Histograma Intermediário de Euler para Estimativa de Seletividade de Multijunções Espaciais

Murilo Cunha dos Santos[1], **Thiago Borges de Oliveira**[1]

[1]Instituto de Ciências Exatas e Tecnológicas (ICET)
Universidade Federal de Goiás (UFG) - Regional Jataí
Jataí, GO – Brasil

murilo_rcc@hotmail.com, thborges@ufg.br

Abstract. *This article presents a new method for building Intermediate Euler Histograms to estimate the selectivity of multiway spatial join queries. The new method is based on the original Euler Histogram and considers that the spatial extent of the spatial datasets is not the same (not aligned), a real scenario for spatial databases. Preliminary results have shown that the proposed method improved the cardinality estimation when compared to Grid Histogram, the most frequently mentioned histogram in the literature.*

Resumo. *Este trabalho apresenta um novo método de construção de Histogramas Intermediários (HIE) para estimativa de seletividade de consultas de multijunção espacial, baseando-se nas técnicas propostas para o Histogramas de Euler e considerando datasets cuja extensão espacial não se alinha, ou seja, um cenário real para banco de dados espaciais. Os resultados preliminares apontam que o método conseguiu estimar a cardinalidade com maior precisão, comparado ao método mais frequentemente referenciado na literatura, o Histograma de Grade.*

1. Introdução

Dados espaciais são usados para representar e descrever aspectos geográficos de fenômenos naturais, como, por exemplo, limites políticos de municípios, trajeto de rios e seu leito, informações do solo e de pragas em cultivos, localização de células de tumores em tomografias, dentre outros. Esses dados são coletados e organizados em *layers*, ou *datasets*, que são armazenados e processados nos Sistemas Gerenciadores de Banco de Dados Espaciais (SGBDE) usando consultas espaciais. Uma importante consulta é a junção espacial (*Spatial Join*), que encontra elementos correlacionados entre dois *datasets*, de acordo com um predicado espacial θ, como interseção ou proximidade [Brinkhoff and Seeger 2006]. Quando a consulta envolve mais de dois *datasets* é chamada de multijunção espacial (*Multiway Spatial Join*) [Mamoulis and Papadias 2001b] e é processada em etapas, processando dois *datasets* de cada vez, e produzindo resultados intermediários.

Devido possuírem múltiplas formas equivalentes de execução, ou planos de execução, cada um com uma ordem específica de *datasets*, as consultas de multijunção espacial passam por um otimizador de consultas que procura selecionar o melhor plano para execução [Mamoulis and Papadias 2001a]. Uma técnica frequentemente empregada dentro do otimizador é o histograma espacial. Histogramas são estruturas de dados que

simplificam os *datasets*, dividindo o espaço em uma grade que contenha diversas células (ou *buckets*). Estes *buckets* podem possuir tamanhos fixos ou variados, dependendo da estratégia adotada na estrutura de dados. Para cada *bucket*, são armazenados metadados a respeito dos objetos espaciais contidos espacialmente, como a quantidade de objetos (cardinalidade) e o tamanho dos objetos (quantidade de pontos) [de Oliveira et al. 2017].

No Histograma de Grade [Mamoulis and Papadias 2001a], um conjunto de células é formado dividindo-se a extensão espacial do *datasets* e os objetos são contados em cada célula do histograma que sobrepõem. Objetos que ocupam ou sobrepõem mais de uma célula são contados múltiplas vezes e isso provoca erros na estimativa de seletividade das consultas. O Histograma de Euler [Sun et al. 2002b], ao contrário do Histograma de Grade, adota métodos em sua estrutura que procuram evitar a contagem múltipla dos objetos alocando *buckets* para identificar a face da célula, as suas laterais ou arestas, e para os cantos da célula, ou vértices. O objeto é contado na estrutura do histograma tanto na face quanto nas arestas e vértices que sobrepõe. Essas contagens adicionais proporcionam uma forma de evitar a contagem múltipla do objeto durante as estimativas das consultas, tornando a estimativa do custo computacional mais assertiva. Entretanto, o Histograma de Euler foi desenvolvido originalmente para consultas de junções espaciais simples, para as quais o histograma é gerado a partir dos *datasets*. Multijunções espaciais utilizam vários *datasets* e possuem um processo de estimativa de seletividade diferenciada devido serem frequentemente executada em etapas [de Oliveira et al. 2017] e necessitarem da criação de histogramas intermediários construídos não a partir dos *datasets*, mas estimados a partir dos histogramas das etapas iniciais.

Neste trabalho, apresentamos um resultado parcial de um projeto de pesquisa que tem como objetivo a implementação de um histograma intermediário para estimativa de seletividade de consultas de multijunções espaciais, baseado no Histograma de Euler. A Seção 2 apresenta os detalhes da elaboração e implementação do histograma proposto, a Seção 3 descreve os parâmetros metodológicos empregados na avaliação, a Seção 4 apresenta os resultados dos experimentos e por fim, a Seção 5 apresenta nossas conclusões e trabalhos futuros.

2. Implementação

O Histograma Intermediário de Euler é criado a partir de dois outros histogramas, observando tal necessidade quando da estimativa de seletividade das etapas intermediárias da multijunções espaciais. Seja H_A e H_B os dois histogramas de uma etapa de multijunção espacial, um histograma vazio H_I é construído e recebe as características originais dos *buckets*, ou seja, os limites espaciais das faces, as arestas, e os vértices de H_A ou H_B, escolhido de acordo com o predicado da próxima etapa da multijunção, sem os valores de cardinalidade originais. Na sequência, estabelece-se o conjunto $S = \{(a, b), a \in H_A, b \in H_B \mid a \cap b \neq \emptyset\}$ e processando seus elementos estima-se o valor das faces, arestas e vértices de H_I, com base nas características de cada par $(a, b) \in S$ e usando as equações descritas a seguir.

O cálculo da estimativa de seletividade para faces, arestas e vértices de H_I é realizado a partir das equações a seguir (Equações 1, 2 e 3), que preservam a estrutura do Histograma de Euler. A Equação 1 foi proposta originalmente por [Mamoulis and Papadias 2001b] e foi usada para atribuir o valor para a face, conside-

rando as adaptações feitas em [de Oliveira 2017] em relação aos valores de $\bar{\bar{a}}$ e $\bar{\bar{b}}$, devido as faces não serem alinhadas[1]. O valor f_i é usado para preencher cada face i do histograma H_I. $\bar{\bar{a}}$ e $\bar{\bar{b}}$ são as cardinalidades estimadas de a e b, conforme a área de interseção entre as mesmas nos histogramas H_A, H_B, respectivamente. É ainda usado o comprimento médio dos objetos, l_{ak}, l_{bk} e l_{ik}, em cada dimensão $k = 1..d$ dos dados espaciais, conforme definido em [de Oliveira 2017].

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

As Equações 2 e 3 utilizam o valor da face f_i, e estabelecem uma proporção baseada na interseção identificada para a face. Consistem na divisão do valor na aresta a, a_a (ou b, a_b, de acordo com o predicado da próxima etapa) pelo valor da face original a (f_a) multiplicado pelo novo valor da face intermediária f_i. O vértice v é calculado de forma análoga. Os valores de a_i e v_i são atribuídos para as aresta e vértices em H_I, respectivamente.

$$a_i = (a_a / f_a * f_i) \tag{2}$$

$$v_i = (v_a / f_a * f_i) \tag{3}$$

3. Metodologia da Avaliação

Para compor o conjunto de dados, utilizou-se datasets reais obtidos nos websites do Instituto Brasileiro de Geografia e Estatística (IBGE)[2] e do Laboratório LAPIG do Instituto de Estudos Sócio-Ambientais da UFG[3]. Os datasets são apresentados na Tabela 1, destacando informações como nome, sigla, o tipo, o valor da cardinalidade e também o tamanho do arquivo em MB no formato SHP ou *Shape File*. Nos experimentos, foram utilizadas as junções espaciais descritas na Tabela 2, destacando os datasets envolvidos e a quantidade de resultados retornados.

Tabela 1. Datasets utilizados nos experimentos

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

[1]Considerou-se neste trabalho um sistema de banco de dados, onde não seria possível que as grades dos histogramas fossem alinhas, devido à heterogeneidade dos *datasets*.

[2]https://www.ibge.gov.br

[3]www.lapig.iesa.ufg.br

Tabela 2. Junções espaciais utilizadas nos experimentos

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

Mediu-se, nos experimentos, a cardinalidade individual de cada estrutura do histograma (c_f, c_a e c_v), obtida através da soma simples do valor de cada respectiva estrutura nos *buckets* do histograma. Para medir a cardinalidade total de um histograma intermediário, ou seja, o tamanho do conjunto resultante de uma etapa de uma multijunção espacial, foi utilizada uma adaptação da Equação de Euler conforme definida para o Histograma de Euler original em [Sun et al. 2002a] e apresentada na Equação 4. Nesta equação, para cada *bucket* $i = 1..n$ do histograma, soma-se a cardinalidade na face f_i, subtrai-se a cardinalidade nas arestas a_i e soma-se a cardinalidade nos vértices v_i. O valor c resultante foi comparado com a cardinalidade esperada da junção espacial real de dois datasets.

$$c = \sum_{i=0}^{n} f_i - a_i + v_i \tag{4}$$

Além da avaliação da cardinalidade resultante para a junção espacial, foi avaliado o erro de cada estrutura individual do histograma em relação a um histograma intermediário construído a partir do conjunto resultante da junção, ou seja, mediu-se o quão distante o histograma estimado utilizando o método proposto é distinto de um histograma construído a partir do dataset resultante da junção. Utilizou-se o Erro Relativo Médio (λ), definido na Equação 5, adaptada para a estrutura do Histograma de Euler, onde I é o conjunto completo de faces, arestas ou vértices, r_i é o valor da estrutura $i \in I$ no histograma real e e_i é o valor estimado para a estrutura $i \in I$ no histograma estimado.

$$\lambda = \frac{\sum_{i \in I} |r_i - e_i|}{\sum_{i \in I} r_i} \tag{5}$$

4. Avaliação

Os experimentos consistiram da execução de cada junção espacial definida na Tabela 2, seguida da captura dos valores especificados na seção anterior. Os resultados são ilustrados por gráficos de linhas a seguir, de forma a evidenciar a comparação. Em cada gráfico, o eixo horizontal indica a junção espacial e o eixo vertical indica a métrica da comparação.

A Figura 1 apresenta a comparação das cardinalidades reais e estimadas para cada consulta de junção (A, B e C), além do erro relativo médio (D) para o Histograma Intermediário de Euler. Analisando os gráficos é possível observar que em (A, B e C) as

cardinalidades estimadas através do HIE ficaram próximas com o resultado real, exceto por alguns casos onde o resultado do HIE foi maior que o real (J5,J6,J7,J13). Devido ao tipo de calculo empregado para estimar as arestas e vértices, o erro da face foi propagado, exceto nos vértices e arestas das junções J19 e J20 do HIE que apresentou resultado menor que o real. O erro relativo médio (λ) em D apresenta a maior parte dos valores entre 0 e 2, que indicam um erro de estimativa pequeno. Algumas consultas merecem atenção nos trabalhos futuros, no entanto, para investigar a fonte dos erros das estimativas, como em J1, J2, J15, J16 e J18. O erro relativo médio foi novamente propagado das faces para as arestas e vértices, indicando que a melhoria da estimativa das faces pode auxiliar na redução do erro médio como um todo ou que equações diferentes para estimar a cardinalidade das arestas e vértices podem ser necessárias.

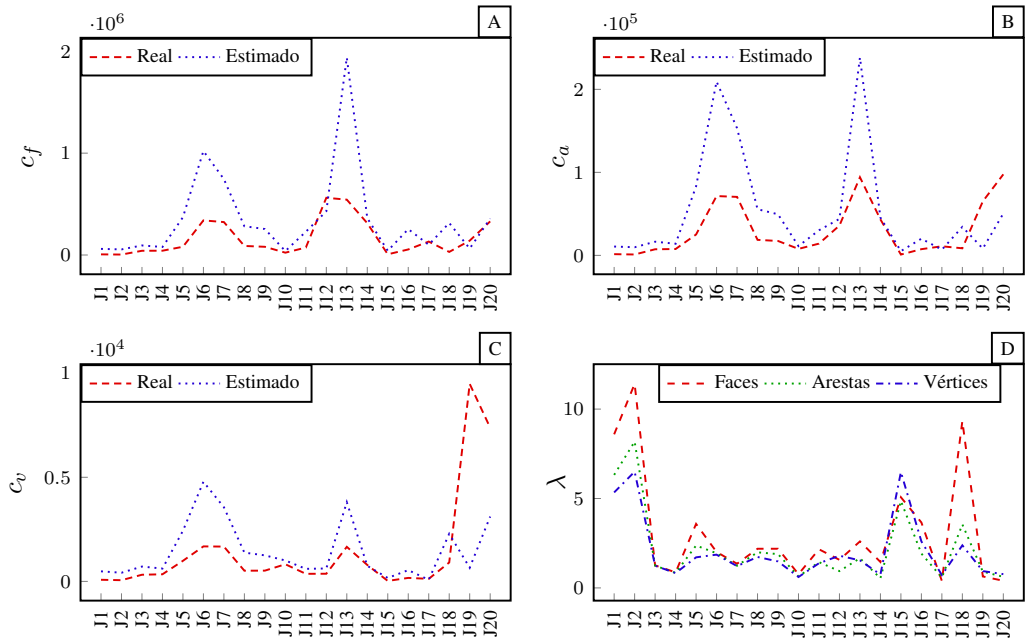

Figura 1. Comparação das cardinalidades reais e estimadas e erro relativo médio para o Histograma Intermediário de Euler. Em A, cardinalidade das faces, em B, cardinalidade das arestas, em C, cardinalidade dos vértices e em D o erro relativo médio para faces, arestas e vértices.

A estimativa da cardinalidade total de cada junção espacial foi avaliada, comparando o erro relativo entre o histograma proposto (HIE), o histograma de grade proposto em [Mamoulis and Papadias 2001a] e a cardinalidade real das junções. O resultado é apresentado na Figura 2. Pelo gráfico é possível observar que o HIE conseguiu estimar a cardinalidade com maior precisão em 13 das 20 consultas. As consultas com estimativas melhores e relevantes foram J1, J2, J5, J11 e J15. Apesar de grande parte dos resultados serem semelhantes, para as consultas onde o HIE tem uma pior estimativa a diferença é pequena. Isso indica que o método proposto é promissor e que melhorias na construção podem gerar melhores resultados.

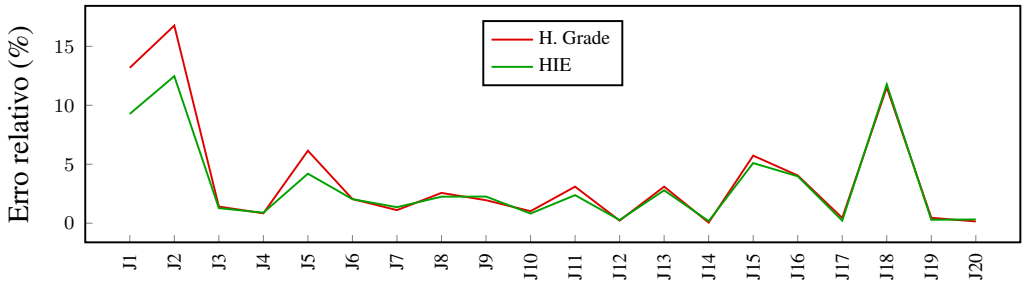

Figura 2. Comparação da cardinalidade estimada para cada junção espacial entre o Histograma de Grade e o Histograma Intermediário de Euler

5. Conclusão

Este trabalho apresentou um novo método de construção de Histogramas Intermediários (HIE) para estimativa de seletividade de consultas de multijunção espacial, baseando-se nas técnicas propostas para o Histogramas de Euler e considerando datasets cuja extensão espacial não se alinha, ou seja, um cenário real para banco de dados espaciais.

O HIE conseguiu estimar a cardinalidade com maior precisão em 13 das 20 consultas analisadas, comparado ao método mais frequentemente referenciado na literatura, o Histograma de Grade. Apesar da diferença ser pequena entre os métodos para algumas consultas, para as consultas onde o HIE tem uma pior estimativa a diferença é pequena. Isso indica que o método proposto é promissor e que melhorias na construção podem melhorar os resultados.

Como trabalhos futuros, deve-se investigar as técnicas propostas em [de Oliveira 2017] para aprimorar as estimativas das faces, considerando os tipos de objetos nos datasets. Também deve-se comparar o método proposto com o histograma intermediário IHWAF, proposto no mesmo trabalho, que usa uma técnica distinta das aqui apresentadas para lidar com o problema da contagem múltipla, chamada de sobreposição proporcional. Tal comparação não foi apresentada neste trabalho devido a necessidade de implementação de novas estruturas e equações no código[4], o que pretende-se fazer no futuro.

Referências

Brinkhoff, T., K. H.-P. and Seeger, B. (2006). *"Parallel processing of spatial joins using R-trees"*. ICDE, pages 258–265. IEEE.

de Oliveira, T. B. (2017). *Efficient Processing of Multiway Spatial Join Queries in Distributed Systems*. PhD thesis, Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brasil.

de Oliveira, T. B., Costa, F. M., and Rodrigues, V. J. S. (2017). Distributed Execution Plans for Multiway Spatial Join Queries using Multidimensional Histograms. *Journal of Information and Data Management*, 7(3):199–214.

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

Mamoulis, N. and Papadias, D. (2001a). *Advances in Spatial and Temporal Databases*, volume 2121 of *Lecture Notes in Computer Science*, chapter Selectivity Estimation of Complex Spatial Queries, pages 155–174. Springer.

Mamoulis, N. and Papadias, D. (2001b). Multiway Spatial Joins. *ACM Transactions on Database Systems*, 26(4):424–475.

Sun, C., Agrawal, D., and El Abbadi, A. (2002a). Exploring spatial datasets with histograms. In *Proceedings 18th International Conference on Data Engineering*, pages 93–102, Washington, DC, USA. IEEE.

Sun, C., Agrawal, D., and El Abbadi, A. (2002b). Selectivity estimation for spatial joins with geometric selections. In *International Conference on Extending Database Technology*, pages 609–626, Prague, Czech Republic. Springer.