Andrey G. França

Precisão da Estimativa de Seletividade de Tarefas de Junção Espacial Distribuída usando Histogramas de Euler

Jataí-Goiás

2018

Andrey G. França

Precisão da Estimativa de Seletividade de Tarefas de Junção Espacial Distribuída usando Histogramas de Euler

Universidade Federal de Goiás - Regional Jataí - UFG-REJ

Instituto de Ciências Exatas e Tecnológicas (ICET)

Bacharelado em Ciências da Computação

Orientador: Prof. Dr. Thiago Borges de Oliveira

Jataí-Goiás

2018

Andrey G. França
Precisão da Estimativa de Seletividade de Tarefas de Junção Espacial Distribuída usando Histogramas de Euler/ Andrey G. França. – Jataí-Goiás, 2018-
62 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. 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, 2018.

1. Banco de Dados Espaciais. 2. Sistemas de Informações Geográficas. 3. Processamento Distribuído. 4. Junção Espacial

Andrey G. França

Precisão da Estimativa de Seletividade de Tarefas de Junção Espacial Distribuída usando Histogramas de Euler

Trabalho aprovado. Jataí-Goiás, data da defesa:

Prof. Dr. Thiago Borges de Oliveira
Orientador

Prof. Mestra Franciny Medeiros Barreto
Avaliador

Prof. Mestre Bruno Moraes Rocha
Avaliador

Jataí-Goiás
2018

Este trabalho é dedicado à minha companheira, meu filho e minha família.

AGRADECIMENTOS

Agradeço à Deus, minha companheira e filho, minha família, ao meu professor orientador que me acompanhou durante quase toda graduação e meus amigos que me acompanharam durante essa caminhada.

"Em vista da ausência de asas, imagine."
(Ramidielque Lima)

RESUMO

O processamento de dados espaciais teve um aumento significativo nos últimos anos, e os dispositivos computacionais dotados de GPS (*Global Positioning System*) e rede de comunicação (2G, 3G e outras) como celulares, smartphones e sensores estão cada vez mais comuns e acessíveis. Há uma grande disponibilidade de dados espaciais: imagens geolocalizadas, dados abertos dos governos federais, estaduais e municipais, mapeamento de lojas comerciais, levantamento de dados georreferenciados por entidades governamentais, dentre outros. Com esses dados espaciais, novas informações podem ser adquiridas a partir desses dados. Um exemplo de processamento de dados espacias é a consulta espacial, que encontra em dois ou mais conjunto de dados informações correlacionadas. A consulta espacial pode ter sua execução complexa devido a quantidade de dados envolvidos. Os sistemas computacionais que realizam esses processamentos não têm evoluído na mesma proporção, de forma a não atender toda a demanda. Sendo assim, é muito adotado na literatura a execução paralela das consultas espaciais por sistemas distribuídos. Em um sistema distribuído uma consulta é particionada, de forma que várias máquinas processam uma parte desta consulta. Dessa forma, é preciso que a distribuição das tarefas em um *cluster* seja otimizada para que a consulta seja executada de forma a despender o menor tempo de execução possível. Um parâmetro utilizado para essa divisão de tarefas é a seletividade. Esta pesquisa apresenta a análise de métodos que estimam a seletividade de uma consulta. Foram realizados diversos experimentos utilizando o histograma de Euler e comparando com o histograma de grade. Nossos experimentos mostraram que no cenário de sistemas distribuídos e com grades distintas para cada *dataset*, o histograma de Euler possui um resultado pior do que o histograma de grade. Em outros cenários, foi confirmado o desempenho superior do método conforme o artigo original.

Palavras-chaves: *Banco de Dados Espaciais; Sistemas de Informações Geográficas; Processamento Distribuído, Junção Espacial.*

ABSTRACT

Spatial data processing has grown significantly in recent years, and computing devices equipped with GPS (Global Positioning System) and communication network (2G, 3G and others) such as mobile phones, smartphones and sensors are increasingly common and affordable. There is a great availability of spatial data: geolocalized images, open data of federal, state and municipal governments, mapping of commercial stores, georeferenced data collection by governmental entities, among others. New information can be acquired from these data. An example of spatial data processing is spatial join query, which finds correlated information in two or more datasets. Spatial join query can be complex because of the amount of data involved. The computational systems that perform these processes have not evolved in the same proportion, so as not to meet the demand. Thus, the parallel execution of spatial queries by distributed systems is much adopted in the literature. In a distributed system a query is partitioned, so that multiple machines process a portion of this query. In this way, the distribution of the tasks in a cluster is optimized so that the query is executed in order to spend the shortest execution time possible. One parameter used for this division of tasks is selectivity. This research presents the analysis of methods that estimate the selectivity of a query. Several experiments were performed using the Euler histogram and comparing with the grid histogram. Our experiments showed that in the distributed systems scenario and with different grids for each dataset, the Euler histogram has a worse result than the grid histogram. In other scenarios, the superior performance of the method was confirmed as per the original article.

Key-words: *Spatial Databases; Geographic Information Systems; Parallel Processing, Spatial Join.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

Para executar consultas de junção espacial eficientemente o otimizador de um banco de dados seleciona o melhor plano de execução com base na estimativa do esforço computacional necessário. Uma métrica frequentemente usada para medir este esforço é a seletividade das junções espaciais, a qual pode ser computada usando histogramas espaciais. Um histograma espacial é um tipo de aproximação do *dataset* real, e portanto contém erros que impactam diversos aspectos da execução da consulta, como por exemplo, a escolha do melhor plano de execução e o balanceamento da execução em sistemas distribuídos. Apesar de existirem propostas de histogramas espaciais que proporcionam estimativas de seletividade mais precisas, como o histograma de Euler (SUN; AGRAWAL; El Abbadi, 2002), geralmente o foco é no calculo do seletividade global da consulta de junção espacial simples (com dois datasets). A execução distribuída desta consulta, no entanto, necessita que o custo individual das atividades de cada processador do *cluster* seja estimado para que se possa executá-la de forma balanceada.

A junção espacial é um tipo de consulta muito estudada em sistemas de banco de dados espaciais (SBDE), não só porque o volume de dados processados é grande, mas também, por envolver predicados espaciais com algoritmos complexos (JACOX; SAMET, 2007). Dessa forma, o processamento em sistema distribuídos é tido como uma solução, pois tem como objetivo dividir o trabalho em várias máquinas para aumentar o desempenho. Para tanto, o sistema distribuído exige que o processamento seja dividido em fragmentos e cada fragmento possui uma complexidade distinta. Tal complexidade deve ser estimada para que o processamento seja alocado no *cluster* de maneira a despender o menor tempo de processamento possível.

Um histograma espacial é uma estrutura de dados que divide o espaço geográfico do *dataset* em células que registram a quantidade de objetos espaciais no fragmento do espaço que a célula representa. Dessa forma, a estimativa de seletividade global pode ser calculada como a soma das estimativas de cada célula (MAMOULIS; PAPADIAS, 2001). Além disso, pode ser empregado em sistemas distribuídos, onde as células do histograma são utilizadas como partições que são distribuídas na máquina de um *cluster*. Um exemplo de histograma multidimensional é o histograma de grade, onde as células possuem tamanhos fixos (OLIVEIRA, 2017). No entanto os histogramas de grade sofrem algumas desvantagens, que impactam diretamente no cálculo da estimativa da seletividade.

Um método proposto na literatura para resolver os problemas do histograma de

grade é o histograma de Euler (SUN; AGRAWAL; El Abbadi, 2002), o qual utiliza resultados da teoria dos grafos. Oliveira (2017) pesquisou, entre outros tópicos, a estimativa de seletividade em junções espaciais e a distribuição de tarefas em um *cluster*, e identificou que para diminuir a taxa de erros, técnicas de histogramas melhoradas devem ser empregadas, e sugeriu como trabalho futuro a utilização do histograma de Euler.

O otimizador de consultas de um SBDE utiliza estruturas de dados para indexar os dados, além de algoritmos específicos para melhorar a eficiência das consultas, como por exemplo decidir em que ordem as operações de uma consulta devem ser executadas. Para tomar essa decisão, o otimizador enumera planos de execução alternativos e escolhe o mais eficiente dentre eles (KOSSMANN, 2000). Praticamente todos os otimizadores de consultas escolhem o melhor plano para uma consulta usando um modelo de custo que depende muito da estimativa precisa de seletividade (MARKL et al., 2004).

Em sistemas distribuídos, a estimativa de seletividade também é usada no escalonamento de partes da consulta, de modo que possa ser executada em vários servidores, dessa forma, o otimizador de consultas também deve especificar em qual nó do *cluster* cada fragmento de consulta será executado, de forma que um conjunto de tarefas possa ser escalonado de maneira moldável: ou seja, ajustando-se o número de processadores a serem usados para otimizar o tempo total (*makespan*) de computação (OLIVEIRA, 2017). No entanto, para determinar o arranjo mais eficiente entre máquinas e tarefas, o otimizador de consultas utiliza um modelo de predição que depende muito da estimativa de seletividade. A Figura 1 mostra uma situação em que o otimizador estima de maneira incorreta o custo de quatro tarefas (J_1, J_2, J_3 e J_4), consequentemente faz a alocação de maneira errada no *cluster*. Em (a) o otimizador de consultas estima de maneira errada o custo de cada tarefa. A alocação ótima para o custo real pode ser visto em (c), por ter realizado uma estimativa incorreta, realiza uma alocação de tarefas baseada em uma informação errada (b), o que resulta em um mau escalonamento, mostrado em (d).

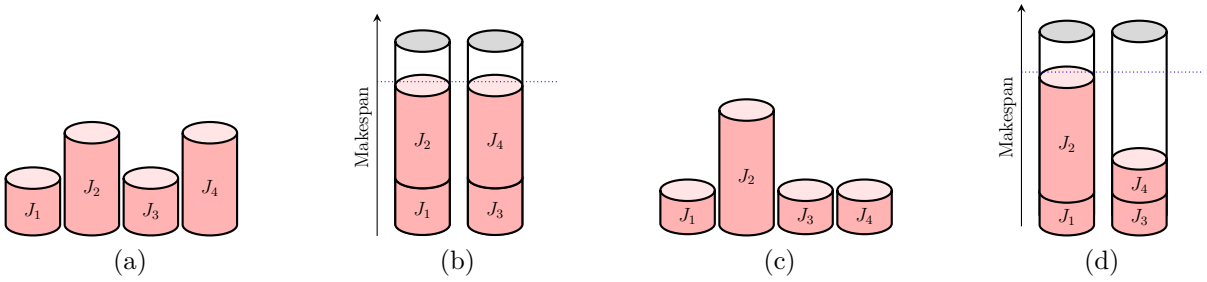


Figura 1 – Situação em que a estimativa imprecisa causa uma alocação de tarefas incorreta.

Todos os métodos propostos na literatura como histogramas, amostragem ou *wavelets* trabalham com aproximações do *dataset* real (SHEKELYAN; DIGNÖS; GAMPER, 2017). Consequentemente erros podem ocorrer por uma série de fatores, como por exem-

plo dados estatísticos imprecisos que representam os dados ou suposições erradas sobre o *dataset*. Portanto, o otimizador de consulta utiliza dados aproximados para realizar suas tarefas. Esse resultado, em conjunção com a crescente complexidade das consultas, demonstra a importância de técnicas mais precisas para o cálculo da estimativa de seletividade. Segundo Acharya, Poosala e Ramaswamy (1999) a utilização de histogramas para a estimativa da seletividade se tornou popular por ser de fácil construção, manutenção e possuir boa precisão em relação aos outros métodos propostos na literatura.

1.2 Objetivo do Trabalho

Este trabalho teve como objetivo avaliar, como o histograma de Euler se comporta em *datasets* reais com uma alto grau de heterogeneidade em relação ao tipo de objetos, como também o quão assertiva é a estimativa da seletividade da junção espacial calculada partição a partição obtida usando o histograma de Euler, comparada a mesma estimativa obtida com o histograma de grade que utiliza o método de sobreposição proporcional, proposto em (OLIVEIRA, 2017). Os objetivos específicos são:

1. avaliar a assertividade do histograma de Euler utilizando *datasets* reais, com objetos do tipo linha e polígono, pois até o momento só foi testado com *datasets* do tipo retângulo.

2. avaliar a assertividade do histograma de Euler no cálculo da seletividade global das consultas de junção espacial;

3. avaliar a assertividade do histograma de Euler no problema da alocação de tarefas, isto é, estimar a seletividade da junção espacial em cada partição e avaliar sua assertividade.

1.3 Contribuição do Trabalho

A principal contribuição deste trabalho é a avaliação do uso do histograma de Euler em um novo cenário, onde os *datasets* possuem grande heterogeneidade em relação aos tipos dos objetos, a utilização do método de enquadramento *clipping*. Além da avaliação, da estimativa de seletividade calculada partição a partição utilizando o histograma de Euler e comparando com o histograma de grade.

1.4 Organização da Monografia

O trabalho está dividido em seis capítulos, descritos resumidamente a seguir: O Capítulo 2 é onde será feita a contextualização dos conceitos e embasamento teórico

do trabalho, apresentando os tipos de histogramas encontrados na literatura que serão utilizados nos experimentos. No Capítulo 3 serão apresentados os trabalhos relacionados a essa pesquisa que motivaram a produção da mesma. O Capítulo 4 apresenta os algoritmos desenvolvidos neste trabalho, e também mostra como foi validada estas implementações. Em seguida, o Capítulo 5 apresenta a metodologia, com a classificação e o detalhamento da forma de como foi conduzida a pesquisa, além disso mostra todos *datasets* utilizados, experimentos realizados durante o processo de avaliação e as métricas utilizadas para verificar a assertividade dos histogramas. Por fim, o Capítulo 6 apresenta a conclusão dos resultados desta pesquisa e trabalhos futuros a serem realizados. Além disso, este no A foram colocados alguns gráficos dos resultados.

2 REFERENCIAL TEÓRICO

Neste capítulo serão apresentados alguns conceitos introdutórios, a fim de que os objetivos deste projeto sejam melhor entendidos. Sendo assim, para a compreensão desta monografia, apresenta-se neste capítulo as definições usadas ao longo do texto. A seção 2.1 traz a definição de dados espaciais e como é a representação destes dados. A seção 2.2 apresenta a definição dos Sistemas de Informação Geográfica. A seção 2.3 traz a definição da consulta do tipo junção espacial. A seção seção 2.4 apresenta o processamento distribuído de consultas espaciais. A seção 2.5 contextualiza a importância da estimativa de seletividade na otimização de consultas espaciais. A seção 2.6 apresenta a definição de histogramas espaciais. Igualmente a seção 2.7 os conceitos relativos ao histograma de Euler. Por fim a seção 2.8 apresenta a suite DGEO ao qual este trabalho está inserido.

2.1 Dados Espaciais

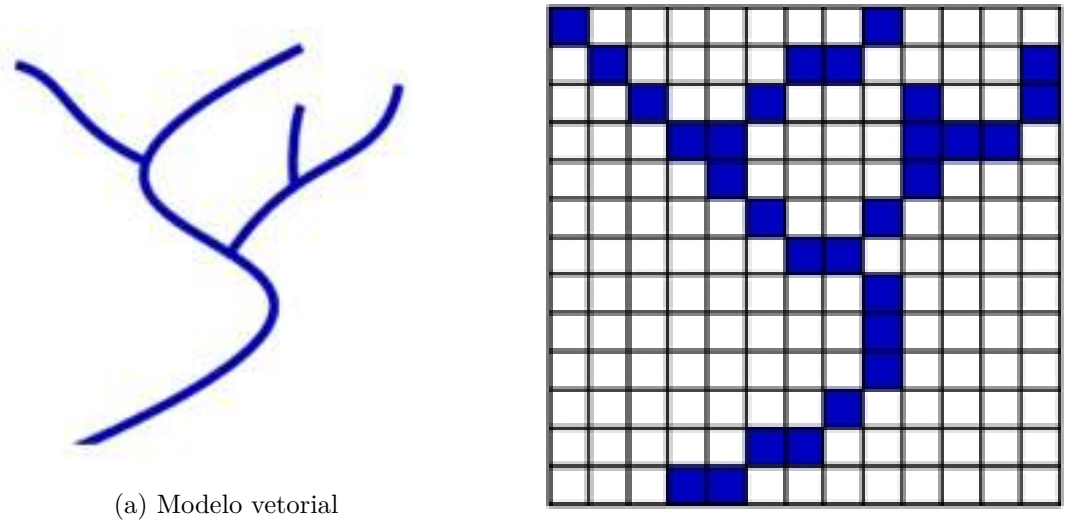
Dados espaciais são os objetos geográficos do mundo real como ruas, edifícios, lagos e países, e suas respectivas localizações. Além disso, cada um desses objetos também possui certos traços de interesse, ou atributos como nome, tamanho, profundidade, população, etc. (HUISMAN; BY, 2009). Por exemplo uma das casas de uma cidade são elementos do espaço geográfico que possuem atributos: proprietário, localização, valor, entre outros. Os atributos estão armazenados num sistema gerenciador de banco de dados.

Os dados espaciais podem ser adquiridos através de imagens, mapas, planos de informação, ou qualquer outra técnica que faz, de alguma forma, referência ao mundo real. A representação dos dados espaciais na forma digital pode ser em estruturas matriciais ou vetoriais (FITZ, 2018).

A estrutura vetorial, como ilustra a Figura 2a, é composta por três primitivas gráficas (pontos, linhas e polígonos) e utiliza um sistema de coordenadas para a sua representação. Os pontos são representados por apenas um par de coordenadas, ao passo que linhas e polígonos são representados por um conjunto de pares de coordenadas (FITZ, 2018).

Os dados espaciais também podem ser armazenados em uma estrutura matricial, ilustrado na Figura 2b, ou em grade (*raster structure*). Essa estrutura de dados é representada por uma matriz com n linhas e m colunas, na qual cada célula, denominada *pixel*, apresenta um valor z que pode indicar, por exemplo uma cor ou um tom de cinza a ele atribuído. Produtos advindos do sensoriamento remoto, como imagens de satélite e fotografias digitais, além de mapas digitalizados, utilizam essa forma de armazenamento

(FITZ, 2018).



(a) Modelo vetorial

(b) Modelo matricial

Figura 2 – Exemplo das representações espaciais. .

2.2 Sistema de Informações Geográficas

Um Sistema de Informações Geográficas (SIG) oferece um conjunto de métodos para tratar de dados espaciais ou georreferenciados incluindo captura, preparação, gerenciamento, armazenamento, manutenção, manipulação, análise e por fim apresentação desses dados. Entende-se como dado espacial toda informação com uma marcação de posicionamento associada. Esta marcação pode-se referir ao planeta Terra (dado geoespacial) ou a um outro referencial, por exemplo, o cosmos, componentes eletrônicos em uma placa de circuito ou dados sobre o corpo humano capturado por imagens médicas. Em cada uma dessas referências espaciais, uma projeção específica é usada para nivelar os dados em uma representação planar. Essa projeção é chamada de Sistemas de Referência das Coordenadas (SRC). Em suma, os dados espaciais são todo e qualquer dado que diz respeito ao espaço(HUISMAN; BY, 2009).

Um SBDE (Sistema de Banco de Dados Espacial) é um tipo de SIG que incorpora funcionalidades que fornecem suporte para bancos de dados que controlam objetos em um espaço multidimensional. Por exemplo, um SBDE pode ser utilizado para gerenciar bases de dados cartográficas que armazenam mapas, os quais incluem descrições espacial multidimensional de seus objetos, rios, cidades, estradas, mares e assim por diante (ELMASRI; NAVATHE, 2010). Para realizar esse gerenciamento, um SBDE armazena os dados de forma estruturada e processam algoritmos espaciais, com o objetivo de dar respostas aos usuários sobre questões sobre os dados, chamados de consultas espaciais.

As consultas espaciais são muito importantes para se obter informação de um *dataset*. Como exemplo, um ambientalista pode conseguir por meio de uma consulta espacial em um mapa geoespacial a quantidade de vegetação próximo a um rio, afim de descobrir o índice de desmatamento do local. Existem cinco categorias principais de consultas (AJI, 2014): i) consultas de agregação de recursos (consultas não espaciais), por exemplo, consultas para encontrar valores médios de atributos ou distribuição de atributos; ii) consultas espaciais fundamentais, incluindo consultas baseadas em pontos, consultas de sobreposição e consultas de janela; iii) consultas espaciais complexas, incluindo consultas de junção espacial, multijunção espacial e vizinhos mais próximos (*K-nearest neighboars*); iv) consultas espaciais e de recursos integradas como, por exemplo, consultas de agregação de feições em certas regiões espaciais; e v) consultas de padrão espacial como, por exemplo, consultas na localização de regiões de alta densidade ou consultas para localizar padrões direcionais de objetos.

A Figura 3 ilusta três exemplos de possíveis consultas. Nas consultas de janela (figura mais a esquerda) inicialmente uma região retangular de um *dataset* é selecionada, essa região recebe o nome de janela de seleção. Dessa forma, dado um conjunto de objetos espaciais e uma janela de seleção, uma consulta de janela recupera todos os objetos que interceptam o retângulo referente a janela de seleção (ACHARYA; POOSALA; RAMASWAMY, 1999). Este trabalho tem como foco as consultas de do tipo junção espacial, não somente por ser uma das consultas que mais consomem tempo de processamento, como também por ser uma das mais utilizadas nos SIGs. Além disso este trabalho também está interessado nas consultas de janela, visto que uma junção espacial pode ser vista como uma coleção de consultas de janela.

Figura 3 – Exemplos de consultas espaciais.

A definição 2.1 apresenta a definição das consultas de janela em SBDE. Matematicamente uma consulta de janela é uma função quer toma como parâmetros um conjunto de objetos multidimensionais e um retângulo de seleção, e retorna um subconjunto de objetos multidimensionais que respeitam algum critério.

Definição 2.1 (Consulta de Janela) *Uma consulta de janela Q é aquela em que dado um dataset $A = \{a_1, a_2, \cdots, a_n\}$ de objetos multidimensionais e um retângulo de seleção $S = \{x_0, y_0, x_f, y_f\}$ sobre A retorna informações sobre esta região. Formalmente Q pode se definida como uma função*

$$Q : A \times S \to A' \tag{2.1}$$

de modo que $A' \subseteq A$, em outras palavras

$$Q(A, S) = \{a : a \subseteq A \wedge a\theta S\} \tag{2.2}$$

sendo θ um predicado espacial.

2.3 Junção Espacial

Uma junção espacial é um tipo de consulta que encontra objetos correlacionados em dois *datasets*, considerando um predicado θ (OLIVEIRA, 2017). Uma descrição formal da consulta junção espacial pode ser encontrada na Definição 2.2. Alguns exemplos de predicados espaciais são (MARK, 2003):

- *Intersecção*: frequentemente usado para identificar características comuns aos *datasets* alvos em relação a um local específico do *dataset* fonte.
- *contém*: Retornará características do *dataset* alvo que estão contidas no *dataset* fonte, mesmo se os limites se sobreporem.
- *é idêntico a*: Esse predicado retorna características do *dataset* alvo que estão na mesma posição geográfica do *dataset* fonte.

Definição 2.2 (θ-junção) *Sejam $A = \{a_1, a_2, \cdots, a_n\}$ e $B = \{b_1, b_2, \cdots, b_n\}$ dois datasets distintos de objetos multidimensionais, então se existem objetos $a \in A$ e $b \in B$ que satisfazem um predicado θ, é possível realizar uma θ-junção espacial entre A, B. Formalmente uma junção espacial pode ser definida como uma função*

$$A \bowtie B : A \times B \to R \tag{2.3}$$

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

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

Para diminuir o custo de CPU e E/S (Entrada e Saída), geralmente uma junção espacial é realizada em dois estágios: *filtro* e *refinamento*, ambos utilizam algoritmos de geometria computacional. No estágio de filtro, pela complexidade dos dados multidimensionais, cada objeto espacial em um *dataset* é aproximado pelo menor retângulo em que

ele está contido, esse retângulo é conhecido por MBRs (*Minimum Bounding Rectangle*) e ilustrado na Figura 4. No entanto, no estágio de filtro podem haver falsos positivos no cálculo da junção, por exemplo, MBRs que satisfizeram o predicado θ, mas não os objetos reais. Desta forma, o estágio de refinamento utiliza determinados algoritmos para remover falsos objetos que foram identificados no estágio de filtro. (JACOX; SAMET, 2007). Neste sentido, o estágio de refinamento é mais caro computacionalmente do que o estágio de filtro. Portanto, para aumentar a eficiência do processamento nas operações de junção, pode ser empregado a utilização de sistemas distribuídos (BRINKHOFF; KRIEGEL; SEEGER, 1996).

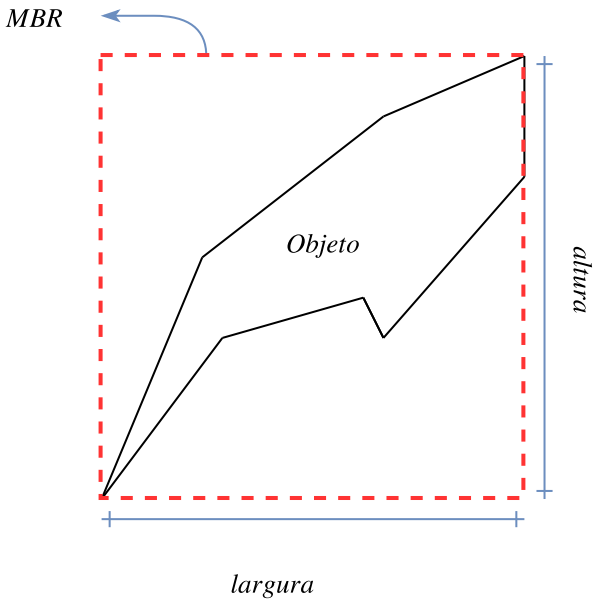


Figura 4 – Exemplo de MBR.

A Figura 5 ilustra a junção espacial entre o *dataset* de rios e o de avenidas de uma determinada cidade, sendo o predicato θ a intersecção. Neste cenário, foi utilizado um método de partição dos dados, cada partição é denominada *bucket* (balde) que armazena as informações do espaço geográfico que representa. Os *buckets* dos *datasets* são disjuntos, de modo que a junção espacial global é calculada como união de todas as junções locais, isto é, sejam $A_1, A_2, \ldots, A_n$ e $B_1, B_2, \ldots, B_n$ os *buckets* de dois *datasets* A e B respectivamente, então, $A \bowtie B = \bigcup_{i=1}^{n} (A_i \bowtie B_i)$.

2.4 Processamento Distribuído de Consultas Espaciais

Um sistema distribuído é um conjunto de computadores independentes que se aprensenta a seus usuários como um sistema único e coerente. Uma classe importante de sistemas distribuídos é utilizada para tarefas de computadores de alto desempenho. Um subgrupo dessa classe é a computação de *cluster*, ilustrado na Figura 6, onde o *hardware*

subjacente consiste em um conjunto de nós $n_1, n_2, n_3, \ldots, n_k$ (estações de trabalho ou PCs semelhantes), conectados por meio de uma rede local de alta velocidade, e cada nó executa o mesmo sistema operacional (SO) (TANENBAUM; STEEN, 2007).

Um SBDE centralizado manipula e processa as solicitações e envia os resultados de volta para cada solicitação. Essa abordagem facilita ao administrador gerenciar e proteger o banco de dados. Mas isso requer um servidor de alto desempenho em termos de capacidade de processamento e capacidade de armazenamento, especialmente com um grande número de solicitações. Para melhorar o processamento do banco de dados, existe outra abordagem, o processamento distribuído em *cluster*. Nesse método, haverá mais de um servidor de banco de dados (nós) no sistema, mas os dados em si estarão em um armazenamento centralizado e compartilhado (ÖZSU; VALDURIEZ, 2011).

Özsu e Valduriez (2011) definem um sistema de banco de dados distribuído (SBDD) como uma coleção de bancos de dados múltiplos e logicamente inter-relacionados distribuídos por uma rede de computadores. Um sistema de gerenciamento de banco de dados distribuído (SGBD distribuído) é então definido como o sistema de *software* que permite o gerenciamento do banco de dados distribuído e torna a distribuição transparente para os usuários.

O custo do processamento de uma junção espacial pode ser muito alto devido aos grandes tamanhos e à complexidade dos objetos espaciais envolvidos. Portanto para melhorar o desempenho é empregado frequentemente a utilização de sistemas distribuídos. Contudo, em um ambiente distribuído, as junções espaciais para dois *datasets* que residem em locais separados geograficamente são dispendiosas em termos de custo de transmissão e custo de computação (JACOX; SAMET, 2007). Nesse sentido, técnicas são necessárias para otimizar o processamento das consultas.

Muitos algoritmos para processar consultas de junção espacial em ambientes com-

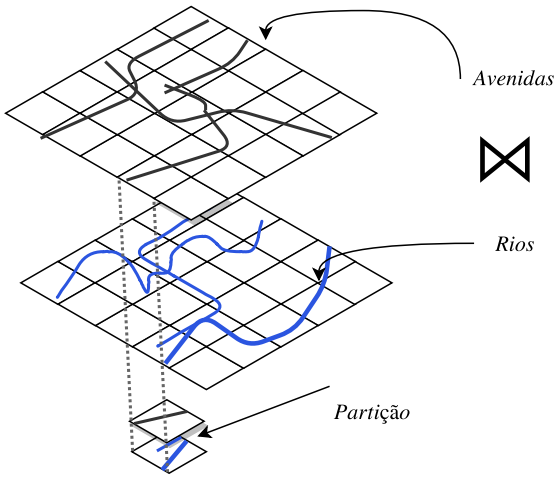


Figura 5 – Exemplo de Junção Espacial entre dois datasets.

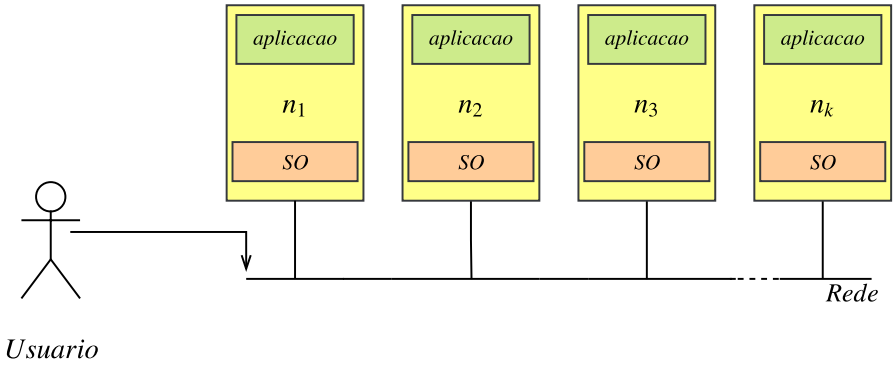


Figura 6 – Exemplo da organização de um *cluster*.

putacionais paralelos ou distribuídos foram propostos. A maioria desses algoritmos usa uma estratégia de particionamento de dados (*declustering*) para dividir objetos de um conjunto de dados em grupos, chamados de partições de dados ou células. Antes ou durante a execução da junção, uma rotina atribui um conjunto de partições a um determinado servidor ou processador que executará a consulta em conjunto. Existem duas categorias principais de métodos de particionamento (OLIVEIRA, 2017):

1. *Particionamento de espaço disjunto*. Esse método utiliza uma grade de células disjuntas para dividir a extensão espacial do *dataset*. Cada célula da grade agrupa os objetos espaciais de acordo com sua interseção com as células. Esse particionamento replica objetos que se cruzam em mais de uma célula;

2. *Particionamento de espaço não disjunto*. as partições podem se sobrepor umas às outras para acomodar a extensão dos objetos que as interceptam e não requerem replicação de objetos. Um exemplo desse tipo de particionamento é o conjunto de MBRs em um determinado nível de um índice R-Tree.

A Figura 7 ilustra o processamento distribuído de uma junção espacial. Primeiramente os *datasets* A e B envolvidos na consulta são particionados em $A_1, A_2, \ldots, A_n$ e $B_1, B_2, \ldots, B_n$ de acordo com um método de partição de dados, em seguida é determinado a junção local de cada partição (tarefa) que será posteriormente atribuída para um nó do *cluster* executar. Neste sentido, para que a execução seja balanceada é necessário um parâmetro que determine quais conjuntos de partições serão executadas por qual nó, de forma a encontrar o melhor arranjo entre nós e tarefas. Na seção seguinte será discutido este parâmetro, bem como métodos para calculá-lo.

2.5 Seletividade das Consultas Espaciais

A cardinalidade de um *dataset*, ou de partes dele, é a chave para prever o tamanho da saída da junção, também conhecida como seletividade da junção ou cardinalidade

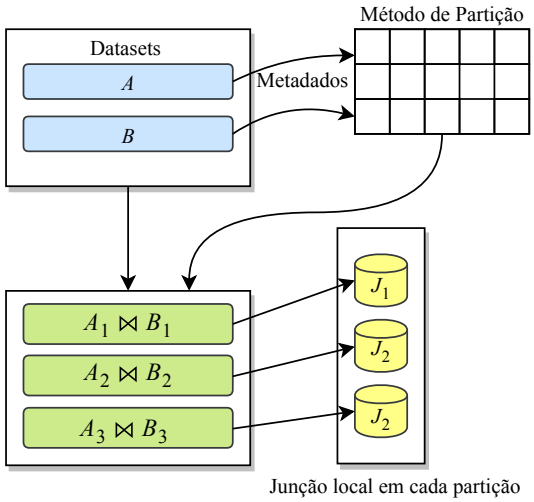


Figura 7 – Diagrama da execução distribuída de uma junção espacial típica.

da junção (OLIVEIRA, 2017). A estimativa de seletividade é usada em um otimizador de consulta para escolher um bom plano de execução para uma determinada consulta. As estimativas de seletividade de junções espaciais podem ser usadas como respostas a consultas especializadas de usuários que buscam valores aproximados. Por exemplo, encontrar o número aproximado de pontes em uma determinada extensão espacial pode simplesmente ser satisfeito fazendo uma estimativa de seletividade de junção entre os conjuntos de dados de ruas e rios para essa extensão (assim o usuário não precisa executar a junção real, economizando tempo)(AN; YANG; SIVASUBRAMANIAM, 2001).

A seletividade pode ser calculada em um fragmento do *dataset* como, por exemplo, o cálculo em apenas uma partição. Portanto, é utilizada como parâmetro para distribuição de tarefas em um *cluster*, posto que partições com seletividade alta seriam mais dispendiosas em termos de tempo de processamento, pois os algoritmos de verificação de predicado da geometria computacional, como interseções, contenções e sobreposições, muitas vezes demoram para processar devido a grande quantidade de objetos (OLIVEIRA, 2017).

2.6 Histogramas Espaciais

Um histograma é uma estrutura de dados que divide o espaço geográfico do *dataset* em células e aloca *buckets* para cada célula. Os *buckets* armazenam dados estatísticos sobre os objetos contidos no fragmento de espaço que esse *bucket* representa, por exemplo a quantidade de objetos. Uma das aplicações dos histogramas espaciais é no cálculo da estimativa da seletividade. Segundo Acharya, Poosala e Ramaswamy (1999) o uso de histograma se popularizou, pois é de fácil construção e consome pouca memória.

A Figura 8 ilustra a representação de um histograma espacial sobre um *dataset*. O retângulo superior da figura representa o *dataset* contendo os objetos, e abaixo dele a

estrutura do histograma, com vários *buckets* de forma a particionar o espaço de dados.

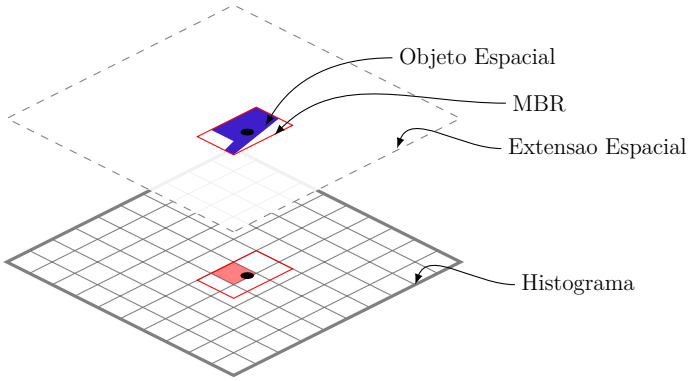


Figura 8 – Exemplo de histograma espacial sobre um *dataset* (OLIVEIRA, 2017).

Muitos histogramas foram propostos no literatura. Acharya, Poosala e Ramaswamy (1999) desenvolveram o histograma *minskew* que tem por objetivo minimizar a variação da densidade no *dataset*. Para isso a técnica realiza sucessivas divisões binárias dos *buckets* com alta densidade, dividindo-os em dois sub-*buckets*, fazendo com que a densidade dos *buckets* fique o mais uniforme possível.

Outro histograma proposto na literatura e muito utilizado, por ser de fácil construção e manutenção, é o histograma de grade que divide o espaço células do mesmo tamanho (AN; YANG; SIVASUBRAMANIAM, 2001). Oliveira (2017) melhorou o histograma de grade utilizando não somente um método para definir o número de células em um histograma de grade, com base nos metadados do *dataset*, mas também um método de enquadramento dos objetos mais eficiente do que o MBR, o que melhorou bastante a estimativa de seletividade das consultas.

No entanto os histogramas supracitados possuem o problema da contagem múltipla de objetos, isto é, objetos grandes que sobrepõe espacialmente mais de um *bucket* são contados múltiplas vezes, o que resulta em erros na estimativa da seletividade, objeto de estudo deste projeto. Na próxima seção, será discutido uma técnica de histograma menos suscetível a este problema em específico, proposta por Sun, Agrawal e El Abbadi (2002).

2.7 Histograma de Euler

O histograma de Euler é projetado para resolver o problema da múltipla contagem de objetos. Comparado ao histograma de grade, um histograma de Euler aloca *buckets* não apenas para células da grade, mas também para cada um dos lados e cantos das células da grade. Sun et al. (2006) apresentaram um algoritmo para estimar a seletividade de consultas de janela usando histogramas de Euler. A base matemática do algoritmo é

baseada na fórmula de Euler da teoria dos grafos. A Figura 9 ilustra a comparação entre o histograma de Euler, e histograma de grade convencional.

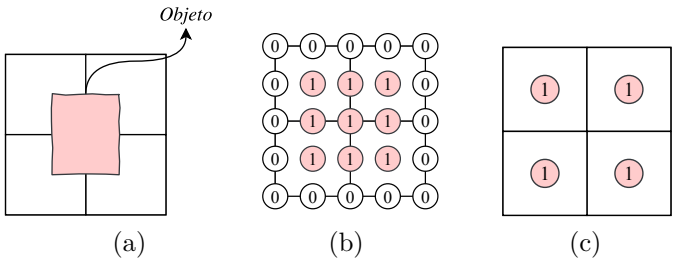


Figura 9 – Em 9a é possível ver um *dataset* trivial, isto é com apenas um objeto. A Figura 9b mostra o histograma de Euler sobre o dataset 9a, as circunferências representam os *buckets* e os números dentro delas a quantidade de objetos naquele fragmento do *dataset*. Em 9c é possível ver como o histograma de grade aloca *buckets* somente para cada célula.

Na Figura 9a é possível ver um objeto e uma grade, a Figura 9c mostra como um histograma de grade convencional aloca *buckets* somente para cada célula, enquanto que o histograma de Euler aloca não somente para a célula, como também para seus lados e cantos.

2.7.1 Teoria dos Grafos e Fórmula de Euler

Um grafo é uma estrutura de abstração bastante útil na representação e solução de diversos tipos de problema. Matemáticamente um grafo formaliza relações de independência existentes entre os elementos de um conjunto. Um grafo representa um conjunto de elementos denominados vértices e suas relações de interdependência ou arestas. Denominando por V o conjunto de vértices da estrutura e por E o conjunto das arestas ou ligações entre os vértices, um grafo pode ser representado por $G = (V, E)$.

Um grafo é conexo se, para qualquer par $\{v, w\}$ de seus vértices, existe um caminho com extremos v e w. Por sua vez, um grafo G é dito planar se seus vértices e arestas podem ser imersos em $\mathbb{R}^2$ tal que suas arestas não se cortem/cruzem. Em outras palavras, G é planar se admite uma representação no plano de modo que nela não exista cruzamento de arestas (GOLDBARG; GOLDBARG, 2012).

A imersão de um grafo planar em um plano divide o mesmo em regiões, estas regiões são chamadas de faces. É preciso ressaltar ainda a existência de uma face que não está limitada por nenhuma aresta, denominada face exterior. A Figura 10 ilustra um grafo planar com as faces f_1, f_2, f_3 e a face exterior f_4. Neste sentido, um vértice ou aresta é considerado exterior se está nos limites do grafo, e interior se não está. Por exemplo, na Figura 10 o vértice v_1 é o único vértice interior do grafo, bem como a aresta v_1v_2 que é a única aresta interior.

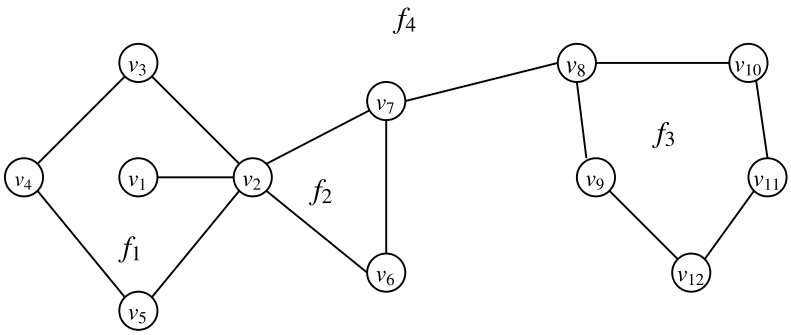


Figura 10 – Grafo planar com 4 faces.

A fórmula de Euler (HARARY, 1969), que está ilustrada na figura 11a, é um resultado importante na teoria dos grafos, que afirma:

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

$$V - E + F = 2 \tag{2.4}$$

Beigel e Tanin (1998) provaram um corolário da fórmula de Euler. Para o propósito deste trabalho, tendo em vista que um histograma trabalha com dados em duas dimensões, será apresentado apenas a versão bidimensional deste corolário.

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

$$V_i - E_i + F_i = 1 \tag{2.5}$$

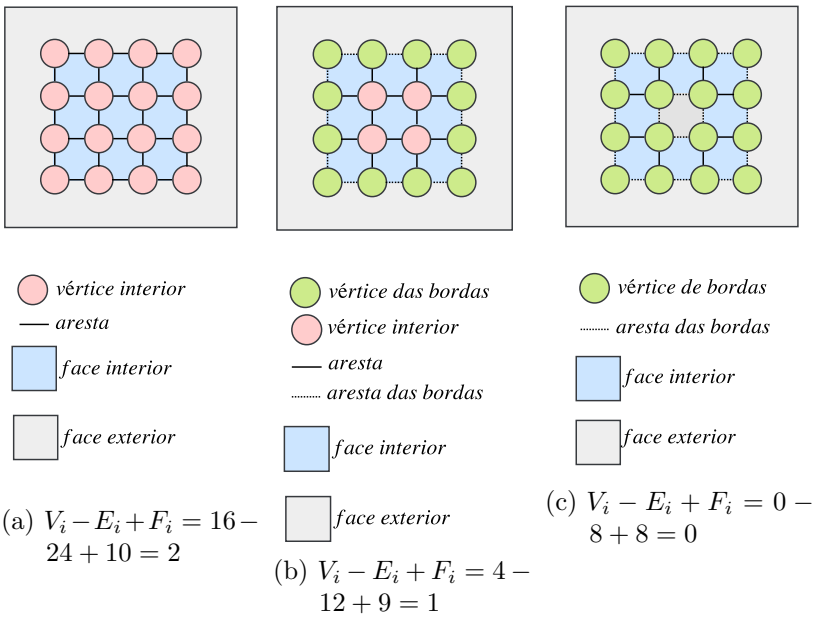
Um exemplo ilustrado do Corolário 2.1 é mostrado na Figura 11b onde é utilizado a mesma grade 3×3 da Figura 9b. Depois de remover a face exterior, tem-se agora 4 vértices interiores, 12 arestas interiores e 9 faces interiores.

Sun et al. (2006) estenderam o corolário de Beigel e Tanin (1998) para manipular grafos com mais de uma face exterior. A Figura 11c mostra um grafo com duas faces externas, correspondendo a uma região com um "buraco". Comparando a Figura 11b e 11c, é possível ver que a face no meio é agora uma face externa. Consequentemente, os vértices e arestas ao redor desta face externa não são mais vértices ou arestas interiores. Depois de remover as duas faces externas e os dois limites, tem-se então 0 vértices internos, 8 arestas internas e 8 faces internas, então $V_i - E_i + F_i = 0$. Para grafos com k faces externas, (SUN et al., 2006) enunciaram o seguinte corolário.

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

$$V_i - E_i + F_i = 2 - k \tag{2.6}$$



(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 11 – Exemplo ilustrado do teorema 2.1 e dos corolários 2.1 e 2.2. Modificado de Sun et al. (2006).

Um histograma de Euler pode ser visto como um grafo, na medida que os *buckets* podem ser vértices, arestas ou faces. A Figura 12 ilustra um grafo correspondente a um histograma de Euler, as arestas correspondem aos lados das células, os vértices os cantos, e as faces a célula propriamente dita. Portanto todos os corolários e teoremas apresentados se aplicam aos histogramas de Euler. Nas subseções seguintes será apresentado a construção do histograma de Euler, bem como o algoritmo para o cálculo da seletividade das consultas de junção espacial.

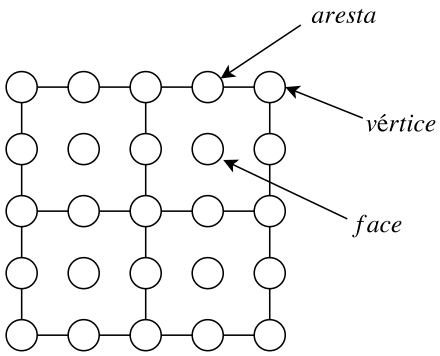


Figura 12 – Um grafo que representa um histograma de Euler.

2.7.2 Construção do Histograma

Dada uma grade no espaço do *dataset*, uma maneira direta de construir um histograma é permitir que cada *bucket* do histograma corresponda a uma célula da grade e, se um objeto sobrepõe uma célula, incrementar o valor do *bucket* correspondente em 1. No entanto, o histograma não consegue distinguir entre um objeto grande que sobrepõe várias células e vários pequenos objetos, que em conjunto sobrepõe várias células. Para resolver este problema, Sun et al. (2006) propuseram um novo tipo de histograma que pode ser construído da seguinte forma:

- Dado uma grade $n_1 \times n_2$ no $\mathbb{R}^2$, alocar $(2n_1 - 1)(2n_2 - 1)$ *buckets* para o histograma $\mathscr{H}$. Um *bucket* de $\mathscr{H}$ 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 que correspondem a arestas.

Dessa forma, dada uma consulta de janela S em um *dataset* A, a seletividade $\mathscr{S}$ dessa consulta pode ser calculada com um histograma de Euler $\mathscr{H}_A$ de acordo com a Equação 2.7, sendo B o conjunto de *buckets* do histograma e $b_{i,j}(S)$ os *buckets* contidos na janela de seleção S (SUN; AGRAWAL; El Abbadi, 2002).

$$\mathscr{S}(S) = \sum_{b_{i,j} \in B} b_{i,j}(S) \tag{2.7}$$

Onde $b_{i,j}$ é o *bucket* da linha i coluna j de $\mathscr{H}$, dentro da consulta de janela S. A Figura 13 apresenta um *dataset* com 3 objetos e uma consulta de janela com os limites (x_1, y_1, x_2, y_2), então pela Equação 2.7, a seletividade da consulta, é a a soma dos valores de todos os *buckets* dentro da janela de seleção. Para maior legibilidade a janela de seleção está representada pelos *buckets* de cor verde, ou seja, $2 - 2 + 2 - 2 + 2 - 2 + 2 - 2 + 3 = 3$. Portanto a estimativa foi calculada sem nenhum erro. Na próxima subseção será demonstrado como o histograma de Euler pode ser utilizado para o cálculo da seletividade nas consultas de junção espacial.

2.7.3 Junção Espacial com Histograma de Euler

A junção espacial pode ser realizada utilizando um histograma de Euler da seguinte forma:

Dado dois *datasets* A, B e uma grade no espaço dos dados, é possível construir um histograma de Euler para cada *dataset*, chamados de $\mathscr{H}_A$ e $\mathscr{H}_B$. Seja $b^A_{i,j}$ e $b^B_{i,j}$ os *buckets* em $\mathscr{H}_A$

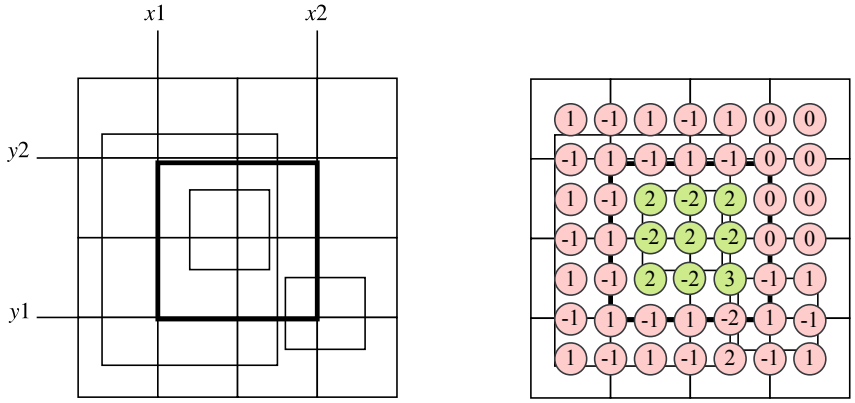


Figura 13 – Histograma de Euler na consulta de janela.

e $\mathscr{H}_B$, respectivamente. O significado do subscrito $b_{i,j}$ é: i é a linha de b no histograma, j é a coluna. Assim para uma consulta Q envolvendo dois *datasets* A e B com uma janela de seleção S, a seletividade da consulta pode ser calculada como mostra a Equação 2.8 (SUN; AGRAWAL; El Abbadi, 2002).

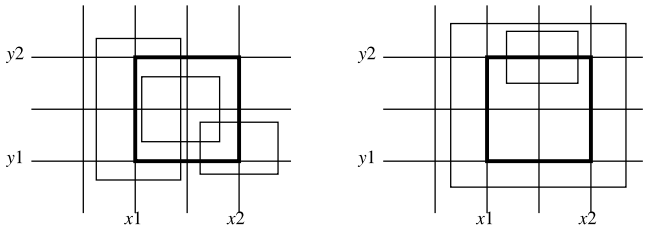
$$\mathscr{S}(Q) = \sum_{i,j} b^A_{i,j}(S) \times b^B_{i,j}(S) \tag{2.8}$$

onde $b^A_{i,j} \in B_A$ e $b^B_{i,j} \in B_B$ são *buckets* de $\mathscr{H}_A$ e $\mathscr{H}_B$ respectivamente, que estão dentro da janela de seleção S. A Figura 14a ilustra dois *datasets* A e B. O *dataset* A consiste de três objetos, e está mais a esquerda, já o *dataset* B consiste de dois e está mais a direita. A janela de seleção está em (x_1, x_2, y_1, y_2). Os histogramas de Euler para A e B são mostrados na figura 14b. Para maior legibilidade foram removidos todos os *buckets* exceto aqueles dentro da janela de seleção. Pela equação 2.8, a seletividade desta consulta é

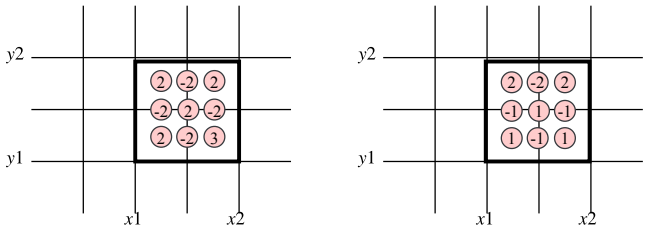
$$\overbrace{2 \times 1}^{\text{vértices}} - \underbrace{(2 \times 2 + 2 \times 1 + 2 \times 1 + 2 \times 1)}_{\text{arestas}} + \overbrace{(2 \times 2 + 2 \times 2 + 2 \times 1 + 3 \times 1)}^{\text{faces}} = 5$$

2.7.4 Histograma de Euler Generalizado

O Histograma de Euler estima corretamente a seletividade quando a janela de seleção da consulta ou os objetos do *dataset* se alinham com a grade do histograma. Assim, como nos exemplos da Figura 14 e 13, onde a janela de seleção se alinha com as grades. No entanto quando a janela de seleção ou objetos não se alinham com as grades do histograma podem haver erros no cálculo da seletividade. Dessa forma, foi proposto por Sun, Agrawal e El Abbadi (2002) o histograma de Euler generalizado para lidar com



(a) *Datasets* A e B e janela de seleção



(b) Histograma de Euler para os *datasets* A e B

Figura 14 – Estimativa de seletividade de junção espacial com histograma de Euler. Modificado de Sun et al. (2006)

esse problema. A equação 2.8 pode ser reescrita da seguinte forma:

$$\mathscr{S}(Q) = \sum_{i,j} b^A_{i,j}(S) \times b^B_{i,j}(S) \times p_{i,j} \tag{2.9}$$

onde $b^A_{i,j}(S) \times b^B_{i,j}(S)$ são as estimativas do número de regiões que se intersectam. Levando em consideração que os objetos se alinhem com a grade, se dois objetos intersectam a mesma face, logo os objetos se intersectam. Então se há h^A objetos do *dataset* A que intersectam a face (i, j), e h^B objetos do *dataset* B que intersecta a mesma face, então o número de regiões de intersecção que intersectam a face (i, j) é exatamenet $h^A \times h^B$. Contudo, com objetos que não se alinham à grade do histograma, o número de de regiões de intersecção pode ser menor do que $h^A \times h^B$.

Neste contexto, $p_{i,j,k}$ é a probabilidade de um conjunto de objetos inersectar outro conjunto de objetos dentro de um vértice, aresta ou face. Para vértices p é sempre 1; para faces e arestas p está entre 0 e 1, dependendo dos tamanhos e distribuições espaciais dos objetos dentro da face ou aresta (SUN; AGRAWAL; El Abbadi, 2002).

É importante ressaltar que a Equação 2.9 não especifica como $p_{i,j}$ deve ser calculado. Em vez disso, serve como uma estrutura onde diferentes modelos probabilísticos podem ser aplicados. Para determinados *datasets*, alguns modelos probabilísticos podem capturar a distribuição de dados melhor do que os outros, o que resultará em uma estimativa de seletividade mais precisa.

Sun, Agrawal e El Abbadi (2002) propuseram um modelo probabilístico para determinar a probabilidade p de arestas e faces que pode ser resumido da seguinte forma,

quando $b_{i,j}$ é uma aresta:

$$p_{i,j} = min(1, \overline{e_1} + \overline{e_2})$$

onde $\overline{e}$ é a projeção média dos objetos nas arestas dos dois *datasets*. E para faces, o fator de intersecção $p_{i,j}$ é

$$p_{i,j} \begin{cases} n_1 \times n_2, & \text{se } \overline{h_1} + \overline{h_2} \text{ e } w_1 + w_2 \geq 1 \\ \overline{a_1} + \overline{a_2} + \overline{h_1} \times \overline{w_2} + \overline{h_2} \times \overline{w_1}, & \text{caso contrário.} \end{cases}$$

onde $\overline{h}, \overline{w}$ e $\overline{a}$ são a altura, a largura e a área médias das regiões de interseção entre os objetos e a célula de cada *dataset*.

2.8 DGEO

Esta seção descreve a aplicação de processamento distribuído de dados espaciais, chamado DGEO, que será usada como aplicação real nos experimentos.

Esta suíte de aplicações foi desenvolvida em um projeto de pesquisa na Universidade Federal de Goiás, em linguagem de programação C e Go, com *threads* nativas para explorar múltiplos níveis de paralelismo e comunicação distribuída através de protocolos próprios, usando serialização Gob, sobre *sockets* TCP. A aplicação processa consulta de junção espacial, de forma paralela e distribuída. Todos os métodos e algoritmos implementados utilizam a biblioteca GEOS (*Geometry Engine - Open Source*) para o processamento do predicado das consultas.

Esta suíte será utilizada por ser uma suíte de pesquisa, a qual já trás implementada além das técnicas de particionamento de dados, histogramas espaciais, estruturas de dados, e rotinas auxiliares que facilitam o desenvolvimento, como por exemplo: processamento de *datasets*, estrutura de dados R-Tree para realizar as consultas de forma a obter o valor real da seletividade.

3 TRABALHOS RELACIONADOS

3.1 Introdução

Com o objetivo de entender e levantar diversos problemas no cálculo da estimativa de seletividade da junção espacial, foram analisadas diversas iniciativas e propostas de métodos e algoritmos. Neste sentido, os trabalhos relacionados foram divididos em três categorias: (i) trabalhos que estudam a estimativa de seletividade na junção espacial (ii) trabalhos que estudam a junção espacial distribuída (iii) trabalhos que utilizam o histograma de Euler no cálculo da seletividade de consultas espaciais. A fim de sedimentar e explicar a metodologia de busca e análise, este capítulo foi dividido em quatro seções. A seção 3.2 apresenta os critérios de busca bem como a metodologia de análise adotada. A seção 3.3 por sua vez apresenta os principais trabalhos relacionados. E por fim a seção 3.4 apresenta um resumo comparativo entre os trabalhos elencados na seção 3.3.

3.2 Metodologia de Análise

Os trabalhos apresentados foram filtrados por meio do sistema de busca de periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que possui bases referenciais importantes para este trabalho. As bases utilizadas forma ACM Digital Library, ACM Computing Reviews, PubMed Central (PMC), Scopus, IEEE Explore. O método utilizado para a busca dos trabalhos apresentados consistiu em aplicar uma *string* de busca com as palavras chaves do presente trabalho nas bases referenciadas supracitadas. A fim de obter publicações que não foram recuperadas pelos mecanismos de busca e que sejam importantes para este trabalho, foi feita uma análise manual das referências dos artigos retornados por essas fontes encontradas previamente.

Foi considerada uma metodologia para a análise dos trabalhos relacionados. Neste sentido, os trabalhos analisados deveriam pertencer ao tema desta pesquisa ou deveriam utilizar outros métodos de resolução para o mesmo objeto de estudo. Os seguintes critérios foram adotados para comparação destes trabalhos:

3.2.1 Estimativa de Seletividade na Junção Espacial(C1)

O primeiro critério adotado levou em consideração trabalhos que propuseram ou utilizaram técnicas para o cálculo da estimativa de seletividade na junção espacial.

3.2.2 Junção Espacial Distribuída (C2)

No que se refere ao segundo critério adotado neste trabalho foi levado em consideração pesquisas que estudaram a junção espacial em um sistema distribuído.

3.2.3 Cálculo da Estimativa de Seletividade em SBDE (C3)

O terceiro critério se refere à trabalhos que estudaram e propuseram técnicas para o cálculo da estimativa de seletividade nas consultas (não somente junção espacial) em SBDEs.

3.2.4 Estimativa de Seletividade utilizando Histograma de Euler(C4)

Relativo ao quarto critério adotado, foram considerados trabalhos que estudaram e utilizaram o histograma de Euler no cálculo da estimativa de seletividade, seja em consultas de janela, seja na junção espacial.

3.3 Trabalhos analisados

Com base nos critérios apresentados na seção anterior, a busca de trabalhos nas respectivas bases citadas resultaram em número superior a dezenas de milhares de trabalhos. No entanto, foi possível perceber a distância destes resultados com o propósito central desta pesquisa. Por meio de filtros, foco em referências primárias em relação ao objeto de estudo e representatividade dos trabalhos, cinco trabalhos foram escolhidos.

3.3.1 Selectivity Estimation in Spatial Databases (T1)

Acharya, Poosala e Ramaswamy (1999) estudaram a estimativa de seletividade em SBDEs, em particular apresentaram várias novos métodos de agrupamento para aproximar dados espaciais. Além disso propuseram novas técnicas baseadas nas noções de densidade espacial.

Alguns *datasets* possuem uma densidade de objetos muito peculiar, ou seja, em algumas áreas possuem muitos objetos e em outras possuem poucos. Assim, técnicas de estimativa de seletividade que supõem uniformidade na distribuição dos objetos no domínio vão sofrer prejuízos no resultado. A Figura 15 ilustra a distribuição dos objetos em um dado domínio, é possível observar que neste *dataset* os objetos estão localizados majoritariamente nos cantos. Consequentemente supor uniformidade em *datasets* como estes resultam em erros que impactam na estimativa de seletividade.

Acharya, Poosala e Ramaswamy (1999) propuseram uma método denominado *Min-Skew* que faz sucessivas divisões binárias no espaço de dados de modo que a grade resul-

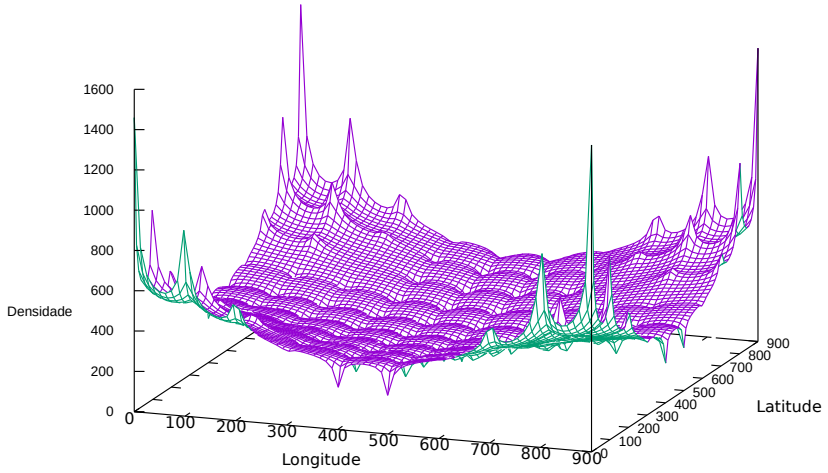


Figura 15 – Exemplo de densidade em um dataset.

tante seja proporcional à densidade no *dataset*. Diferentemente de outros métodos como por exemplo, o histograma de grade que divide o espaço em células de tamanhos iguais e supõe uniformidade em cada uma delas.

3.3.2 On Spatial Joins in MapReduce (T2)

Sabek e Mokbel (2017) em seu trabalho fornece a primeira tentativa de um otimizador de consulta completo para algoritmos de junção espacial baseados em MapReduce. O otimizador desenvolve sua própria taxonomia que abrange quase todas as maneiras possíveis de fazer uma junção espacial para quaisquer dois conjuntos de dados de entrada.

O otimizador funciona de duas formas: baseado em custo e baseado em regras. Com dois conjuntos de dados de entrada, o otimizador de consulta baseado em custo avalia os custos de todas as opções possíveis na taxonomia desenvolvida e seleciona aquele com o menor custo. O otimizador de consultas baseado em regras abstrai os modelos de custo desenvolvidos do otimizador baseado em custo em um conjunto de regras heurísticas simples e fáceis de verificar. Em seguida, aplica suas regras para selecionar a opção de menor custo. Ambos os otimizadores de consulta são implantados e avaliados experimentalmente dentro de um sistema de dados espaciais baseado em MapReduce de código aberto amplamente utilizado (ELDAWY; MOKBEL, 2015).

No que se refere ao modelo de custo das junções espaciais adotados no referido trabalho é importante notar que utiliza a estimativa de seletividade como parâmetro, e calcula esta estimativa utilizando um método proposto a mais de duas décadas que utiliza R-Trees.

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

Sun, Agrawal e El Abbadi (2002) propuseram a utilização do histograma de Euler no cálculo da estimativa de seletividade em consultas de junção espacial. Além disso, mostraram que para dois *datasets* com grades do histograma alinhadas e para objetos que também se alinham com a grade de seu respectivo histograma, a seletividade neste caso é calculada sem erro. No entanto em bancos de dados reais, que manipula *datasets* com objetos reais, as grades dos histogramas criados para dois *datasets* distintos provavelmente não irão se alinhar, visto que possuem dimensões espaciais diferentes. Assim os autores, tentaram desenvolver um método que fizesse com que os objetos dos *datasets* se alinhassem, assim propuseram a Suposição de Quantização em que os objetos não alinhados com a grade passam a se alinhar depois de efetuada uma transformação. No entanto esta transformação aumentaria o tempo de processamento devido ao tamanho dos objetos.

Para que a utilização do Histograma de Euler pudesse ser feita em *datasets* reais foi introduzido na equação da estimativa de seletividade um fator que determina, dado uma face, aresta ou vértice de dois histogramas que se intersectam, a probabilidade dos objetos desta face, aresta ou vértice se intersectar.

3.3.4 Multiway Spatial Join (T4)

Mamoulis e Papadias (2001) estudaram extensivamente a junção espacial e a multijunção espacial. Neste trabalho foram propostos algoritmos, métodos de partição dos dados, técnicas para enumerar um conjunto de planos de execução, para estimar seus custos e selecionar um bom plano. Além disso propuseram uma fórmula para o cálculo da seletividade da junção espacial, e multijunção espacial. No entanto seu trabalho se aplica apenas para junção espacial não distribuída.

3.3.5 Efficient Processing of Multiway Spatial Join Queries in Distributed Systems (T5)

Oliveira (2017) realizou um estudo completo não somente nas consultas de junção espacial sequencial e distribuída como também na multijunção espacial sequencial e distribuída. Identificou as características de *datasets* e distribuição dos dados que são relevantes para o processamento eficiente da junção espacial e propôs um modelo estatístico mais eficaz para a estimativa da seletividade nas consultas de junção levando em consideração estas características. Em conjunção a isso, propôs o método de sobreposição proporcional para o enquadramento dos objetos, e o histograma IHWAF que utiliza o método de sobreposição proporcional e um método que particiona o *dataset* em uma grade que não possui tamanho fixo. Além disso identificou que para melhorar a estimativa

de seletividade, histogramas melhorados como o de Euler podem ser empregados em um sistema distribuído.

3.4 Resumo Comparativo

Observa-se nos trabalhos recentes que estudaram a junção espacial em sistemas distribuídos (SABEK; MOKBEL, 2017; NOBARI et al., 2013; OLIVEIRA; COSTA; RODRIGUES, 2015), que os métodos para o cálculo da estimativa de seletividade foram propostos há quase duas décadas, em contrapartida pesquisas recentes apontam que para melhorar a eficiência do processamento de junções espaciais distribuída, métodos que calculam de maneira mais assertivas a estimativa de seletividade devem ser utilizados (OLIVEIRA, 2017). O método proposto por Acharya, Poosala e Ramaswamy (1999) resolve o problema da densidade espacial em alguns *datasets*, no entanto ainda sofre com o problema da contagem múltipla de objetos. Sun, Agrawal e El Abbadi (2002) estudaram a estimativa de seletividade nas junções espaciais, no entanto, não houve a utilização do mesmo em um sistema distribuído.

Como mostra a Tabela 1, no cruzamento de critérios e trabalhos e consequentemente na comparação com a proposta deste trabalho é possível perceber alguns direcionamentos: Os trabalhos T1 e T3 estudaram principalmente métodos para aumentar a assertividade no cálculo da seletividade. Os trabalhos T2, T4 e T5 realizaram estudos em relação a junção espacial distribuída. Neste sentido percebe-se que nenhum dos trabalhos elencados neste capítulo utilizou os benefícios do histograma de Euler no cálculo da seletividade para melhorar a assertividade e consequentemente executar junções espaciais de maneira mais eficiente. Em conjunção a isso, nenhum trabalho utilizou o histograma de Euler em junções espaciais em um sistema distribuído.

Tabela 1 – Comparativo entre trabalhos

	C1	C2	C3	C4
T1			X	
T2		X		
T3	X			X
T4	X	X		
T5	X	X	X	
Este Trabalho	X	X	X	X

4 IMPLEMENTAÇÃO E VALIDAÇÃO DOS ALGORITMOS

4.1 Introdução

Este capítulo apresenta os algoritmos desenvolvidos neste trabalho. Portanto serão apresentados os algoritmos para a construção do histograma de Euler, enquadramento de objetos, as estruturas de dados necessárias. Além disso apresenta os algoritmos do histograma de grade, pois este será o objeto de comparação. Neste sentido, a seção 4.2 apresenta todas as estruturas de dados e algoritmos utilizados para o cálculo da estimativa de seletividade para as consulta de janela e junção espacial.

4.2 Implementação do Histograma de Euler

Um histograma é uma grade multidimensional, neste sentido faz sentido utilizar uma matriz para armazenar cada *bucket* do mesmo. Um *bucket* armazena metadados referente a região espacial que ele representa. Um histograma de Euler armazena mais *buckets*, pois além das células da grade é preciso armazenas também para os cantos (vértices) e bordas (arestas).

O Algoritmo 1 apresenta as estruturas de dados do histogramas de Euler. Foram criadas quatro estruturas para compôr o histograma. A primeira delas foi a estrutura do MBR que possui as variáveis $min[d]$ e $max[d]$ onde d significa a dimensão, por exemplo, quando $d = 2$ temos um MBR de duas dimensões. A estrutura de faces apresentada na linha 6, possui as variáveis *cardinalidade* que armazena a quantidade total de objetos que a intersectam e a variável $comprimento\_medio[d]$ que representa a média das larguras dos objetos que intersectam aquela face na dimensão d. A linha 11 mostra a estrutura que representa uma aresta do histograma, ela possui um MBR. É importante ressaltar que uma aresta pode ser vertical ou horizontal, na aresta vertical o MBR terá valores nulos nas dimensões x, e na horizontal nulos na dimensão y. Além disso, a estrutura aresta possui uma variável *cardinalidade* e $projecao\_media$, a qual armazena a projeção média dos objetos naquela aresta. A linha 17 exibe a estrutura Vértice que possui as coordenadas nas variáveis x e y, e a *cardinalidade*.

O Algoritmo 2 completa o Algoritmo 1 e apresenta a estrutura do histograma de Euler. Na linha 23 está declarado o MBR que representa os limites do histograma. As linhas 24 e 25 mostram as variáveis $xqtd$ e $yqtd$ que representam o tamanho da grade. Desta forma, um histograma com um tamanho de grade, por exemplo, 50×50 terá 50

Algoritmo 1 Estruturas de Dados Histograma de Euler Parte 1

```
1: estrutura MBR
2:     double min[d]
3:     double max[d]
4: fim estrutura
5:
6: estrutura Face
7:     double cardinalidade
8:     double comprimento_medio[d]
9: fim estrutura
10:
11: estrutura Aresta
12:     MBR mbr
13:     double projecao_media
14:     double cardinalidade
15: fim estrutura
16:
17: estrutura Vertice
18:     double x
19:     double y
20:     double cardinalidade
21: fim estrutura
```

faces na horizontal e 50 na vertical. As 30, 31 e 32 apresentam os vetores que contém as faces, arestas e vértices do histograma.

Algoritmo 2 Estruturas de Dados Histograma de Euler Parte 2

```
22: estrutura Histograma_de_Euler
23:     MBR mbr
24:     int xqtd
25:     int yqtd
26:     double xtam
27:     double ytam
28:     double xtics[x_1, x_2, ..., x_n]
29:     double ytics[y_1, y_2, ..., y_n]
30:     Face faces[f_1, f_2, ··· , f_n]          ▷ Vetor de faces do histograma
31:     Aresta arestas[a_1, a_2, ··· , a_n]      ▷ Vetor de arestas
32:     Vertice vertices[v_2, v_2, ··· , v_n]    ▷ Vetor de vértices
33: fim estrutura
```

O Algoritmo 3 apresenta o método de enquadramento do histograma de Euler. O procedimento Enquadramento-Histograma-De-Euler recebe como parâmetros um *dataset* A, um histograma $\mathscr{H}$ e um predicado espacial, por exemplo intersecção. Os laços encadeados a partir da linha 6 podem ser entendidos da seguinte forma, considerando θ como intersecção: para cada objeto no *dataset*, percorra o histograma de Euler e se este objeto intersecar uma face, aresta ou vértice, então incremente em uma unidade a

cardinalidade da face e do vértice, e decremente em uma unidade a cardinalidade da aresta. De forma a diminuir os erros causados pelos MBRs, ou seja, quando os MBRs se intersectam mas os objetos reais não, foi utilizado um método que recorta o objeto pela face i, j apresentado na linha 9.

Algoritmo 3 Enquadramento de objetos de um *dataset* A em um histograma de Euler $\mathscr{H}$ de acordo com um predicado θ

```
1:  procedimento ENQUADRAMENTO-HISTOGRAMA-DE-EULER(A, ℋ, θ)
2:      Seja f(i, j) a face na linha i e coluna j de ℋ
3:      Seja a(i, j) a aresta na linha i e coluna j de ℋ
4:      Seja v(i, j) o vértice na linha i e coluna j de ℋ
5:
6:      para cada a ∈ A faça
7:          para i ← 0 até ℋ.xsize faça
8:              para j ← 0 até ℋ.ysize faça
9:                  a ← CLIP(a, f(i, j))                ▷ Recorte no objeto.
10:                 se aθf(i, j) então
11:                     f(i, j).cardinalidade+ = 1
12:                 fim se
13:                 se aθa(i, j) então
14:                     a(i, j).cardinalidade− = 1
15:                 fim se
16:                 se aθa(i, j) então
17:                     v(i, j).cardinalidade+ = 1
18:                 fim se
19:             fim para
20:         fim para
21:     fim para
22: fim procedimento
```

Após o enquadramento dos objetos do *dataset* no histograma de Euler, é possível realizar as consultas de janela e de junção espacial. O Algoritmo 4 apresenta a o passo a passo para se estimar a seletividade de dois *datasets*. A função ESTIMA-CARDINALIDADE-JUNÇÃO-ESPACIAL recebe como parâmetros dois histogramas de Euler. Para cada partição do histograma $\mathscr{H}_a$ são verificadas nas partições de $\mathscr{H}_b$ se existe intersecção entre ambas. Caso exista intersecção entre alguma partição dos histogramas, então é necessário calcular a estimativa seletividade das mesmas. As funções `OBTER_FACE`, `OBTER_ARESTA`, `OBTER_VERTICE` são funções auxiliares para se obter a referência para determinada aresta, face ou vértice dos histogramas. Assim, para calcular a seletividade de duas partições utilizou-se o método proposto por (MAMOULIS; PAPADIAS, 2001).

Mamoulis e Papadias (2001) mostraram que dado um histograma H_A, a cardinalidade de saída $O^{\bar{w}}$ de uma consulta de janela $\bar{w}$ pode ser estimada com a Equação 4.1, onde $\bar{\bar{a}}$ é a cardinalidade de a, d é o número de dimensões dos dados, l_{ak} representa o comprimento médio do conjunto de objetos em a em uma dimensão em k, $l_{\bar{w}k}$ é o comprimento

Algoritmo 4 Estima a cardinalidade da junção espacial de dois *datasets*, de acordo com um predicado θ

```
1: função ESTIMA-CARDINALIDADE-JUNÇÃO-ESPACIAL(ℋ_A, ℋ_B)
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:         v_a ← OBTER_VERTICE(ℋ_a, i, j)       ▷ Obtém o vértice de ℋ_a na linha i coluna j
9:         a_a ← OBTER_ARESTA(ℋ_a, i, j)        ▷ Obtém a aresta de ℋ_a na linha i coluna j
10:        para cada p_b(r,s) ∈ P_b faça
11:            se p_a(i,j).mbr ∩ p_b(r,s).mbr então
12:                f_b ← OBTER_FACE(ℋ_b, r, s)
13:                resultado += ESTIMA-CARDINALIDADE-METODO-MP(f_a, f_b)
14:
15:                v_b ← OBTER_VERTICE(ℋ_b, r, s)
16:                se v_a.x = v_b.x e v_a.y = v_b.y então
17:                    resultado+= v_a.cardinalidade * v_b.cardinalidade
18:                fim se
19:
20:                a_b ← OBTER_ARESTA(ℋ_b, r, s)
21:                se a_a.mbr ∩ a_b.mbr então
22:                    resultado -= ESTIMA-CARDINALIDADE-METODO-MP(a_a, a_b)
23:                fim se
24:            fim se
25:        fim para
26:    fim para
27:    retorne resultado
28: fim função
```

de $\bar{w}$ na dimensão k, e l_{uk} é o comprimento de a na dimensão k, $l_{uk} \neq 0$.

$$O^{\bar{w}} = \bar{\bar{a}} \cdot \prod_{k=1}^{d} min\Big(1, \frac{l_{ak} + l_{\bar{w}k}}{l_{uk}}\Big) \tag{4.1}$$

Para junções espaciais, dado um par de partições que se intersectam $\{a, b\}$, dos histogramas H_A e H_B, geradas a partir dos *datasets* A e B, a cardinalidade de saída da junção espacial O^j para o par de partições pode ser estimada pela Equação 4.2, onde i é a intersecção dos MBRs de a e b, l_{ik} é o comprimento de i na dimensão k, $l_{ik} \neq 0$, e os outros termos são definidos de acordo com a equação anterior (4.1). Esta é uma extensão de (4.1), em que a cardinalidade dos dois *datasets* é limitada por sua área comum (i) e o número esperado de interseções é reduzido pela aplicação de consultas de janela de tamanho l_{bk} em a. O espaço de trabalho, usado para normalizar os comprimentos, é também limitado

pela área comum, isto é, determinado pelo comprimento de i (l_{ik}).

$$O^{j}(a,b) = O^{\bar{w}}(a,i) \cdot O^{\bar{w}}(b,i) \cdot \prod_{k=1}^{d} min\left(1, \frac{l_{ak} + l_{bk}}{l_{ik}}\right) \tag{4.2}$$

Algoritmo 5 Cálculo da estimativa de seletividade para duas partições do histograma.

1: **função** ESTIMA-CARDINALIDADE-METODO-MP(a, b)
2: Seja $a.mbr_k$ o comprimento de $a.mbr$ da dimensão k
3: Seja $b.mbr_k$ o comprimento de $b.mbr$ da dimensão k
4:
5: $l_{ak} = a.comprimento\_medio$
6: $l_{bk} = b.comprimento\_medio$
7:
8: $\bar{\bar{a}} = a.cardinalidade * \prod_{k=1}^{d} min\left(1, \frac{l_{ak}+inters_k}{a.mbr_k}\right)$
9: $\bar{\bar{b}} = b.cardinalidade * \prod_{k=1}^{d} min\left(1, \frac{l_{ak}+inters_k}{b.mbr_k}\right)$
10: `estimado` $= \bar{\bar{a}} * \bar{\bar{b}} * \prod_{k=1}^{d} min\left(1, \frac{l_{ak}+l_{bk}}{l_{ik}}\right)$
11: **retorne** `estimado`
12: **fim função**

4.3 Avaliação dos Algoritmos

No que se refere a avaliação dos avaliação dos histogramas gerados pelos algoritmos apresentados na seção 4.2 foi utilizado o *software Quantum Geographic Information System* (QGIS). O QGIS é um Sistema de Informações Geográficas de Código Aberto. O projeto nasceu em maio de 2002 e foi estabelecido como um projeto no *SourceForge*[1] em junho do mesmo ano. O QGIS tem como objetivo ser um SIG amigável ao usuário, fornecendo funções e recursos comuns. O objetivo inicial do projeto era fornecer um visualizador de dados SIG. O QGIS permite aos usuários analisar e editar informações espaciais, além de compor e exportar mapas gráficos. Em conjunção a isso, suporta camadas matriciais e vetoriais; os dados vetoriais são armazenados como recursos de ponto, linha ou polígono. Vários formatos de imagens matriciais são suportados e o software pode georreferenciar imagens (QGIS, 2011).

A Figura 16 mostra a interface do *software* QGIS apresentado o *dataset* de municípios brasileiros e o histograma de Euler para este *dataset*. Nesse *software* é possível verificar os metadados calculados pelos algoritmos, com os metadados reais do *dataset*. Além disso, permite uma visão da estrutura física do histograma, o que permite validar a criação de um histograma para um determinado *dataset*.

A Figura 17 apresenta dois *datasets* o de municípios brasileiros e o de alertas de queimadas, além disso os dois histogramas dos *datasets* também estão na figura. O

[1] SourceForge é um repositório de código fonte baseado em Web.

software também permite cortar um *dataset* por outro, e esta função foi utilizada nas seções abaixo.

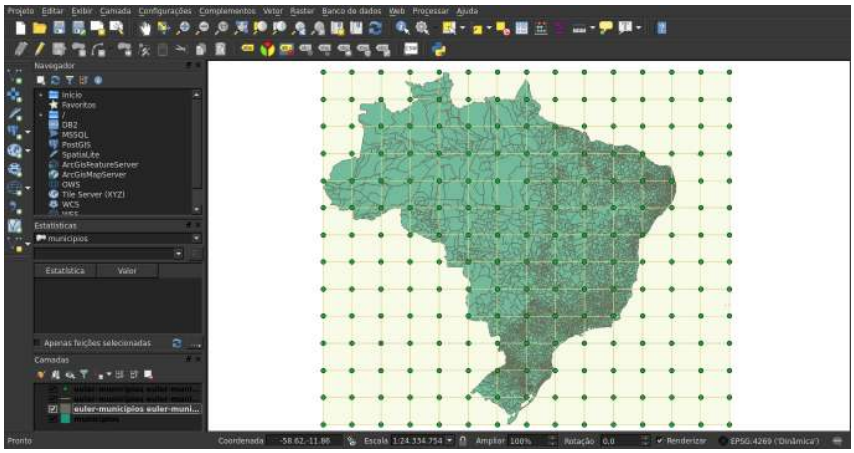


Figura 16 – QGIS com o *dataset* municípios brasileiros e o histograma de Euler.

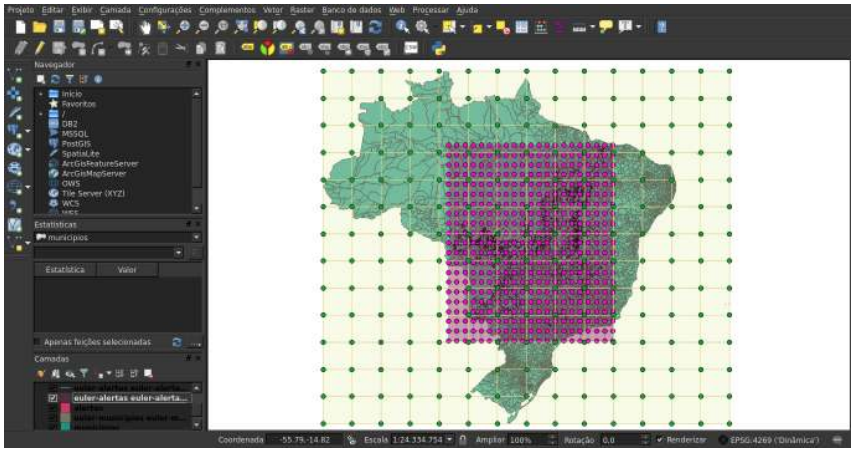


Figura 17 – QGIS com os *datasets* de munícipios e alertas com seus respectivos histogramas de Euler.

5 AVALIAÇÃO COMPARATIVA DO HISTOGRAMA DE EULER E DO HISTOGRAMA DE GRADE

5.1 Introdução

Neste capítulo será apresentada toda a metodologia de avaliação e os experimentos realizados para avaliar a proposta apresentada. As seções foram organizadas da seguinte forma: A seção 5.2 apresenta a metodologia de avaliação, isto é, explica detalhadamente como foi realizado os experimentos. A seção 5.3 mostra o resultado dos experimentos realizados nas consultas de janela. Na seção 5.4 estão os resultados e discussão dos experimentos das consultas de junção espacial. E por fim, a seção 5.5 são onde os experimentos são discutidos e avaliados de forma mais ampla.

5.2 Metodologia de Avaliação

De forma a avaliar a utilização do histograma de Euler nas consultas de janela e de junção esta seção apresenta as técnicas utilizadas para realizar os experimentos. A subseção 5.2.1 mostra os *datasets* utilizados para realizar os experimentos. Na subseção 5.2.2 as consultas espaciais realizadas são detalhadas. E por fim, a subseção 5.2.3 apresenta as métricas pelas quais os histogramas serão avaliados.

5.2.1 Dados Utilizados

Para avaliar a proposta apresentada foi escolhido um conjunto de datasets reais, obtidos no site do IBGE (<www.ibge.org>), do Laboratório LAPIG do Instituto de Estudos Sócio-Ambientais da UFG (<www.lapig.iesa.ufg.br/lapig/>) e do Digital Chart of the World (<https://gis-lab.info/qa/vmap0-eng.html>). Os *datasets* empregados e suas características estão descritos na Tabela 2. *Datasets* com cardinalidade pequena foram escolhidos para testar cenários onde um dos datasets restringe completamente a seletividade da consulta, retornando nenhum ou quase nenhum resultado. Os *datasets* com objetos mais complexos, do tipo linha, foram escolhidos devido a dificuldade e inexatidão de sua representação nos histogramas.

5.2.2 Consultas Espaciais

A Tabela 3 apresenta as consultas de junção, as consultas de janela foram realizadas em todos os *datasets* da Tabela 2. Tais consultas foram escolhidas, pelo fato de terem sido também utilizadas em (OLIVEIRA, 2017), o que cria a possibilidade de comparação da eficiência entre o histograma de grade utilizado pelo autor e o histograma de Euler no cálculo da seletividade.

Tabela 2 – Datasets utilizados nos testes

Nome	**Sigla**	**Tipo**	**Cardinalidade**	**Tamanho SHP(MB)**
Vegetação	V	Polígonos	2.140	4,7
Municípios	M	Polígonos	5.564	38,8
Alertas desmat. cerrado	A	Polígonos	32.578	11,2
Rodovias	R	Linhas	51.646	15,2
Hidrografia	H	Linhas	226.963	64,5
Culturas	CU	Polígonos	123.746	69,3
Ferrovias	FM	Linhas	194.261	28,7
Represas de água	RA	Polígonos	338.860	136,7
Contorno de relevo	CR	Linhas	703.574	572,5
Hidrografia Mundial	HM	Linhas	943.638	243,2

Tabela 3 – Junções espaciais realizadas no experimento.

Nome	**Consulta**	**Cardinalidade**	**Nome**	**Consulta**	**Cardinalidade**
J_1	$A \bowtie H$	4,868	J_{11}	$HM \bowtie FM$	58,885
J_2	$A \bowtie R$	3,395	J_{12}	$HM \bowtie RA$	530,782
J_3	$A \bowtie C$	34,261	J_{13}	$HM \bowtie CR$	449,309
J_4	$A \bowtie V$	34,672	J_{14}	$HM \bowtie CU$	269,301
J_5	$H \bowtie R$	55,766	J_{15}	$FM \bowtie RA$	5,975
J_6	$H \bowtie C$	268,369	J_{16}	$FM \bowtie CR$	47,106
J_7	$H \bowtie V$	252,830	J_{17}	$FM \bowtie CU$	121,007
J_8	$R \bowtie C$	70,304	J_{18}	$RA \bowtie CR$	22,128
J_9	$R \bowtie V$	63,339	J_{19}	$RA \bowtie CU$	79,002
J_{10}	$C \bowtie V$	15,678	J_{20}	$CR \bowtie CU$	234,900

5.2.3 Métricas

A fim de identificar a assertividade de uma consulta, foi utilizada uma métrica conhecida como Erro Relativo Médio (ERM), cuja função é comparar a resposta esperada de uma consulta com a resposta obtida, sua representação matemática pode ser observada na Equação 5.1, onde Q é o conjunto de consultas avaliadas, r_i é o valor real da seletividade

na i-ésima consulta e e_i é o valor estimado.

$$ERM(Q) = \frac{\sum_{q_i \in Q} |r_i - e_i|}{\sum_{q_i \in Qt} r_i} \tag{5.1}$$

Outra métrica a ser utilizada é o desvio padrão do erro, cujo objetivo é mostrar o quanto variou o resultado obtido com a consulta (quantidade de objetos) com o resultado esperado (quantidade exata de objetos), dessa forma o desvio padrão para um conjunto de consultas Q é $\sigma^2 = (\{r_i - e_i\} \forall q_i \in Q)$

5.3 Avaliação a Seletividade das Consultas de Janela

Esta seção apresenta a avaliação do histograma de Euler nas consultas de janela. Faz-se necessário a avaliação do histograma de Euler nas consultas de janela pois não foi verificado na literatura sua utilização com *datasets* reais com objetos do tipo linha e polígonos. Portanto, foi realizado um experimento comparando o histograma de Euler com o histograma de grade. O experimento consiste em criar, para cada *dataset* na Tabela 2, dois conjuntos de histogramas multidimensionais: o conjunto A, usando o histograma de Euler e o conjunto B, usando o histograma de grade. Cada conjunto de histograma tem tamanho $r \times r, r \in R = \{10, 20, 30, 50, 70, 100\}$. Para cada histograma $t \in \{A, B\}$ e tamanhos r, foi gerado um conjunto $\mathcal{Q}_{rs}^{t}$ de consultas de janela com seus lados medindo $s\%$ da extensão espacial do *dataset*, $s \in \{5, 10, 15, 20, 25, 30\}$. As consultas foram geradas em quantidades e posições de forma a cobrir completamente o *dataset*.

Os gráficos da Figura 18 apresentam o resultado do experimento para o *dataset* alertas. Foi plotado o erro relativo médio no eixo y e o tamanho do histograma, ou seja $r \times r$ no eixo x, além disso a linha tracejada azul representa o histograma de grade, e a linha preta com pontos e traços representa o histograma de Euler. Para $s = 5\%$ e $s = 15\%$ a diferença do erro relativo médio nos dois histogramas não foi tão expressiva. No entanto para $s = 10\%$ e $s = 20\%$ é possível notar uma grande diferença, enquanto o erro relativo médio do histograma de Euler tende a zero o histograma de grade segue em um crescimento quase linear.

A Figura 19 apresenta o resultado experimento para o *dataset* municípios. Neste o crescimento do erro relativo médio do histograma de grade é ainda mais acentuado comparando com o *dataset* alertas, por exemplo para $s = 15\%$ e $r = 100$ o erro relativo médio do histograma de Euler é $0,20$ com $\sigma^2 = 31,6$ enquanto para o histograma de grade o erro relativo médio é $3,1$ com $\sigma^2 = 467,1$. O histograma de Euler mantém um erro relativo médio próximo de zero. Em contraposição o histograma de grade tende a um crescimento quase quadrático.

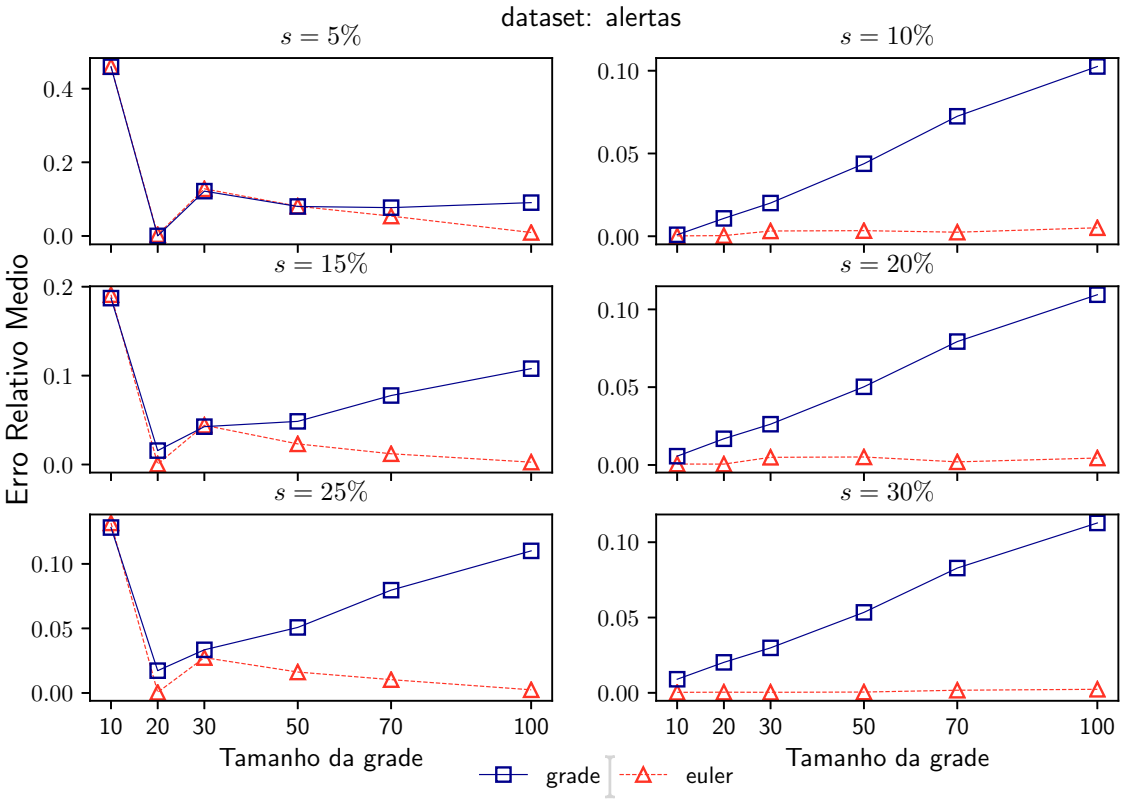


Figura 18 – Resultado das consultas de janela para o *dataset* Alertas

A Figura 20 apresenta a interpolação curva do histograma de grade do *dataset* municípios para $s = 30\%$, nesta situação o erro cresce na ordem de $x^{1,39}$ e aumenta proporcionalmente ao aumento de r.

O *dataset* de hidrografias diferentemente dos dois apresentados nos gráficos acima possui objetos somente do tipo linha. A Figura 21 apresenta o resultado da comparação no *dataset* de hidrografias. Assim como nos outros o histograma de Euler é melhor, isto é, possui um erro relativo médio abaixo do histograma de grade. Para $s = 10\%$ e $s = 20\%$ o histograma de grade tem um crescimento quase linear, enquanto que novamente o erro relativo médio do histograma de Euler se aproxima de zero. É importante notar também que no histograma de grade, para $s = 15\%$ $\sigma^2 = 1007,8$ enquanto no histograma de Euler $\sigma^2 = 26,7$ e essa diferença aumenta proporcionalmente ao tamanho de r.

As outras consultas de janela nos *datasets* restantes seguem o comportamento da consultas apresentadas neste capítulo, dessa forma foram movidas para o Anexo A.

5.4 Avaliação da Seletividade das Consultas de Junção Espacial

Esta seção apresenta a avaliação do histograma de Euler no cálculo da seletividade nas junções espaciais. Os resultados obtidos com o histogramas de Euler nos experimen-

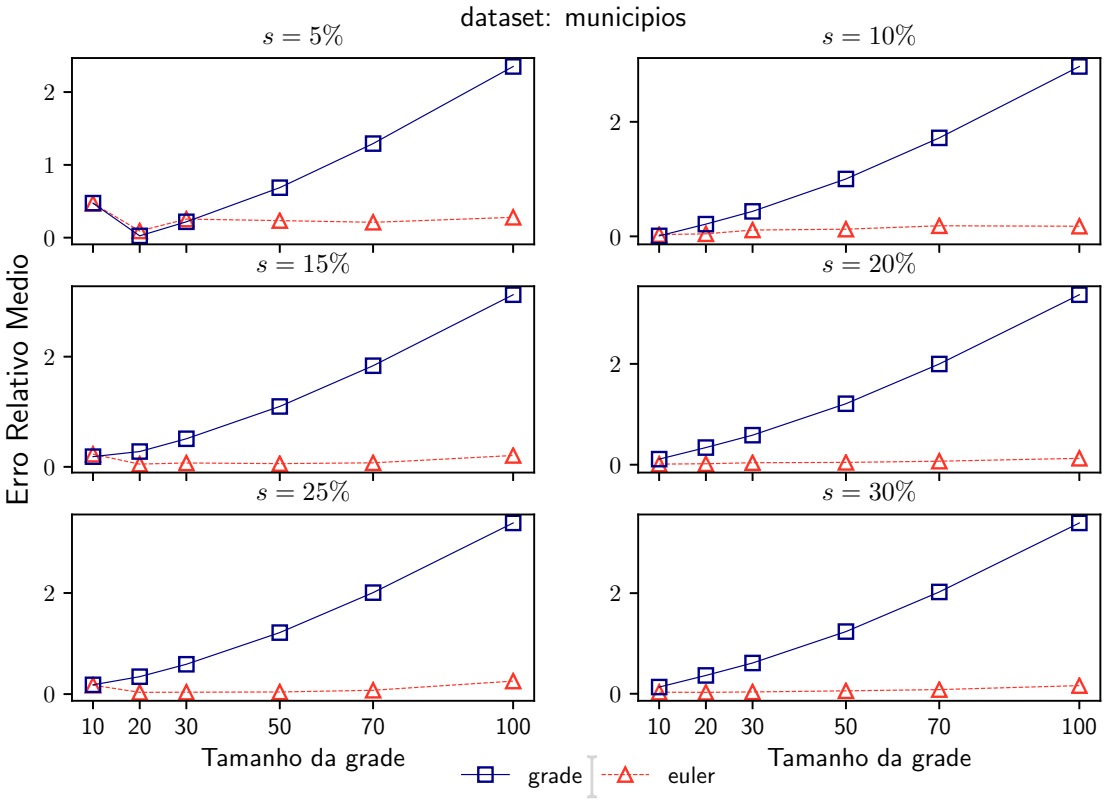


Figura 19 – Resultado das consultas de janela para o *dataset* Municipios

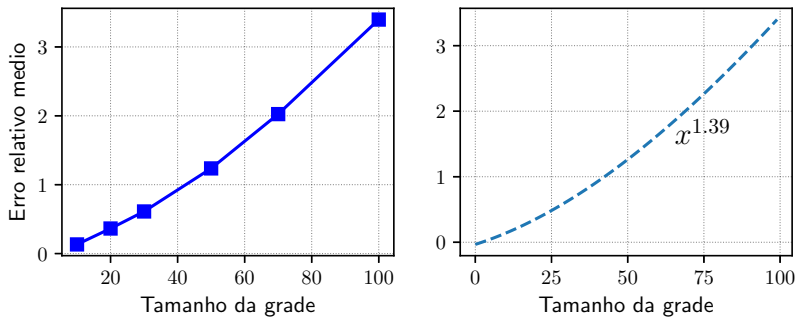


Figura 20 – Regressão do crescimento no *dataset* municípios para $s = 30\%$

tos realizados foram comparados com os resultados do histograma de grade utilizando o método de enquadramento sobreposição proporcional.

Foram construídos histogramas de Euler para os *datasets* das junções J_1 até J_{10} da Tabela 3 para cada tamanho de histograma $r \in \{10, 20, 30, 50, 70, 100\}$. Assim, para cada partição do histograma de Euler $a \in \mathscr{H}_A$ e $b \in \mathscr{H}_B$ foi calculada a estimativa de seletividade da junção das duas partições, utilizando a função ESTIMA-CARDINALIDADE-METODO-MP do Algoritmo 5. Em seguida as consultas foram realizadas de forma a obter o valor real. Dessa forma, com o valor real r_{ab} e o valor estimado e_{ab} da junção das partições $\{a, b\}$ foi calculado o ERM para cada tamanho de histograma r. O mesmo experimento foi realizado para o histograma de grade.

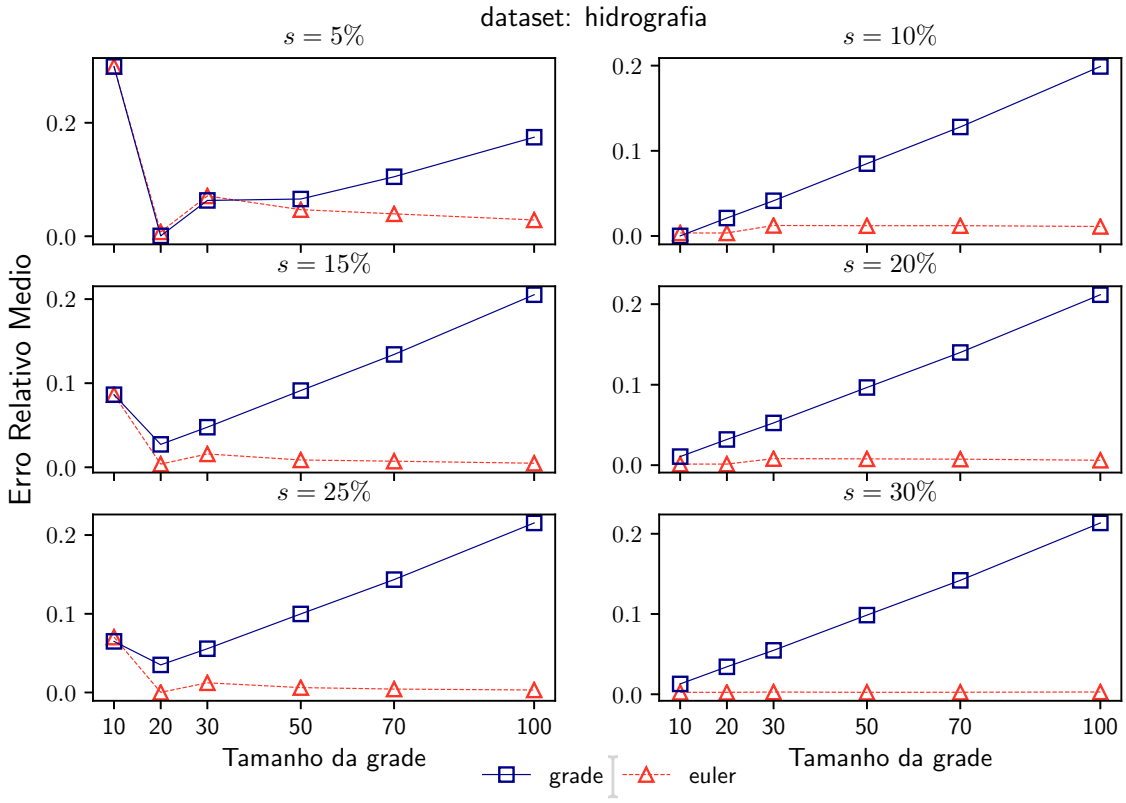


Figura 21 – Resultado das consultas de janela para o *dataset* Hidrografia

Os gráficos da Figura 22 apresentam os resultados do experimento. No eixo y é possivel ver o ERM, e no eixo x o tamanho r dos histogramas. Os valores do histograma de Euler estão em azul, e do histograma de grade em vermelho. As junções $A \bowtie H$, $A \bowtie R$ e $H \bowtie R$ são onde os ERMs estão mais próximos, além disso em $A \bowtie H$ apesar de pequena a diferença, é a única junção em que o histograma de Euler tem um ERM menor do que o histograma de grade. Nas outras junções o histograma de grade é sempre superior ao histograma de Euler, isso ocorre principalmente pois as grades dos histogramas não coincidem, ou seja, para toda aresta $a_A \in \mathscr{H}_A$ e $a_B \in \mathscr{H}_B$ então $a_A.mbr \cap a_B.mbr = \emptyset$ portanto as arestas no histograma de Euler não irão ser subtraídas, uma vez que a subtração só ocorre quando as as arestas se intersectam.

Para verificar a assertividade do histograma de Euler quando as grades se alinham, foi realizado outro experimento, que consistiu em cortar os *datasets* utilizando o *software* QGIS, de modo que após o corte os *datasets* ficassem com os mesmos limites e do mesmo tamanho, dessa forma para quaisquer *datasets* A, B cortados com $r = k \times k$ as grades se alinham. Este processo de corte foi realizado para as consultas J_1 até J_4, não foi realizado com as demais junções devido ao esforço necessário para realizar os cortes com o *software*.

A Figura 23 apresenta o resultado do experimento, o histograma de grade com a cor azul, histograma de Euler quando as grades não se alinham em vermelho e o histograma de

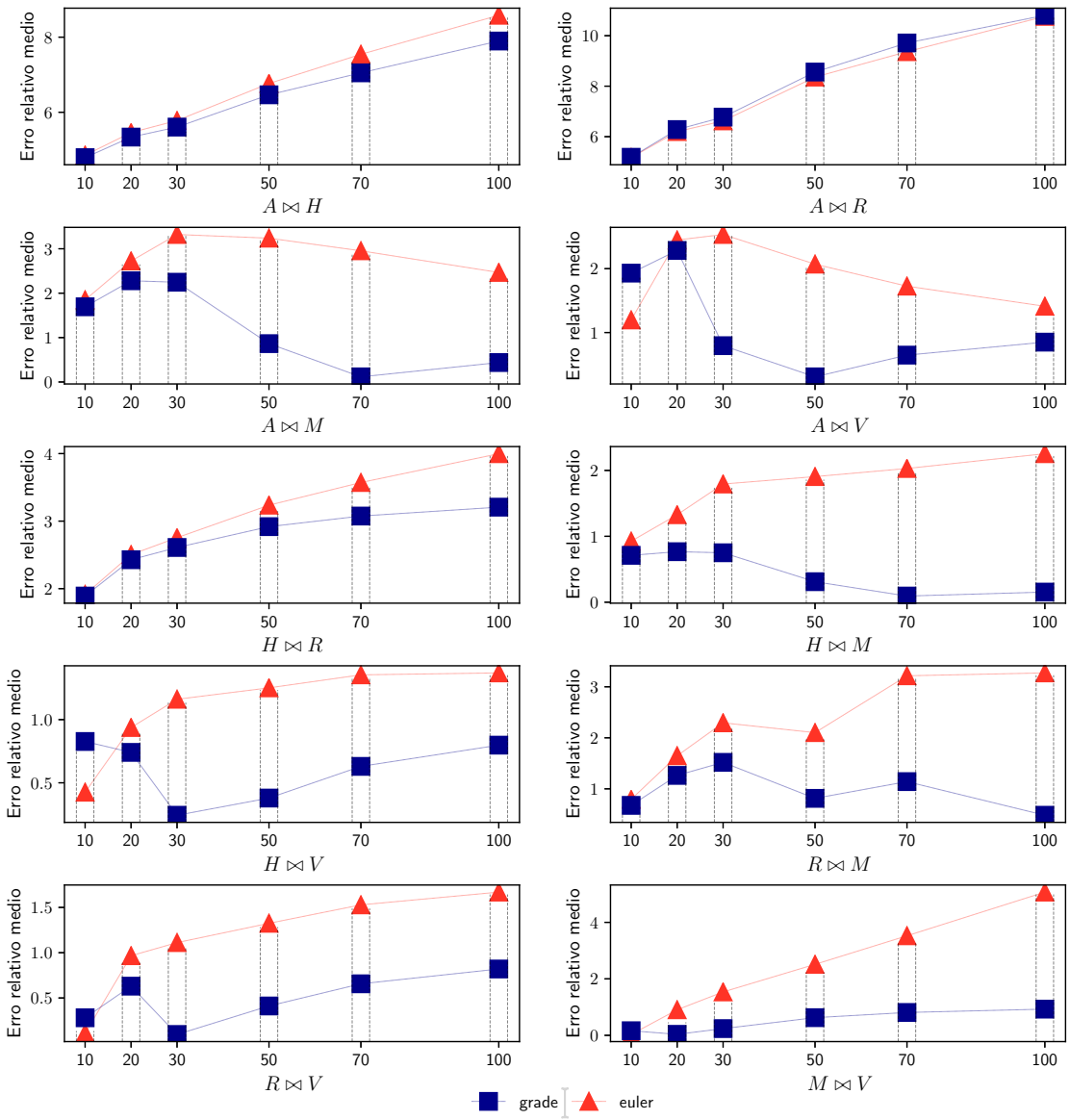


Figura 22 – Precisão da seletividade das junções J_1 até J_{10} por partição.

Euler com as grades alinhadas em verde. No eixo y estão os ERMs e no eixo x o tamanho da grade. É possível ver uma diferença grande em relação ao ERM do histograma de Euler alinhado com o dos outros histogramas, por exemplo na consulta $A \bowtie M$ enquanto o histograma de euler não linhado chega a ter um ERM acima de 3.0, o histograma de Euler com as grades alinhadas se aproxima de um ERM $= 0$. O histograma de grade só tem um ERM menor que o histograma de Euler com as grades alinhadas em apenas uma ocasião em $A \bowtie M$ quando $r = 70$.

Foi realizado outro experimento a fim de comparar o histograma de Euler com o histograma de grade melhorado denominado IHWAF (OLIVEIRA, 2017), que utiliza além do método de enquadramento sobreposição proporcional, outros métodos que visam diminuir o erro na seletividade como por exemplo, um método de particionamento que não

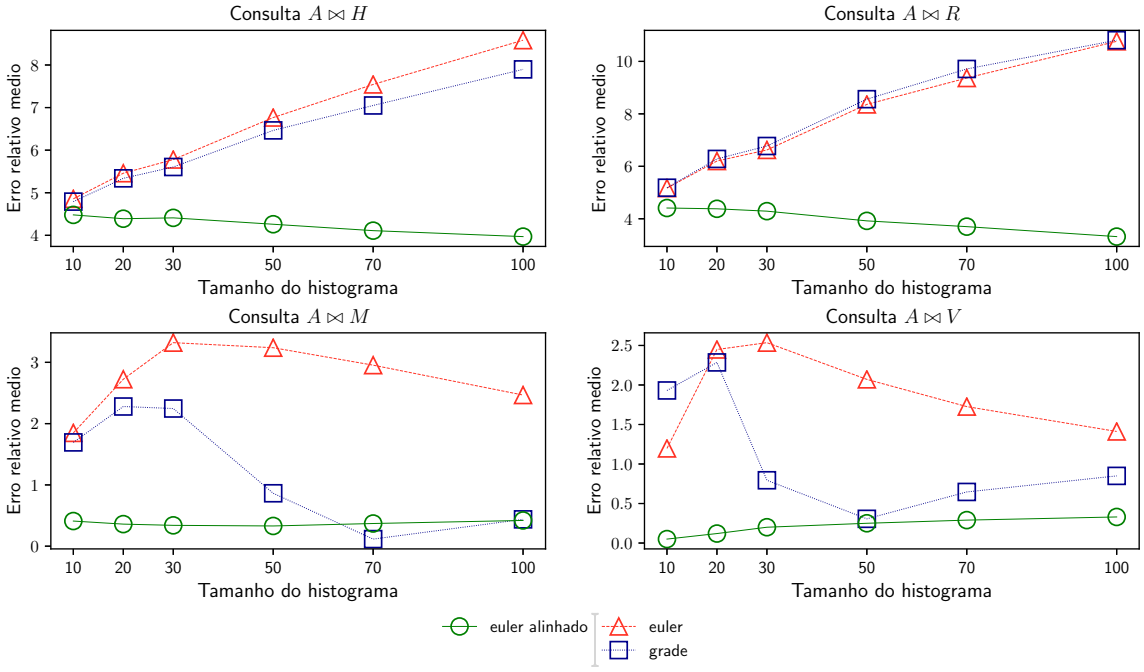


Figura 23 – Resultados quando os histogramas de Euler se alinham.

divide a grade do histograma em tamanhos iguais. Foram criados histogramas de grade e de Euler para os as junções de J_1 até J_{20}, o tamanho das grades foram determinados pela Tabela 3.3 proposta em (OLIVEIRA, 2017). A cardinalidade estimada de cada junção foi obtida e comparadas com a cardinalidade real. Igualmente, erro relativo de cada junção também foi calculado.

A Tabela 4 apresenta os resultados do experimento. Eles são agrupados pelos tipos de objetos espaciais na consulta ($L \bowtie L$ para *datasets* com objetos de linha, $P \bowtie P$ para dois *datasets* com objetos do tipo polígonos e $L \bowtie P$ ou $P \bowtie L$ para *datasets* com objetos de linha e polígono) ordenada pelo erro relativo do método IHWAF. As primeiras três colunas indicam a consulta, o tipo de objetos e a cardinalidade real daquela junção. As seguintes colunas apresentam a cardinalidade estimada do histograma de grade com método de sobreposição proporcional e histograma de Euler, e por último a coluna do erro relativo para cada um dos histogramas, além de apresentar também o erro relativo do método IHWAF.

Nas primeiras quatro linhas onde o tipo de objeto dos *datasets* são do tipo linha, o histograma de Euler tem um erro relativo menor apenas em J_{13} se comparado ao histograma de grade, mas não ao IHWAF que tem $1,3$. Das 20 junções realizadas o histograma de Euler possui um erro relativo menor do que o histograma de grade em 7 ocasiões, e se comparado ao IHWAF em apenas 2. Além disso o histograma de Euler vence os dois outros métodos ao mesmo tempo em apenas 1 junção (J_{10}). É importante notar que o histograma de Euler tem seus melhores resultados, isto é, erros relativos menores, quando

Tabela 4 – Resultado da cardinalidade estimada utilizando o histograma de Euler, grade e IHWAF

Consulta	**Tipo**	**C. Real**	**C. Estimada**		**Erro Relativo (%)**		
			Grade	**Euler**	**Grade**	**Euler**	**IHWAF**
J_{13}	L⋈L	449.309	2.000.405	1.938.826	345,2	33,1	1,3
J_{11}	L⋈L	58.885	258.538	298.162	339,0	406,3	2,3
J_5	L⋈L	55.766	370.194	316.381	563,8	467,3	21,5
J_{16}	L⋈L	47.106	269.433	382.334	471,9	711,6	31,0
J_4	P⋈P	34.672	46.563	36.874	34,2	6,3	1,8
J_3	P⋈P	34.261	93.656	60.991	173,3	78,0	4,2
J_{10}	P⋈P	15.678	21.391	15.754	36,4	0,4	11,3
J_{19}	P⋈P	79.002	61.030	97.232	22,7	23,0	32,6
J_{17}	L⋈P	121.007	106.328	153.148	12,1	26,5	4,3
J_7	L⋈P	252.830	410.382	309.036	62,3	22,2	4,4
J_8	L⋈P	70.304	209.087	282.429	197,4	301,7	7,1
J_9	L⋈P	63.339	90.757	124.731	43,2	96,6	7,6
J_6	L⋈P	268.369	1.020.742	651.380	280,3	142,7	8,1
J_{14}	L⋈P	269.301	360.629	371.904	33,9	38,0	20,1
J_{15}	P⋈L	5.981	38.633	142.513	545,9	2282,7	26,0
J_{20}	L⋈P	234.900	357.002	431.497	51,9	83,6	26,7
J_{18}	L⋈P	22.128	311.861	374.414	1309,3	1592,0	28,2
J_2	P⋈L	3.395	53.679	46.993	1481,1	1284,1	33,0
J_1	P⋈L	4.868	57.954	50.262	1090,5	932,4	57,9
J_{12}	L⋈P	531.269	466.601	426.330	12,1	19,7	88,1
Média					355,3	442,3	20,9
σ^2					447,5	627,7	21,1

os objetos são do tipo polígono, por exemplo, nas junções $P \bowtie P$ o histograma de Euler venceu o histograma de grade em 3 ocasiões, e venceu o IHWAF em duas. No entanto, na média o histograma de Euler não se mostrou superior aos demais métodos, novamente é importante notar que devido ao tamanho, as grades dos histogramas de Euler em cada junção neste experimento não se sobrepunham, então as arestas não eram subtraídas no cálculo da seletividade. A média dos erros relativos para o histograma de grade foi $355,3\%$, no histograma de Euler $442,3\%$ e com o método IHWAF $20,9\%$. O desvio padrão (σ^2) do erro do histograma de grade foi $447,4\%$ do histograma de Euler $627,7\%$ e do IHWAF $21,1\%$. Em suma, o método IHWAF se mostrou consistente e obteve os melhores resultados em comparação ao histograma de Euler e o de grade utilizando o método de sobreposição proporcional.

5.5 Considerações Finais

Neste capítulo foi investigado a assertividade da estimativa de seletividade calculada utilizando o histograma de Euler nas consultas de janela e junção espacial. Foram realizados diversos experimentos que tiveram como objetivo principalmente comparar a assertividade do histograma de Euler com outros histogramas propostos na literatura. Além disso uma atenção especial foi dedicada a assertividade do histograma de Euler na estimativa da junção de cada partição do histograma.

Nas consultas de janela o histograma de Euler se mostrou mais consistente em relação ao histograma de grade utilizando o método de sobreposição proporcional, na maioria das consultas realizadas apresentou um ERM menor. No que se refere às consultas de junção espacial, o histograma de Euler não obteve o mesmo resultado das consultas de janela. No entanto foi identificado a fonte dos erros na estimativa: na junção espacial os *datasets* possuem tamanhos diferentes, portanto um *dataset A* com tamanhos de grade 10×10 e outro *dataset B* com tamanhos de grade 10×10, apesar de possuírem tamanhos de grades iguais, estas grades não são do mesmo tamanho pois os *datasets* não são do mesmo tamanho. Dessa forma, quando as grades não se alinham, consequentemente as arestas do histograma não vão se alinhar, então não vão ser subtraídas na conta da seletividade.

Para confirmar a suposição das grades que não se alinham, foi realizado um experimento em que os *datasets* foram cortados, de forma a ficarem do mesmo tamanho. Os resultados mostraram uma melhora significativa da assertividade do histograma de Euler quando as grades se alinham comparando com o histograma de grade.

A fim de comparar o histograma de Euler com o histograma de grade melhorado IHWAF, foi realizado outro experimento, onde histograma de Euler obteve o pior resultado comparando com o IHWAF e o histograma de grade.

6 CONCLUSÕES E TRABALHOS FUTUROS

6.1 Introdução

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.

6.2 Conclusões

Neste trabalho foi avaliado a assertividade da estimativa de seletividade do histograma de Euler nas consultas de janela e de junção espacial. Para isso foram realizados diversos experimentos, e os resultados foram comparados com outros métodos propostos na literatura. A estimativa foi avaliada utilizando diversos *datasets* reais e com diferentes tipos de objetos, alguns possuindo polígonos e outros apenas linhas.

Nas consultas de janela o histograma de Euler se mostrou mais assertivo do que o histograma de grade, em praticamente todas as consultas realizadas, em alguns momentos o erro relativo médio (ERM) do histograma de grade teve um crescimento quase quadrático, enquanto que o ERM do histograma de Euler se manteve em um crescimento linear quase tendendo a zero. Assim, foi possível constatar que o histograma de Euler é assertivo em *datasets* com uma grande heterogeneidade em relação aos tipos e formas dos objetos.

Nas consultas de junção espacial, a fim de avaliar a utilização do histograma de Euler em um sistema distribuído, foram realizadas junções com vários *datasets*, partição a partição. Nestes experimentos o histograma de Euler não obteve o mesmo resultado das consultas de janela, e na maioria dos casos o ERM era muito maior do que o histograma de grade, que além de se mostrar mais assertivo, se mostrou mais consistente em relação aos resultados. Por fim, foram realizados experimentos para avaliar a assertividade na junção espacial global do histograma de Euler. Para isso foi realizado um experimento comparando o histograma de Euler tanto com o histograma de grade quando com método IHWAF proposto recentemente na literatura. Novamente o histograma de Euler obteve um resultado abaixo dos outros métodos. No entando, foram identificadas fontes de erros no histograma de Euler, a principal está ligada diretamente ao posicionamento da grade dos histogramas na junção de dois *datasets*. Outras fontes de erros existem como por exemplo, a distribuição dos objetos em um *dataset* pode não ser muito densa em alguns pontos e pouco densa em outros.

Os resultados obtidos indicaram que apesar do histograma de Euler ter um embasamento matemático interessante, e resolver o problema da contagem múltipla de objetos, o mesmo não apresenta uma melhor assertividade no cálculo da seletividade para consultas do tipo junção espacial se comparado com outros métodos propostos na literatura. Os experimentos indicaram que o histograma de grade melhorado IHWAF é superior e possui mais consistência nas consultas de junção em comparação com o histograma de Euler e histograma de grade utilizando apenas o método de sobreposição proporcional.

6.3 Trabalhos futuros

Os estudos realizados neste trabalho proporcionam novas perspectivas quando ao processamento de consultas espaciais distribuídas. Como parte de trabalhos futuros pretende-se resolver as fontes de erros no histograma de Euler. A primeira delas são as grades que não se alinham, para isso métodos de particionamento podem ser desenvolvidos, de forma que o após particionar os *datasets* as grades dos histogramas de Euler passam a se alinhar. A segunda fonte de erro é o problema da densidade espacial dos objetos em um *dataset*. De modo a diminuir estes erros, técnicas de histogramas desenvolvidas para resolver este problema podem ser incorporadas ao histograma de Euler, como por exemplo o Minskew.

REFERÊNCIAS

ACHARYA, S.; POOSALA, V.; RAMASWAMY, S. Selectivity estimation in spatial databases. In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 1999. (SIGMOD '99), p. 13–24. ISBN 1-58113-084-8. Disponível em: <http://doi.acm.org/10.1145/304182.304184>. Citado 6 vezes nas páginas 15, 19, 24, 25, 34 e 37.

AJI, A. *High Performance Spatial Query Processing for Large Scale Spatial Data Warehousing*. [S.l.: s.n.], 2014. Citado na página 19.

AN, N.; YANG, Z.-Y.; SIVASUBRAMANIAM, A. Selectivity estimation for spatial joins. In: *Proceedings 17th International Conference on Data Engineering*. [S.l.: s.n.], 2001. p. 368–375. ISSN 1063-6382. Citado 2 vezes nas páginas 24 e 25.

BEIGEL, R.; TANIN, E. The geometry of browsing. In: SPRINGER. *Latin American Symposium on Theoretical Informatics*. [S.l.], 1998. p. 331–340. Citado na página 27.

BRINKHOFF, T.; KRIEGEL, H.-P.; SEEGER, B. Parallel processing of spatial joins using r-trees. In: *Proceedings of the Twelfth International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 1996. (ICDE '96), p. 258–265. ISBN 0-8186-7240-4. Citado na página 21.

ELDAWY, A.; MOKBEL, M. F. Spatialhadoop: A mapreduce framework for spatial data. *2015 IEEE 31st International Conference on Data Engineering*, p. 1352–1363, 2015. Citado na página 35.

ELMASRI, R.; NAVATHE, S. *Fundamentals of database systems*. [S.l.]: Addison-Wesley Publishing Company, 2010. Citado na página 18.

FITZ, P. R. *Geoprocessamento sem complicação*. [S.l.]: Oficina de textos, 2018. 80–89 p. Citado 2 vezes nas páginas 17 e 18.

GOLDBARG, M.; GOLDBARG, E. *Grafos: Conceitos, algoritmos e aplicações*. [S.l.]: Elsevier, 2012. Citado na página 26.

HARARY, F. Graph theory. addison wesley publishing company. *Reading, MA, USA*, 1969. Citado na página 27.

HUISMAN, O.; BY, R. D. *Principles of geographic information systems*. [S.l.: s.n.], 2009. v. 1. 17 p. Citado 2 vezes nas páginas 17 e 18.

JACOX, E. H.; SAMET, H. Spatial join techniques. *ACM Transactions on Database Systems*, v. 32, n. 1, 2007. ISSN 03625915. Citado 3 vezes nas páginas 13, 21 e 22.

KOSSMANN, D. The state of the art in distributed query processing. *ACM Computing Surveys (CSUR)*, ACM, v. 32, n. 4, p. 422–469, 2000. Citado na página 14.

MAMOULIS, N.; PAPADIAS, D. Multiway spatial joins. *ACM Transactions on Database Systems*, v. 26, n. 4, p. 424–475, 2001. ISSN 03625915. Disponível em:

<http://portal.acm.org/citation.cfm?doid=503099.503101>. Citado 3 vezes nas páginas 13, 36 e 40.

MARK, D. M. Geographic information science: Defining the field. *Foundations of geographic information science*, Taylor and Francis, New York, v. 1, p. 3–18, 2003. Citado na página 20.

MARKL, V. et al. Robust query processing through progressive optimization. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data - SIGMOD '04*, p. 659, 2004. ISSN 07308078. Citado na página 14.

NOBARI, S. et al. Touch: In-memory spatial join by hierarchical data-oriented partitioning. In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2013. (SIGMOD '13), p. 701–712. ISBN 978-1-4503-2037-5. Disponível em: <http://doi.acm.org/10.1145/2463676.2463700>. Citado na página 37.

OLIVEIRA, T. B. de. *Efficient Processing of Multiway Spatial Join Queries in Distributed Systems*. Tese (Doutorado) — Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brasil, 11 2017. Citado 13 vezes nas páginas 9, 13, 14, 15, 20, 23, 24, 25, 36, 37, 45, 50 e 51.

OLIVEIRA, T. B. de; COSTA, F. M.; RODRIGUES, V. J. do S. Definição de planos de execução distribuídos para consultas de junção espacial usando histogramas multidimensionais. In: *SBBD*. [S.l.: s.n.], 2015. p. 89–100. Citado na página 37.

ÖZSU, M. T.; VALDURIEZ, P. *Principles of distributed database systems*. [S.l.]: Springer Science & Business Media, 2011. Citado na página 22.

QGIS, D. Quantum gis geographic information system. *Open Source Geospatial Foundation Project*, v. 45, 2011. Citado na página 42.

SABEK, I.; MOKBEL, M. F. On spatial joins in mapreduce. In: ACM. *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. [S.l.], 2017. p. 21. Citado 2 vezes nas páginas 35 e 37.

SHEKELYAN, M.; DIGNÖS, A.; GAMPER, J. DigitHist: A histogram-based data summary with tight error bounds. *Proceedings of the VLDB Endowment*, v. 10, n. 11, p. 1514–1525, 2017. ISSN 21508097. Citado na página 14.

SUN, C.; AGRAWAL, D.; El Abbadi, A. Selectivity estimation for spatial joins with geometric selections. In: *Advances in Database Technology—EDBT 2002*. [S.l.: s.n.], 2002. p. 609–626. Citado 8 vezes nas páginas 13, 14, 25, 29, 30, 31, 36 e 37.

SUN, C. et al. Exploring spatial datasets with histograms. *Distributed and Parallel Databases*, v. 20, n. 1, p. 57–88, 2006. ISSN 09268782. Citado 6 vezes nas páginas 9, 25, 27, 28, 29 e 31.

TANENBAUM, A. S.; STEEN, M. V. *Distributed systems: principles and paradigms*. [S.l.]: Prentice-Hall, 2007. Citado na página 22.

Anexos

ANEXO A – RESULTADO DAS CONSULTAS DE JANELA

Neste anexo são apresentados os resultados dos experimentos para as consultas de janela que não foram apresntadas no Capítulo 5.

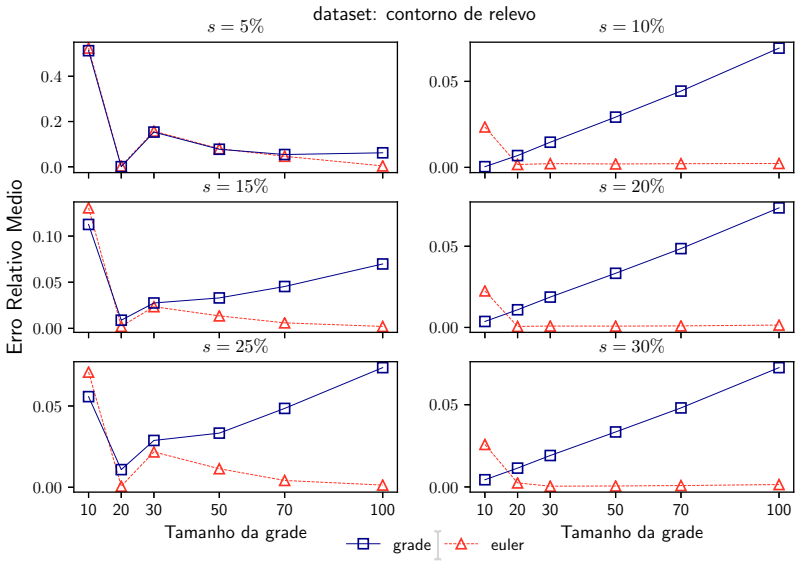


Figura 24 – Resultados das consultas de janela para o *dataset* Contornos de Relevo.

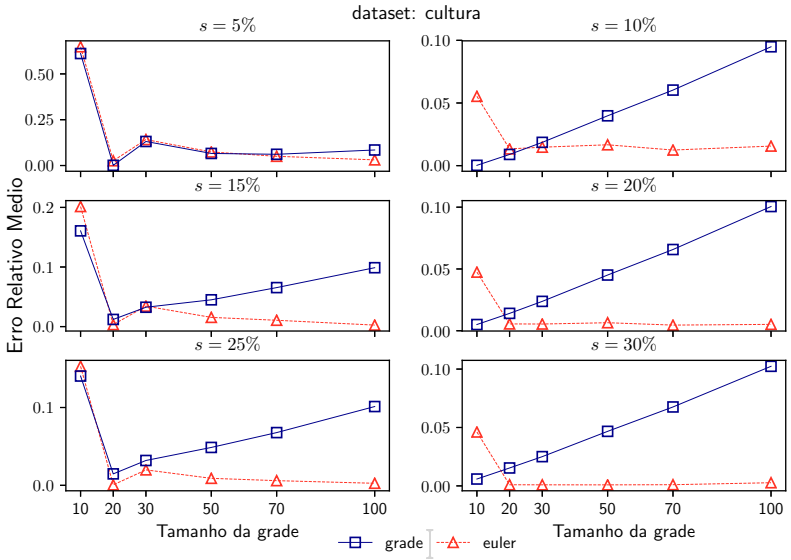


Figura 25 – Resultados das consultas de janela para o *dataset* Culturas

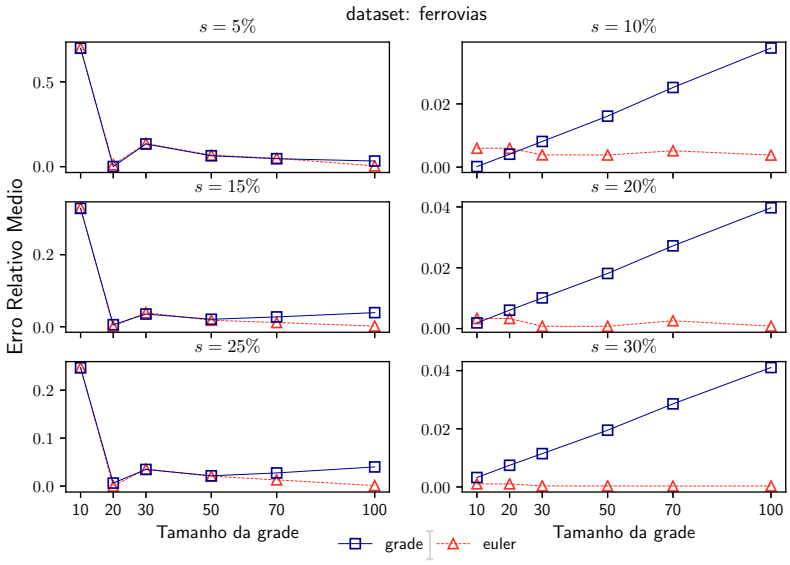


Figura 26 – Resultados das consultas de janela para o *dataset* Ferrovias

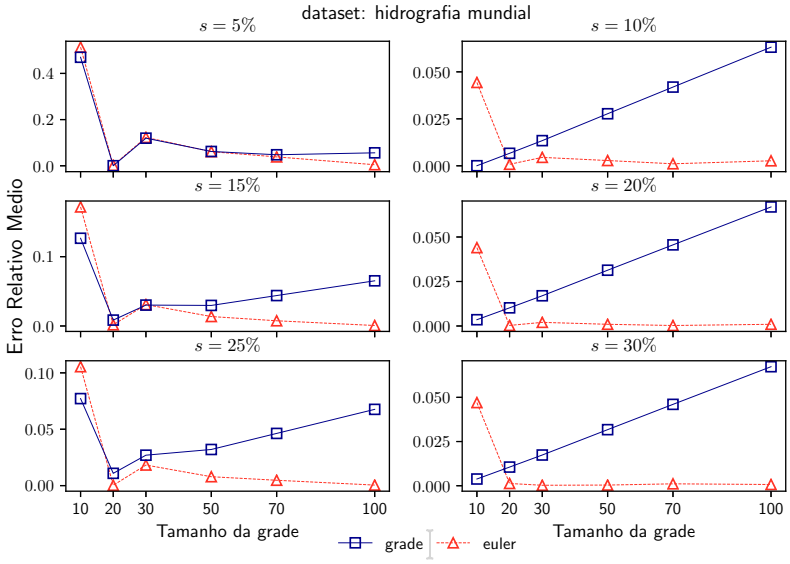


Figura 27 – Resultados das consultas de janela para o *dataset* Hidrografia Mundial

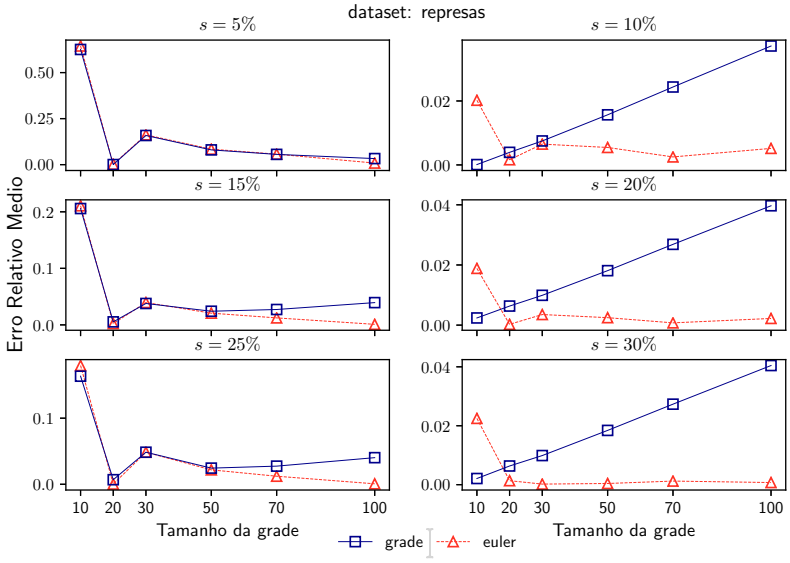


Figura 28 – Resultados das consultas de janela para o *dataset* Represas

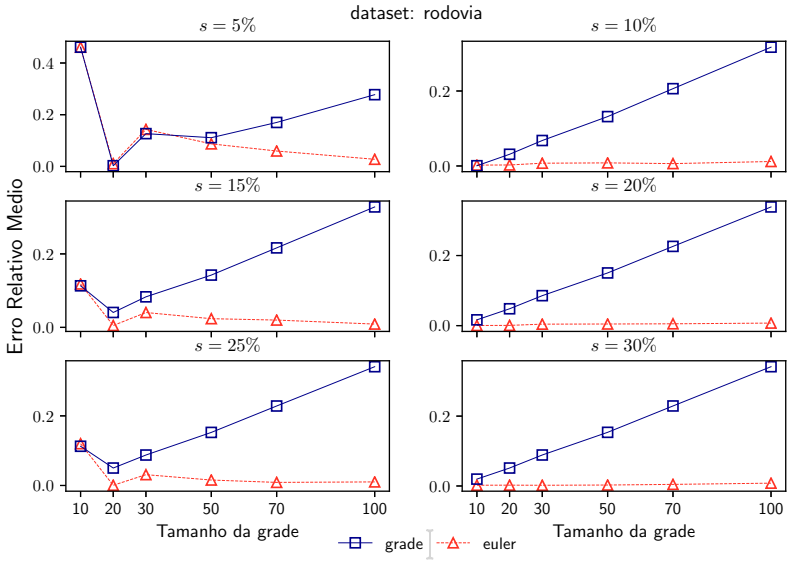


Figura 29 – Resultados das consultas de janela para o *dataset* Rodovias

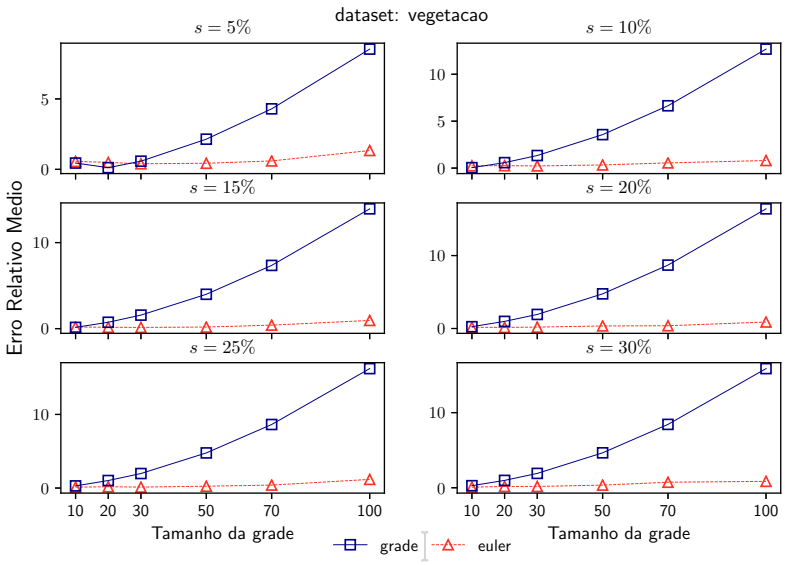


Figura 30 – Resultados das consultas de janela para o *dataset* Vegetação