.arestas
15:        H_I.vertices = g.vertices
16:
17:        H_I.faces.avgArea = g.faces.avgArea
18:        H_I.faces.avgHeight = g.faces.avgHeight
19:        H_I.faces.avgWidth = g.faces.avgWidth
20:    fim para
21:
```

da cardinalidade nas faces do Histograma Intermediário. Nas linhas 3 e 4 é definido estruturas que recebem os histogramas H_A e H_B que irão ser percorridos posteriormente. A partir da linha 6 contém estruturas de repetições cuja função é percorrer cada histograma. Assim, primeiramente irá ser percorrido cada célula do histograma H_A obtendo cada face com uma função auxiliar OBTER_FACE que retorna a face. E para cada célula do histograma H_A será percorrido o histograma H_B completo. Neste segundo passo de percorrer o histograma H_B sempre é verificado a existência de alguma interseção entre as células dos histogramas H_A e H_B. Consequentemente quando existe alguma interseção entre ambos então é calculado a face do histograma H_B para poder ser passada como parâmetro da função Estima-Cardinalidade-Com-AvgLengthFix definida anteriormente pelo Algoritmo 3.3 proposto por Oliveira (2017). O resultado desta função é atribuído na face do Histograma Intermediário.

Última parte da construção do histograma é definida na linha 20, que chama uma função auxiliar que é demonstrada no Algoritmo 3 e também é uma continuação dos algoritmos 1 e 2. Nesta parte é definida a atribuição das cardinalidades das arestas e vértices. Foi definido que o cálculo das arestas e vértices seriam efetuados posteriormente as faces, devido chegar a uma conclusão de que o valor da face influenciaria nos resultados das arestas e vértices. Posteriormente a função INSERE_DADOS_ARESTAS_VERTICES retorna o Histograma Intermediário, para poder ser utilizado em outra etapa de multijunção.

Algoritmo 2 Inserção dos dados no Histograma Intermediário - Parte 2

```
1:
2:     double resultado ← 0
3:     Seja p_a(i, j) ∈ P_a a partição na linha i coluna j de H_A
4:     Seja p_b(r, s) ∈ P_b a partição na linha r coluna s de H_B
5:
6:     para cada p_a(i, j) ∈ P_a faça
7:         f_a ← OBTER_FACE(H_A, i, j)          ▷ Obtém a face de H_A na linha i coluna j
8:         f_i ← OBTER_FACE(H_I, i, j)          ▷ Obtém a face de H_I na linha i coluna j
9:         para cada p_b(r, s) ∈ P_b faça
10:            se p_a(i, j).mbr ∩ p_b(r, s).mbr então
11:                f_b ← OBTER_FACE(H_B, r, s)  ▷ Obtém a face de H_B na linha i coluna j
12:
13:                f_i += Estima-Cardinalidade-Com-AvgLengthFix(f_a, f_b)
14:                resultado += Estima-Cardinalidade-Com-AvgLengthFix(f_a, f_b)
15:
16:            fim se
17:        fim para
18:    fim para
19:
20:    INSERE_DADOS_ARESTAS_VERTICES(H_a, H_i)
21:
22:    retorne H_I
23: fim função
```

Logo, o Algoritmo 3 apresenta a função INSERE_DADOS_ARESTAS_VERTICES que recebe como parâmetro o Histograma Intermediário H_I e o histograma utilizado como base para o intermediário H_A. Na linha 2 é definido uma estrutura que recebe o H_I para ser percorrido e atribuir os devidos valores para as arestas e vértices. Portanto, a partir da linha 4 encontra-se uma estrutura de repetição que percorre o histograma. Funções auxiliares como OBTER_FACE, OBTER_VERTICE e OBTER_ARESTA retornam cada face, vértice e aresta. Uma variável chamada percentual irá receber a redução da cardinalidade da face do intermediário com base na cardinalidade do histograma passado como parâmetro percorrendo a quantidade de faces em cada estrutura. Por exemplo, um vértice no extremo x e y contém apenas uma face, vértices que não estão nos extremos contém mais de uma face. Assim sendo, verificado o percentual de redução das faces será calculado a cardinalidade dos vértices e arestas estabelecendo uma proporção baseada nesta redução. Esta proporção consiste na divisão da redução percentual pela quantidade de faces, logo esta divisão é multiplicado pelo valor original do vértice ou aresta.

Algoritmo 3 Inserção dos dados no Histograma Intermediário - Parte 3

1: **função** INSERE_DADOS_ARESTAS_VERTICES(H_A, H_I)
2: Seja $p_i(i,j) \in P_a$ a partição na linha i coluna j de H_I
3:
4: **para cada** $p_i(i,j) \in P_a$ **faça**
5: $f_a \leftarrow$ `OBTER_FACE`(H_A, i, j) ▷ Obtém a face de H_A na linha i coluna j
6: $v_a \leftarrow$ `OBTER_VERTICE`(H_A, i, j) ▷ Obtém o vértice de H_A na linha i coluna j
7: $a_a \leftarrow$ `OBTER_ARESTA`(H_A, i, j) ▷ Obtém a aresta de H_A na linha i coluna j
8:
9: $f_i \leftarrow$ `OBTER_FACE`(H_I, i, j) ▷ Obtém a face de H_I na linha i coluna j
10: $v_i \leftarrow$ `OBTER_VERTICE`(H_I, i, j) ▷ Obtém o vértice de H_I na linha i coluna j
11: $a_i \leftarrow$ `OBTER_ARESTA`(H_I, i, j) ▷ Obtém a aresta de H_I na linha i coluna j
12:
13: double percentual = $\sum_{n=0}^{qtdFace}(f_i.cardin/f_a.cardin)$
14:
15: **se** $v_i \neq vazio$ **então**
16: $v_i = v_a * (percentual/qtdFace)$
17: resultado = $v_a.cardin * (percentual/qtdFace) + resultado$
18: **fim se**
19:
20: **se** $a_i \neq vazio$ **então**
21: $a_i = a_a * (percentual/qtdFace)$
22: resultado = $a_a.cardin * (percentual/qtdFace) - resultado$
23: **fim se**
24: **fim para**
25:
26: **fim função**

4.2 Técnicas Avançadas de Construção do Histograma

Descrito anteriormente, equações foram propostas com base nos objetos dos *datasets*. Primeiramente, utilizamos a Equação 2.9 que endereça a imprecisão dos MBR que induzem no campo de comprimento médio sobre objetos espaciais complexos (linhas e polígonos). Posteriormente, verifica-se qual o tipo do *dataset*, sendo polígono ou linha. Assim, verificado a ocorrência de interseção entre dois *datasets* com objetos do tipo linha e outro com objetos do tipo polígono ($J_{l,p}$), utiliza-se a Equação 2.12. Utilizar a Equação 2.15 quando ocorre a interseção entre dois *datasets* com objetos do tipo linha com linha ($J_{l,l}$). Por fim, quando a interseção ocorre com *datasets* do tipo polígonos ($J_{p,p}$) apenas endereça a impressão dos MBR.

Então, este conjunto de equações descritas estão contidos no Algoritmo 3.3 definido em Oliveira (2017) e é utilizado neste trabalho para o cálculo realizadas para as faces no histograma.

4.3 Considerações Finais

Neste capítulo, apresentou-se a elaboração de algoritmos utilizados para a construção do Histograma Intermediário de Euler Avançado. Em primeiro lugar, foi projetado o algoritmo cujo conteúdo contém a simplificação dos MBR que causavam erros de cardinalidade, como também as fórmulas propostas por Oliveira (2017), que estima a seletividade de junção quando os dois conjuntos de dados possuem objetos de linha ou quando os objetos de um conjunto de dados são do tipo linha e os objetos do outro são de um tipo de polígono.

Além disso, foi projetado a estrutura do Histograma Intermediário com propósito de melhorar a assertividade da seletividade de multijunções espaciais. Pelo algoritmo definiu-se estruturas de dados para a alocação de memória para as propriedades do histograma como também a atribuição de alguns dados. Posteriormente a atribuição da cardinalidade das faces utilizando um método de estimativa de custo mais avançada, como também a atribuição da cardinalidade das arestas e vértices com base em uma redução das faces sobre o valor original das arestas e vértices.

O Capítulo 5 conclui a avaliação do Histograma Intermediário apresentando uma avaliação geral das estimativas para consultas de multijunção espacial.

5 AVALIAÇÃO DA CRIAÇÃO DO HISTOGRAMA INTERMEDIÁRIO DE EULER E RESULTADOS

Neste capítulo é apresentado uma análise da construção do Histograma Intermediário de Euler Avançado. Para aspectos técnicos esta pesquisa foi classificada como sendo metodológica devido a implementar as fórmulas especificas dos *datasets* para o Histograma de Euler e de caráter quantitativo pois os resultados gerados serão um conjunto de valores de técnicas estatísticas, medindo assim a assertividade do histograma proposto. Logo, o restante deste capítulo está definido na seguinte forma: A seção 5.1 abrange o conjunto de dados utilizados com suas propriedades, as multijunções que serão realizadas e também as equações estatísticas utilizadas para a avaliação. A seção 5.2 descrevem e ilustram os resultados obtidos dos experimentos. Por fim, a seção 5.3 discute sobre os resultados de forma mais ampla.

5.1 Metodologia e Amostras de Avaliação

Esta seção abrange os métodos de avaliação assim como o conjunto de dados que são utilizados no Histograma Intermediário de Euler Avançado e além disso esta pesquisa também executou os experimentos em histogramas intermediários de Grade, Euler e o IHWAF. Assim a subseção 5.1.1 apresenta os dados utilizados e suas particularidades e as multijunções que serão realizadas nos experimentos. A subseção 5.1.2 apresenta as métricas utilizadas para avaliação.

5.1.1 Dados utilizados

Para compor o conjunto de dados utilizados para gerar a população da pesquisa, que é de caráter quantitativo, utilizou os *datasets* reais obtidos no *website* do Instituto Brasileiro de Geografia e Estatística (IBGE)[1], do Laboratório LAPIG do Instituto de Estudos Sócio-Ambientais da Universidade Federal de Goiás (UFG)[2] e do *Digital Chart of the World*[3]. Os *datasets* estão demonstrados na Tabela 2, e são encontrados informações específicas de cada *dataset* contendo o nome, a sigla utilizada, o tipo, o valor da cardinalidade e também o seu tamanho de memória. Os arquivos baixados serão utilizados no formato de arquivo

[1] https://ww2.ibge.gov.br/home/.
[2] www.lapig.iesa.ufg.br.
[3] https://gis-lab.info/qa/vmap0-eng.html.

Shapefile (SHP) e utilizando as bibliotecas GDAL[4] e GEOS[5] para extrair e processar a geometria de cada objeto espacial dentro dos arquivos.

Tabela 2 – *Datasets* utilizados nos experimentos

Nome	**Sigla**	**Tipo**	**Cardin.**	**Tam. Arq. SHP(MB)**
		***Datasets* Brasileiros**		
Alertas desmat. cerrado	A	Polígono	32.578	11,2
Hidrografia	H	Polígono	226.963	64,5
Rodovia	R	Linha	51.646	15,2
Municípios	M	Polígono	5.564	38,8
Vegetação	V	Polígono	2.140	4,7
		***Datasets* mundiais**		
Hidrografia Mundial	HM	Linha	943.638	243,2
Ferrovias	FM	Linha	194.261	28,7
Represas de água	RA	Polígono	338.860	136,7
Contorno de Relevo	CR	Linha	703.574	572,5
Cultura	CU	Polígono	123.746	69,3

A Tabela 3 apresenta as junções espaciais usadas nos experimentos e suas respectivas cardinalidades. As junções foram geradas a partir da combinação dos cinco primeiros *datasets* e também todas as combinações dos últimos *datasets*. O tamanho dos histogramas para os *datasets* escolhidos foram gerados e definidos pelo método de *Split* definido em (OLIVEIRA, 2017) e são apresentados na Tabela 4.

Tabela 3 – Junção Espacial utilizadas nos experimentos

Nome	**Consulta**	**Cardin. Junção**	**Nome**	**Consulta**	**Cardin. Junção**
J1	A $\bowtie$ H	4.868	J11	HM $\bowtie$ FM	58.885
J2	A $\bowtie$ R	3.395	J12	HM $\bowtie$ RA	530.782
J3	A $\bowtie$ M	34.261	J13	HM $\bowtie$ CR	449.309
J4	A $\bowtie$ V	34.672	J14	HM $\bowtie$ CU	269.301
J5	H $\bowtie$ R	55.766	J15	FM $\bowtie$ RA	5.975
J6	H $\bowtie$ M	268.369	J16	FM $\bowtie$ CR	47.106
J7	H $\bowtie$ V	252.830	J17	FM $\bowtie$ CU	121.007
J8	R $\bowtie$ M	70.304	J18	RA $\bowtie$ CR	22.128
J9	R $\bowtie$ V	63.339	J19	RA $\bowtie$ CU	79.002
J10	M $\bowtie$ V	15.678	J20	CR $\bowtie$ CU	234.900

O ambiente para a execução dos experimentos será em um computador fornecido pela Universidade Federal de Goiás, Regional Jataí, do Laboratório de Redes de Computadores e Sistemas Distribuídos. O computador tem um processador Intel Core I5-3470

[4] www.gdal.org
[5] https://trac.osgeo.org/geos/

Tabela 4 – Tamanhos determinados para cada *Dataset*

Dataset	**Tamanho Histograma**
Vegetação	7 × 7
Alertas desmat. cerrado	133 × 219
Municípios	14 × 13
Rodovia	77 × 75
Hidrografia	126 × 130
Cultura	203 × 78
Ferrovias	249 × 118
Represas de água	214 × 137
Contorno de Relevo	181 × 161
Hidrografia Mundial	233 × 125

CPU (Ive Bridge), rodando a 3.40GHz x 4 núcleos, 8 GB de memória RAM, o disco de 500 GB e o sistema operacional Linux Mint 18.1 (Serena).

5.1.2 Métricas

Para a elaboração do Histograma Intermediário de Euler Avançado foi considerado as propriedades estatísticas descritas em (OLIVEIRA, 2017) para o Histograma de Grade, no procedimento *Build-Intermed-Histogram* (Algoritmo 3.3). Porém, estas propriedades serão adaptadas segundo a especificidade do Histograma de Euler em relação à existência de células de borda adicionais. Dentre as propriedades que foram adaptadas estão o cálculo estimado da seletividade de junções espaciais para cada célula do histograma, o tamanho médio esperado dos objetos espaciais em cada célula, a largura média dos objetos em cada dimensão de cada célula e a área total dos objetos em cada célula.

Mediu-se, nos experimentos, a cardinalidade individual de cada estrutura do histograma, então (c_f, c_a e c_v) são as cardinalidades das faces, arestas e vértices respectivamente e foram obtidas através da soma simples do valor de cada estrutura nos *buckets* do histograma. Para medir a cardinalidade total de um histograma intermediário, ou seja, o tamanho do conjunto resultante de uma etapa de uma multijunção espacial, foi utilizada uma adaptação da Equação de Euler conforme definida para o Histograma de Euler original em (SUN; AGRAWAL; ABBADI, 2002a) e apresentada na Equação 5.1. Nesta equação, para cada *bucket* $i = 1..n$ do histograma, soma-se a cardinalidade na face f_i, subtrai-se a cardinalidade nas arestas a_i e soma-se a cardinalidade nos vértices v_i. O valor c resultante foi comparado com a cardinalidade esperada da junção espacial real de dois *datasets*.

$$c = \sum_{i=0}^{n} f_i - a_i + v_i \tag{5.1}$$

Além da avaliação da cardinalidade resultante para a junção espacial, foi avaliado o erro de cada estrutura individual do histograma em relação a um histograma intermediário

construído a partir do conjunto resultante da junção, ou seja, mediu-se o quão distante o histograma estimado utilizando o método proposto é distinto de um histograma construído a partir do *dataset* resultante da junção. Utilizou-se o Erro Relativo Médio (λ), definido na Equação 5.2, adaptada para a estrutura do Histograma de Euler, onde I é o conjunto completo de faces, arestas ou vértices, r_i é o valor da estrutura $i \in I$ no histograma real e e_i é o valor estimado para a estrutura $i \in I$ no histograma estimado.

$$\lambda = \frac{\sum_{i \in I} |r_i - e_i|}{\sum_{i \in I} r_i} \tag{5.2}$$

Outra medida realizada com o intuito de complementar o erro relativo médio do histograma como um todo e também para cada estrutura é apresentado a média dos erros de cada estrutura e o respectivo desvio padrão para analisar quanto distante estão os valores da média.

Também foi avaliado a porcentagem de erro por célula do histograma estimado em relação ao histograma real. Assim, exibiu os valores máximo e mínimo das faces, assim como a média das cardinalidades, o desvio padrão e a porcentagem das faces que apresentaram uma taxa erro menor que 5%, 10%, 20% e 50%.

5.2 Análise dos Resultados Obtidos

Esta seção engloba os resultados dos experimentos e estão divididos com base no tipo de avaliação. Os experimentos consistiram na execução de cada junção espacial definida na Tabela 3, seguida da captura dos valores especificados na seção anterior. Logo, a subseção 5.2.1 apresenta os resultados da análise das particularidades do Histograma Intermediário de Euler Avançado, sendo a verificação das faces, arestas e vértices. A subseção 5.2.2 descreve os resultados da cardinalidade total de cada multijunção. E por fim, na subseção 5.2.3 é apresentado uma avaliação sobre o erro por célula do Histograma Intermediário de Euler Avançado. Os resultados são ilustrados por gráficos de linhas a seguir, de forma a evidenciar a comparação. Em cada gráfico, o eixo horizontal indica a junção espacial e o eixo vertical a métrica da comparação.

5.2.1 Avaliação da Estrutura de Cada Multijunção Espacial

A Figura 17 apresenta a comparação das cardinalidades reais e estimadas de cada consulta de junção (A, B e C), além do erro relativo médio (D) para o Histograma Intermediário de Euler Avançado (HIEA). Analisando os gráficos é possível observar que em (A, B e C) é ilustrado as cardinalidades estimadas (em vermelho) e reais (em azul) de cada estrutura do histograma, sendo A a cardinalidade das faces, em B a cardinalidade das arestas e em C a cardinalidade dos vértices. No gráfico A é analisado que os resultados da cardinalidade do estimado em relação ao real obtiveram resultados que apresentaram um

comportamento ruim se distanciando do real, que podem ser melhores vistos nas junções ($J_3, J_5, J_6, J_{12}, J_{13}, J_{19}$ e J_{20}). Nos gráficos B e C, foi verificado que o erro foi propagado para as arestas e vértices cuja cardinalidades apresentaram piores resultados nas junções descritas das faces. O erro relativo médio (λ) em D apresenta valores acima de 30%, que indicam um erro de estimativa alto e a estrutura que apresentou as menores taxas de erros foi as faces. Algumas consultas merecem atenção nos trabalhos futuros, no entanto, para investigar a fonte dos erros das estimativas, como nas junções que apresentaram maiores distâncias das arestas e vértices em relação as faces. Como o erro relativo médio apresentou valores acima de 30%, indica que uma melhoria da estimativa das arestas e vértices podem auxiliar na redução do erro médio como um todo ou que equações que não consideram o cálculo das faces para o cálculo das arestas e vértices para estimar a cardinalidade podem ser necessárias.

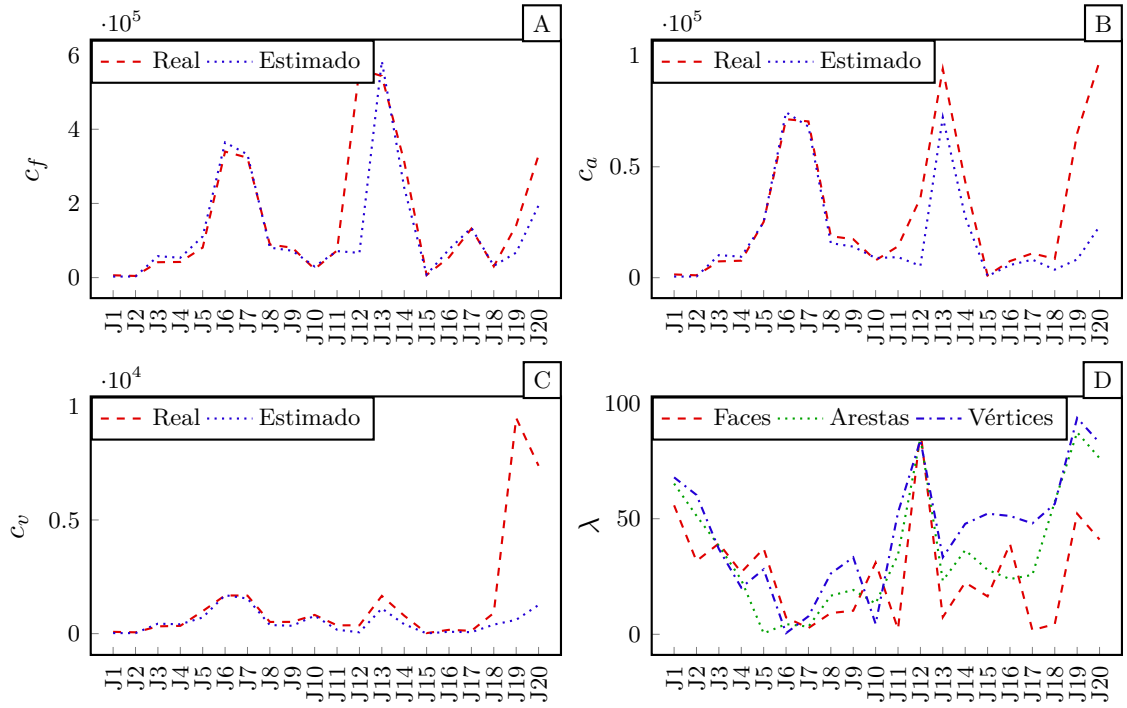


Figura 17 – Comparação das cardinalidades reais e estimadas e erro relativo médio para o HIEA. Em A, cardinalidade das faces, em B, cardinalidade das arestas, em C, cardinalidade dos vértices e em D o erro relativo médio para faces, arestas e vértices.

A Tabela 5 apresenta os resultados dos experimentos das cardinalidades reais e estimadas para cada consulta de junção. A tabela contém na primeira coluna as junções realizadas, seguido do tipo de cada objeto dos *dataset*, cada estrutura do Histograma de Euler, que são faces, arestas e vértices com suas cardinalidades reais e estimadas e também apresenta o λ para estrutura. Para complementar, é apresentado também a média dos λ e o seu desvio padrão (σ^2). A tabela está agrupada pelos tipos dos objetos dos *datasets* (L⋈L para *datasets* com objetos do tipo linha, P⋈P para *datasets* com objetos do tipo

polígono e L⋈P ou P⋈L para *datasets* com objetos do tipo polígono e linha) e ordenados pelo resultado do λ da face devido a ter apresentado melhores resultados que as arestas e vértices.

Tabela 5 – Resultado da cardinalidade de cada estrutura do Histograma de Euler

	Tipo	**Face**		**Aresta**		**Vértice**		(λ) (%)		
		Real	**Est.**	**Real**	**Est.**	**Real**	**Est.**	**Face**	**Are.**	**Vér.**
J_{11}	L⋈L	73.035	71.452	14.154	9.275	364	172	2,2	34,5	52,7
J_{13}	L⋈L	543.425	583.644	94.133	72.201	1.660	1.104	7,4	23,2	33,3
J_5	L⋈L	80.909	110.938	25.153	25.035	989	709	37,1	0,4	28,1
J_{16}	L⋈L	54.583	75.899	7.484	5.695	168	82	39,1	23,9	51,2
J_4	P⋈P	42.178	53.461	7.659	9.462	342	410	26,8	23,6	20,2
J_{10}	P⋈P	22.920	30.031	7.824	8.824	823	749	31,0	13,3	4,5
J_3	P⋈P	41.493	57.801	7.363	10.181	320	437	39,3	38,3	36,9
J_{19}	P⋈P	140.912	67.364	65.107	8.220	9.486	609	52,2	87,4	93,6
J_{17}	L⋈P	131.971	134.372	10.964	8.144	123	63	1,8	25,7	48,0
J_7	L⋈P	323.271	332.103	70.452	68.070	1.672	1.542	2,7	3,3	7,8
J_{18}	L⋈P	30.244	31.573	8.525	3.630	908	395	4,4	57,4	56,4
J_6	L⋈P	339.809	364.534	71.453	74.591	1.681	1.690	7,3	4,4	0,5
J_8	L⋈P	88.962	80.818	18.713	15.590	514	378	9,2	16,7	26,3
J_9	L⋈P	80.636	72.442	17.351	14.011	517	344	10,2	19,2	33,3
J_{15}	P⋈L	6.928	8.066	996	718	23	10	16,4	27,9	52,2
J_{14}	L⋈P	312.623	242.508	43.328	27.542	790	412	22,4	36,4	47,8
J_2	P⋈L	4.439	3.032	1074	522	58	23	31,7	51,3	60,3
J_{20}	L⋈P	330.024	194.760	97.617	23.201	7.382	1.258	41,0	76,2	82,9
J_1	P⋈L	6.285	2.781	1.469	510	78	25	55,8	65,2	67,9
J_{12}	L⋈P	561.772	66.633	35.939	5.339	368	59	88,1	85,1	84,0
Média								26,3	35,7	44,4
σ^2								22,5	26,5	26,1

As junções com *datasets* com objetos somente do tipo linha apresentou resultados do λ próximos a 60% o que indica uma taxa de erro alta. A estrutura do histograma que apresentou as maiores taxas de erros das junções foram os vértices cuja taxa de erros foram acima de 30% e posteriormente as arestas e faces, porém, em alguns casos o erro ficou abaixo de 10%. Analisando as junções com apenas objetos do tipo polígono, os λ demonstrou taxas de erros maiores que as junções do tipo linha com linha, com o maior resultado do λ nos vértices e obtendo uma taxa de erro maior que 90%, indicando um erro altíssimo. Neste cenário, diferente das junções com objetos do tipo linha, a estrutura que apresentou as maiores taxas de erros foram os arestas e não os vértices. Finalmente, as junções com objetos do tipo linha e polígono apresentaram casos com taxas de erros inferiores a 10% manifestando uma assertividade boa. Porém, casos como as junções J_1, J_{12} e J_{20} demonstraram resultados superiores a 50%, o que acusa uma assertividade ruim.

As junções J_7 e J_6, ambas com objetos do tipo linha e polígono, obtiveram os melhores resultados de todas as junções e o λ se manteve abaixo de 10% apontando uma assertividade alta. O pior cenário de linha com polígono foi a junção J_{12} que apresentou em todas as estruturas o resultado maior que 80% demonstrando ser muito ruim. Outro caso muito ruim na assertividade foi a junção J_{19} com objetos do tipo polígono que as taxas de erros de todas as estruturas resultaram em uma média de 77,73%. Analisando a média de todos os λ, foram um total de 26,3% para as faces, 35,7% para as arestas e para os vértices um total de 44,4% sendo a estrutura com as maiores taxas de erros. O desvio padrão para cada estrutura, foram 22,5% para face, 26,5% para as arestas e 26,1% para os vértices.

5.2.2 Avaliação da Cardinalidade Total de Cada Multijunção

A estimativa da cardinalidade total de cada junção espacial foi avaliada, comparando o erro relativo entre o histogramas intermediários porém cada um com sua particularidade. Assim os histogramas analisados foram o Histograma Intermediário de Euler (HIE) (SANTOS; OLIVEIRA, 2019a; SANTOS; OLIVEIRA, 2019b), o Histograma Intermediário de Grade (HIG) proposto em Mamoulis e Papadias (2001a), o Histograma Intermediário IHWAF proposto por Oliveira (2017), este trabalho, o Histograma Intermediário de Euler com a estimativa de custo avançada (HIEA) e a cardinalidade real das junções. O resultado é apresentado na Figura 18 e mostra o Erro Relativo dos histogramas.

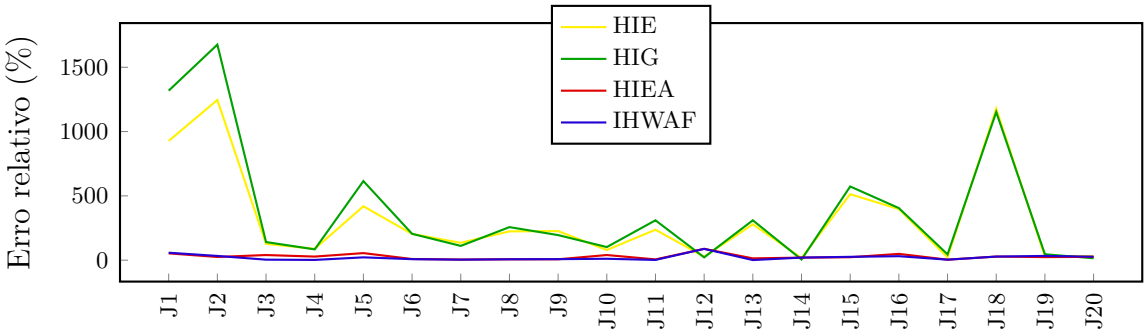


Figura 18 – Comparação da cardinalidade estimada para cada junção espacial entre o HIG, o HIE, o HIEA e o Histograma IHWAF

Na observação do gráfico é possível dividir os histogramas em dois grupos que são os que contém uma estimativa que considera o tipo dos objetos (HIEA e IHWAF) e os que não contém a estimativa (HIE e HIG). Os histogramas que apresentam uma estimativa que levam em consideração o tipo dos objetos obtiveram uma taxa de erro baixa comparada com os histogramas que não consideram os objetos. Assim, é evidenciado que dentre todos os histogramas avaliados o histograma HIG apresentou uma taxa de erro maior e a junção J_2 o erro chegou a ser 1500% maior que real. O método HIE que não apresenta uma estimativa de custo avançada obteve em alguns casos melhores resultados

que o HIG que podem ser melhor visualizados nas junções J_1, J_2, J_5, J_{11} e J_{15}. Entretanto seu comportamento foi semelhante ao HIG ao decorrer das junções, ocasionando que quando o erro presente no HIG era muito elevado, o HIE também errava porém com uma taxa menor. Os histogramas que contém uma estimativa de custo que considera os objetos como o HIEA e o IHWAF as taxas de erros se demonstraram muito baixas, o que conclui que os resultados destes histogramas foram muito bons. Porém, os resultados de ambos se mantiveram bem próximos indicando que é necessário uma avaliação mais profunda entre ambos.

Uma comparação mais aprofundada sobre os histogramas HIEA e IHWAF é apresentada pela Figura 19. No gráfico a análise do Erro Relativo dos histogramas é possível verificar que algumas ocasiões da comparação houve uma melhoria por parte de um histograma e vice versa. Com isso, a comparação dos histogramas foi dividida em três casos que são: a) HIEA obteve melhores resultados que o IHWAF; b) HIEA obteve piores resultados que o IHWAF; e c) ambos histogramas apresentaram resultados próximos (dito como próximos devido a ambos histogramas obterem uma diferença de taxa de erro inferior a 1%).

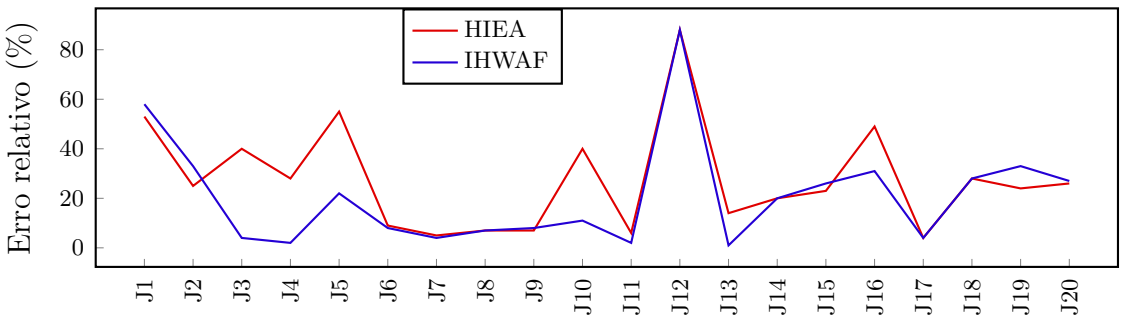


Figura 19 – Comparação da cardinalidade entre o Histograma IHWAF e o Histograma Intermediário de Euler Avançado

Pelo gráfico é possível observar que as junções que apresentaram melhores resultados no HIEA foram as junções J_1, J_2, J_{15} e J_{19} e apenas a J_{19} apresenta objetos do tipo polígono e o restante das junções consideram objetos do tipo linha com polígono. Para contemplar o caso em que o HIEA obteve resultados abaixo do IHWAF, estão as junções J_3, J_4, J_5, J_{10}, J_{11}, J_{13} e J_{16}. Destas junções referem-se as junções que os objetos são combinações de linha com linha e polígono com polígono, não contendo junção do tipo linha com polígono. No último caso, que são os histogramas com resultados próximos, as junções J_6, J_7, J_8, J_9, J_{12}, J_{14}, J_{17}, J_{18} e J_{20}. Todas estas junções descritas são junções com objetos do tipo linha com polígonos.

Entretanto, é constatado que houveram junções J_3, J_4, J_6, J_7, J_8, J_9, J_{11} e J_{13} que o IHWAF obteve uma taxa erro inferior a 10% que significa que o histograma obteve uma assertividade altíssima. Dentre essas junções, o HIEA não conseguiu alcançar ótimos resultados nas junções J_3 e J_4 que possuem objetos do tipo polígono.

Os resultados de todas as junções realizadas são apresentadas na Tabela 6 que contém as cardinalidades dos histogramas e a cardinalidade real e o Erro Relativo (%). Assim a tabela possui parâmetros que são a primeira coluna as consultas realizadas, seguindo do tipo de *dataset* de cada junção (L⋈L para *datasets* com objetos do tipo linha, P⋈P para *datasets* com objetos do tipo polígono e L⋈P ou P⋈L para *datasets* com objetos do tipo polígono e linha). E cada consulta contém a sua cardinalidade real (Real), as cardinalidades estimadas de modelos propostos sendo o Histograma IHWAF, o Histograma Intermediário de Grade (HIG), Histograma Intermediário de Euler (HIE) e este trabalho o Histograma Intermediário de Euler Avançado (HIEA), assim como o erro relativo de cada modelo em relação a cardinalidade real. Foram calculados também a média dos erros de cada histograma, assim como o desvio padrão (σ^2). A tabela está organizada e ordenada pelo erro relativo do IHWAF e separados pelo tipo do *dataset*.

Tabela 6 – Resultado da cardinalidade estimada utilizando o histograma de HIG, HIE, HIEA e IHWAF

	Tipo	**Real**	**Cardinalidade Estimada**				**Erro Relativo (%)**			
			HIG	**HIE**	**HIEA**	**IHWAF**	**HIG**	**HIE**	**HIEA**	**IHWAF**
J_{13}	L⋈L	449.309	1.843.256	1.703.895	512.548	443.251	310,2	279,2	14,1	1,3
J_{11}	L⋈L	58.885	241.473	198.756	62.350	57.553	310,1	237,5	5,9	2,3
J_5	L⋈L	55.766	398.869	289.816	86.613	67.781	615,3	419,7	55,3	21,5
J_{16}	L⋈L	47.106	237.809	234.438	70.286	61.722	404,8	397,7	49,2	31,0
J_4	P⋈P	34.672	63.746	65.373	44.410	35.281	83,9	88,56	28,1	1,8
J_3	P⋈P	34.261	82.560	78.013	48.057	32.837	141,0	127,7	40,3	4,2
J_{10}	P⋈P	15.678	31.707	28.164	21.957	17.442	102,2	79,6	40,0	11,3
J_{19}	P⋈P	79.002	42.986	55.768	59.753	53.255	45,6	29,4	24,4	32,6
J_{17}	L⋈P	121.007	66.298	93.907	126.292	126.257	45,2	22,4	4,4	4,3
J_7	L⋈P	252.830	534.048	595.651	265.575	264.011	111,2	135,6	5,0	4,4
J_8	L⋈P	70.304	250.912	228278	65.607	65.310	256,9	224,7	6,7	7,1
J_9	L⋈P	63.339	186.602	206.616	58.776	58.505	194,6	226,2	7,2	7,6
J_6	L⋈P	268.369	819.124	816.426	291.633	290.029	205,2	204,2	8,7	8,1
J_{14}	L⋈P	269.301	282.484	317.206	215.378	215.062	4,9	17,8	20,0	20,1
J_{15}	P⋈L	5.981	40.272	36.675	7.359	7.535	573,3	513,2	23,0	26,0
J_{20}	L⋈P	234.900	269.542	310.836	172.818	172.173	14,7	32,3	26,4	26,7
J_{18}	L⋈P	22.128	277.125	282.815	28.339	28.374	1.152,4	1.178,1	28,1	28,2
J_2	P⋈L	3.395	60.287	45.712	2.533	2.274	1.675,8	1.246,4	25,4	33,0
J_1	P⋈L	4.868	69.045	50.006	2.295	2.048	1318,3	927,2	52,8	57,9
J_{12}	L⋈P	531.269	414.137	386.062	61.353	63.287	22,0	27,3	88,5	88,1
Média							379,4	320,7	27,7	20,9
σ^2							473,5	375,0	21,6	21,7

A partir da tabela é observado que os maiores valores estão contidos no HIG ao longo de todas as junções realizadas. O HIE obteve uma vantagem sobre o HIG em que é possível observar na tabela que o HIE conseguiu estimar a cardinalidade com maior precisão em 13 das 20 consultas e as consultas com estimativas melhores e relevantes foram J_1, J_2, J_5, J_{11} e J_{15}. O HIEA e o IHWAF foram os histogramas que obtiveram melhores resultados dentre as junções, tendo um acerto de 17 das 20 junções em relação aos histogramas HIG e HIE, portanto, foi analisado com maior foco entre os histogramas

HIEA e IHWAF. Assim, quando as consultas foram com *datasets* do tipo linha com linha, o HIEA não conseguiu obter vantagem comparado com o IHWAF, logo os resultados do HIEA permaneceram abaixo do IHWAF. Nesta situação é possível observar uma tendência que as cardinalidades do HIEA sempre foram superiores as cardinalidades reais. Para as consultas do tipo polígono com polígono, apenas em uma ocasião o HIEA levou vantagem sobre o IHWAF e a única que apresentou uma cardinalidade menor que a real, na junção J_{19}. Para as consultas entre polígonos com linhas, ou vice-versa, o HIEA conseguiu obter dentre as 12 junções, apenas 3 ganhos presentes nas junções J_1,J_2 e J_{15} sobre o IHWAF e as outras 9 foram consideradas próximas devido a possuírem uma diferença de valor inferior a 1% e neste cenário o IHWAF não obteve nenhum ganho sobre o HIEA. Analisando com mais clareza os resultados é possível perceber que o HIG conseguiu uma vantagem sobre os demais histogramas nas junções J_{12},J_{14} e J_{20}. Esse comportamento foi verificado também no HIE porém apenas nas junções J_{12} e J_{14}. Ao final, é exibido as médias dos erros relativos para cada histograma e seus desvios padrão. Os resultados das médias dos modelos, HIG, HIE, HIEA e IHWAF foram 379,4%, 320,7%, 27,7% e 20,9% respectivamente. O desvio padrão do HIG foi 473,5%, para o o HIE foi 375%, para o HIEA foi 21,6% e IHWAF foi 21,7%.

5.2.3 Avaliação da seletividade de junção por célula do histograma

Também foi averiguado a precisão da cardinalidade estimada para cada célula do histograma resultante, calculado pelo procedimento de construção do Histograma Intermediário de Euler Avançado presente no Algoritmo 1. Então, utilizou a cardinalidade estimada para cada célula do histograma em todas as consultas de junção (J_1 a J_{20}) e a comparou com os valores reais obtidos, assim o cálculo para cada célula é utilizando a Equação 5.1.

Os resultados são mostrados na Tabela 7. A tabela contém nas quatro primeiras colunas que descrevem os valores máximos e mínimos, média e o desvio padrão σ^2. E nas últimas quatro colunas é apresentado uma porcentagem de células do histograma resultantes para as quais o erro se encaixa em um intervalo de 5%, 10%, 20% e 50%. Os valores mínimos possuem valores negativos devido a construção do histograma que quando percorre as células apenas a aresta inferior e a aresta lateral esquerda que são utilizadas no cálculo, assim é possível, em alguns casos, que os valores das arestas serem maiores que a face, resultando em um número negativo. Assim o menor valor é -24 e o maior valor é 996 ambos presentes na junção J_{10}. As quatro primeiras junções da tabela apresentaram 90% da quantidade de células com uma taxa de erro inferior a 5%, desde modo a maioria das células estimaram a cardinalidade com poucos erros. A taxa do erro nas junções seguintes foi diminuindo gradativamente até o fim com mais de 60% das células com erro menor que 5%, com apenas a junção J_{10} que apresentou uma taxa de 49,5% das células com erro

de 5%. A melhor consulta desta tabela a J_{15} apresentou um comportamento diferente da Tabela 6, que conseguiu estimar com 94,2% das células com o erro menor que 5% porém apresenta um Erro Relativo de 23% na tabela anterior. Este caso ocorre devido a baixa cardinalidade da junção que mesmo com um menor número de estimativas incorretas ou um grande número de erros menores causa um erro substancial ao somar os valores para obter a cardinalidade da consulta. Averiguando e analisando a consulta a J_{10} apresentou um comportamento igual nas duas tabelas, na tabela anterior quando analisado apenas os *datasets* com objetos polígonos com polígonos, apresentarão um dos piores resultados e nesta tabela apresentou a maior taxa de erro. Isto acontece porque está relacionado à compensação que ocorre quando a estimativa oscila entre alta e baixa, comparada com a cardinalidade real da célula.

Tabela 7 – Estatísticas para resultados estimados de cardinalidade por célula de histograma para consultas de junção.

	Estatísticas Globais				**% das células com erro**			
Junção	Min	Max	σ^2	Média	≤ 5	≤ 10	≤ 20	≤ 50
J_{15}	-17	125	2,4	0,3	94,2	94,3	94,7	95,9
J_2	-3	80	0,9	0,1	93,7	93,7	93,8	94,3
J_1	-1	35	0,6	0,1	90,6	90,7	90,8	91,4
J_{17}	-4	877	27,3	4,3	90,5	91,1	92,3	95,7
J_{19}	-19	206	9,6	2,0	89,0	89,5	90,6	94,1
J_{16}	-4	236	10,3	2,4	88,7	89,4	90,7	93,5
J_{11}	-3	152	8,2	2,1	87,8	88,9	90,8	95,0
J_{20}	-22	740	30,1	5,9	87,5	88,0	89,2	92,4
J_{18}	-17	207	5,0	1,0	86,1	86,4	87,1	89,5
J_{14}	-4	828	30,9	7,4	83,6	84,5	86,4	91,5
J_4	-1	91	4,7	1,5	79,9	82,7	87,0	97,8
J_3	-1	88	5,0	1,6	77,7	81,2	84,6	94,0
J_8	-1	255	25,8	11,4	74,7	78,7	84,5	95,6
J_9	-1	272	23,4	10,2	74,6	78,2	84,2	96,1
J_{13}	-11	497	40,8	17,6	71,7	73,6	77,5	86,5
J_{12}	-2	154	9,1	2,1	68,9	69,0	69,2	69,9
J_5	-3	81	10,0	5,3	68,6	70,3	73,5	81,4
J_7	-2	508	26,0	16,2	64,2	68,7	77,8	95,2
J_6	-2	455	28,0	17,8	64,0	68,3	76,1	94,8
J_{10}	-24	996	200,5	120,6	49,5	53,8	59,3	76,4

5.3 Considerações Finais

Este capítulo apresentou a base de dados utilizados neste trabalho utilizando *datasets* reais obtidos de fontes de instituições conceituadas para analisar a assertividade do histograma em uma ocasião onde os objetos contidos nos *datasets* eram diversificados e com suas próprias particularidades. Descreveu-se os experimentos que foram realizados e

as fórmulas estatísticas empregadas para analisar os experimentos, as quais permitiram uma melhor apresentação e interpretação dos resultados.

Apresentou-se os resultados dos experimentos abrangendo os resultados de cada junção e verificando a assertividade da estrutura das junções como também a cardinalidade total de cada junção. Analisando a comparação das cardinalidades reais com as estimadas de cada estrutura de cada junção foi verificado que o erro da face foi propagado para as arestas e vértices e o λ apresentou uma taxa de erro média para as faces de 26,3%, para as arestas 35,7% e vértices 44,4%, indicando que a construção do cálculo das arestas e vértices necessitam de mais atenção por possuírem uma taxa de erro alta.

Analisando a cardinalidade total das junções, foi possível comparar o HIE com o HIG e verificou-se uma melhora relativa na estimativa da seletividade, na qual o HIE conseguiu estimar a cardinalidade mais assertivamente em 13 das 20 consultas analisadas comparado ao HIG. Analisando os histogramas que contêm uma estimativa de custo que considera o tipo dos objetos (HIEA e IHWAF) como o HIE e HIG, foi verificado que o HIG e HIE ficam em desvantagem em 17 das 20 consultas em relação aos histogramas que consideram os tipos dos objetos. Comparando os histogramas HIEA e o IHWAF foi possível classificar em três casos que foram: os que o HIEA obteve vantagem sobre o IHWAF, os que o HIEA não obteve vantagem sobre o IHWAF e que ambos demonstraram uma assertividade próxima. Todavia, se compararmos a quantidade de junções presentes em cada caso, temos 4, 7 e 9 respectivamente. Assim é possível identificar que o HIEA não obteve uma vantagem relevante sobre o IHWAF, portanto concluímos que o HIEA não se sobressaiu ao IHWAF.

Apesar da diferença ser pequena entre os métodos para algumas consultas, para as consultas que o HIEA tem a pior estimativa a diferença é pequena. Isso indica que o método proposto não obteve resultados ruins e que melhorias na construção podem melhorar os resultados. Um caso foi verificado que a estimativa de custo que considera os objetos pode retornar a cardinalidade para as faces abaixo do que deveria, assim consequentemente o cálculo que subtrai das arestas irá diminuir ainda mais a cardinalidade total. Uma possível fonte de erro pode ser da distribuição dos objetos em um *dataset* podendo ser muito densa em alguns pontos e pouco densa em outros, motivo este que teve casos em que o HIEA melhorou a estimativa. A construção e a implementação da estrutura do HIEA encontra-se disponível em plataforma de hospedagem de código-fonte, chamada Github[6].

[6] O código do HIEA e dos experimentos está disponível em https://github.com/thborges/dgeohistogram.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este capítulo tem como objetivo apresentar os principais pontos discutidos no trabalho, relacionar os possíveis trabalhos futuros advindos desta pesquisa e avaliar a principal contribuição deste trabalho para a área científica.

6.1 Conclusões

Este trabalho apresentou um novo método de construção de Histogramas Intermediários de Euler (HIE) com a utilização de técnicas avançadas (HIEA) para estimativa de seletividade de consultas de multijunção espacial, baseando-se nas técnicas propostas para o Histogramas de Euler e considerando *datasets* cuja extensão espacial não se alinha, ou seja, um cenário real para banco de dados espaciais. Os *datasets* escolhidos consistem em dados reais, como um *dataset* de limites políticos de municípios brasileiros, e também apresentam características distintas uns dos outros contendo objetos com o formato de linhas ou polígonos.

Experimentos foram realizados com um total de 20 consultas e buscou-se analisar a assertividade da estrutura de cada junção, a cardinalidade total de cada junção e o erro presente em cada célula do histograma intermediário resultante. Nos resultados da estrutura do histograma foi verificado que o erro das faces propagou-se para as arestas e vértices, o que pode ter sido ocasionado pelo cálculo das arestas e vértices utilizar o valor das faces. Nas comparações das cardinalidades das faces, arestas e vértices constatou-se que os maiores erros estavam presentes nas arestas e vértices. Em alguns casos o erro passou de 50% e foram considerados ruíns. Também foi verificado que o Erro Relativo Médio ficou com uma média de 40% indicando que o erro não foi baixo, com isso, necessitando de uma atenção para o cálculo das estruturas.

Na comparação das cardinalidades totais dos histogramas foi exibido quatro histogramas com suas particularidades e ao final seu erro relativo. Assim, foi constatado as piores estimativas da seletividade ficaram com os histogramas Intermediário de Grade (HIG) e o Intermediário de Euler (HIE) que obtiveram uma desvantagem sobre os demais em 17 das 20 junções. Porém, entre estes histogramas, o HIE conseguiu uma estimativa mais assertiva que o HIG. Em relação aos histogramas que obtiveram uma melhora na assertividade estão o Histograma Intermediário de Euler Avançado (HIEA) e o IHWAF, os quais apresentam uma estimativa que considera o tipo dos objetos. Neste cenário, foi apontado que ambos os histogramas tiveram uma melhora na estimativa da seletividade

comparado com outros histogramas que não continham uma estimativa de custo que considera os tipos de objetos. E quando comparados entre si, foi constatado que o HIEA não conseguiu se sobressair significativamente sobre o IHWAF.

Finalmente, foi observado que a melhoria na precisão da estimativa da seletividade ocorreu devido a utilização de técnicas avançadas para a estimativa de seletividade do histograma que consideraram o tipo dos objetos presentes nos *datasets*, sendo estes objetos do tipo linha ou polígonos. Deste modo, concluímos que o que se demonstrou mais eficiente na melhoria da estimativa foi tal técnica e não o tipo de histograma utilizado.

6.2 Trabalhos futuros

Como trabalhos futuros, deve-se investigar e averiguar novas técnicas para estimar as cardinalidades das arestas e vértices no histograma intermediário de Euler, e também verificar se aprimoramentos no comprimento dos objetos influenciariam no cálculo equivalente, similar ao que acontece no cálculo das faces e da própria multijunção.

É possível ainda analisar como o HIEA propaga o erro para as próximas etapas das multijunções. Devido termos avaliado neste trabalho apenas a primeira etapa da multijunção, o Histograma Intermediário gerado foi construído a partir de histogramas originados dos próprios *datasets*. Posteriormente, outras etapas consistem na criação do Histograma Intermediário a partir de outros dois ou mais Histogramas Intermediários. Neste cenário, seria interessante verificar qual tipo de plano de multijunção, definidos na subseção 2.2.2, que o Histograma Intermediário de Euler obteria melhores resultados.

Outro trabalho futuro interessante é uma possível integração ao HIEA com um histograma existente, cuja fundamentação baseia-se na distribuição dos objetos de acordo com sua densidade, como o histograma Minskew.

REFERÊNCIAS

AN, N.; YANG, Z.-Y.; SIVASUBRAMANIAM, A. Selectivity estimation for spatial joins. In: IEEE. *Proceedings 17th International Conference on Data Engineering*. Heidelberg, Germany, 2001. p. 368–375. Citado 4 vezes nas páginas 24, 25, 27 e 31.

BEIGEL, R.; TANIN, E. The geometry of browsing. In: SPRINGER. *Latin American Symposium on Theoretical Informatics*. Campinas, Brasil, 1998. p. 331–340. Citado 2 vezes nas páginas 30 e 33.

BRINKHOFF, T.; KRIEGEL, H.-P.; SEEGER, B. Parallel processing of spatial joins using r-trees. In: IEEE. *Proceedings of the Twelfth International Conference on Data Engineering*. Washington, DC, USA, 1996. p. 258–265. Citado 3 vezes nas páginas 16, 21 e 22.

CAMPBELL, J. E. *Geographic information system basics*. [S.l.]: The Saylor Foundation, 2015. Citado 3 vezes nas páginas 9, 19 e 20.

CORMODE, G. et al. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends® in Databases*, Now Publishers, Inc., v. 4, n. 1–3, p. 1–294, 2011. Citado na página 25.

FITZ, P. R. *Geoprocessamento sem complicação*. [S.l.]: Oficina de textos, 2018. ISBN 978-85-86238-82-6. Citado 2 vezes nas páginas 19 e 20.

FORNARI, M. R.; COMBA, J. L. D.; IOCHPE, C. Query optimizer for spatial join operations. In: ACM. *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*. Arlington, VA, USA, 2006. p. 219–226. Citado 2 vezes nas páginas 24 e 25.

FRANÇA, A. G. *Preciso 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 4 vezes nas páginas 9, 22, 31 e 32.

HARARY, F. *Graph theory. 1969*. [S.l.]: Addison-Wesley, Reading, MA, 1969. Citado na página 30.

HUISMAN, O.; BY, R. D. Principles of geographic information systems. *ITC Educational Textbook Series*, v. 1, p. 17, 2009. Citado na página 20.

IOANNIDIS, Y. E.; POOSALA, V. Histogram-based approximation of set-valued query-answers. In: VLDB. *Proceedings of the 25th International Conference on Very Large Data Bases*. San Francisco, CA, USA, 1999. v. 99, p. 174–185. ISBN 1-55860-615-7. Citado na página 26.

JACOX, E. H.; SAMET, H. Spatial join techniques. *ACM Transactions on Database Systems (TODS)*, Acm, v. 32, n. 1, p. 7, 2007. Citado na página 22.

LIU, Q.; YUAN, Y.; LIN, X. Multi-resolution algorithms for building spatial histograms. In: AUSTRALIAN COMPUTER SOCIETY, INC. *Proceedings of the 14th Australasian database conference-Volume 17*. Adelaide, Australia, 2003. p. 145–151. Citado na página 16.

MAMOULIS, N.; PAPADIAS, D. Advances in spatial and temporal databases. volume 2121 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, p. 424–475, 2001. Citado 7 vezes nas páginas 16, 17, 25, 26, 37, 38 e 51.

MAMOULIS, N.; PAPADIAS, D. Multiway spatial joins. *ACM Transactions on Database Systems (TODS)*, ACM, v. 26, n. 4, p. 424–475, 2001. Citado 5 vezes nas páginas 16, 21, 22, 23 e 25.

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 18 vezes nas páginas 9, 17, 23, 24, 25, 26, 27, 28, 33, 34, 37, 38, 41, 43, 44, 46, 47 e 51.

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 Brazilian Symposium on Databases*. Petrópolis, RJ, Brazil: Laboratório Nacional de Computação Científica, 2015. p. 89–100. Citado 3 vezes nas páginas 16, 24 e 27.

PAPADIAS, D.; MAMOULIS, N.; THEODORIDIS, Y. Processing and optimization of multiway spatial joins using r-trees. In: ACM. *Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. Philadelphia, PA, USA, 1999. p. 44–55. Citado na página 23.

PITA, L. C. *Avaliação do Particionamento de Dados através de Histogramas Espaciais em Sistemas Distribuídos*. 50 p. Monografia — Universidade Federal de Goiás, Regional Jataí, Jataí, GO, Brasil, 2016. Citado 2 vezes nas páginas 9 e 23.

RIGAUX, P.; SCHOLL, M.; VOISARD, A. *Spatial databases: with application to GIS*. [S.l.]: Elsevier, 2001. Citado 2 vezes nas páginas 19 e 20.

SANTOS, M. C.; OLIVEIRA, T. B. de. Estimativa de custo de multijunções espaciais usando histogramas de euler intermediários. In: CONPEEX (2019). *Anais do Congresso de Pesquisa, Ensino e Extensão - ISSN 2447-8695*. Goiânia (GO), Brasil, 2019. To appear. Citado 2 vezes nas páginas 39 e 51.

SANTOS, M. C.; 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 2 vezes nas páginas 39 e 51.

SUN, C.; AGRAWAL, D.; ABBADI, A. E. Exploring spatial datasets with histograms. In: IEEE. *Proceedings 18th International Conference on Data Engineering*. Washington, DC, USA, 2002. p. 93–102. Citado 4 vezes nas páginas 28, 30, 32 e 47.

SUN, C.; AGRAWAL, D.; ABBADI, A. E. Selectivity estimation for spatial joins with geometric selections. In: SPRINGER. *International Conference on Extending Database Technology*. Prague, Czech Republic, 2002. p. 609–626. Citado 6 vezes nas páginas 17, 27, 28, 32, 38 e 39.

WEST, D. B. et al. *Introduction to graph theory*. [S.l.]: Prentice hall Upper Saddle River, NJ, 1996. v. 2. Citado na página 29.