Similaridade Raster de Polígonos

[1]Léo Antunes, [1,2]Leonardo Guerreiro Azevedo

[1]Departamento de Informática Aplicada – DIA
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Avenida Pasteur, 458 - Urca - Rio de Janeiro - RJ - Brasil
[2]Núcleo de Pesquisa e Prática em Tecnologia – NP2Tec

{leo.antunes,azevedo}@uniriotec.br

Abstract. *The similarity concept is fundamental for learning, knowledge and thought. Many science areas develop their own notions of similarity. This work presents an algorithm to compute similarity between polygons through their Four-Color Raster Signatures (4CRS). Implementations required for this algorithm were performed in SECONDO, an extensible DBMS platform. Experimental tests were executed in order to evaluate algorithm precision with respect to the algorithm to compute similarity through real polygons.*

Resumo. *O conceito de similaridade é fundamental para a aprendizagem, conhecimento e reflexão. Muitas áreas da ciência têm suas próprias noções de similaridade. Este trabalho apresenta um algoritmo para cálculo da similaridade entre polígonos a partir de suas Assinaturas Raster de Quatro Cores (4CRS). As implementações necessárias para este algoritmo foram realizadas no SECONDO, um banco de dados extensível. Testes experimentais foram realizados a fim de avaliar a precisão do algoritmo em relação ao cálculo da similaridade utilizando os próprios polígonos.*

1. Introdução

Segundo Quine [1969] e Cakmakov e Celakoska [2004], o conceito de similaridade é fundamental para a aprendizagem, conhecimento e reflexão. Uma métrica de similaridade fornece uma medida de semelhança entre pares de coisas que permite identificar a que classes elas pertencem. Cakmakov e Celakoska [2004] apresentam que muitos autores concordam que diferentes áreas da ciência desenvolvem suas próprias noções para similaridade, e que avaliações de similaridade vão se tornando crucialmente dependentes do contexto em que elas ocorrem.

Áreas de aplicação para uso da similaridade incluem: bancos de dados de imagens médicas, reconhecimento de gestos/movimentos humanos, sistemas de informações geológicas, comércio eletrônico, proteção de direitos autorais, design gráfico, arte criativa etc. [Sako e Fujimura, 2000]. A similaridade espacial pode ser vista como um caso particular da similaridade na qual as entidades que estão sendo comparadas possuem componentes espaciais [Holt, 2003]. Este trabalho trata da similaridade entre objetos espaciais.

Dados espaciais consistem de pontos, linhas, regiões, retângulos, superfícies e volumes [Samet, 1990]. São exemplos de dados espaciais: cidades, rios, estradas, países, estados, áreas de plantio, cadeias de montanhas etc. Freqüentemente, atributos espaciais aparecem associados com atributos não espaciais. Exemplos de dados não espaciais são: nomes de estradas, endereços, números de telefone, nomes de cidades etc.

[Azevedo *et al*., 2006]. Segundo [Güting, 1994], Sistemas Gerenciadores de Banco de Dados Espaciais (SGBDE) provêem a tecnologia de banco de dados fundamental para Sistemas de Informações Geográficas (SIG) e outras aplicações.

Uma questão principal na área de banco de dados é o processamento eficiente de consultas. Existem muitos casos onde uma resposta rápida pode ser mais importante para o usuário do que receber uma resposta exata. O processamento aproximado de consultas surgiu como uma alternativa para o processamento de consultas em ambientes nos quais o fornecimento de uma resposta exata pode demandar muito tempo. O objetivo é fornecer uma resposta estimada em ordem de magnitude de tempo menor do que o tempo necessário para computar uma resposta exata, evitando ou minimizando o número de acessos a disco ao banco de dados [Gibbons *et al*. 1997]. Hellerstein *et al.* [1997] enfatizam o uso de processamento aproximado de consultas em Sistemas de Apoio à Decisão onde tornam-se necessárias técnicas de apresentação dos dados acumulados para os tomadores de decisão em tempo hábil. Eles também propõem o uso de processamento aproximado de consultas durante uma sequência de consultas *drill-down* em mineração de dados *ad-hoc*, onde as consultas anteriores na sequência são usadas somente para determinar quais são as consultas de interesse. Papadias *et al*. [2001] propõem o processamento aproximado de consultas para OLAP (Online Analytical Processing) espacial.

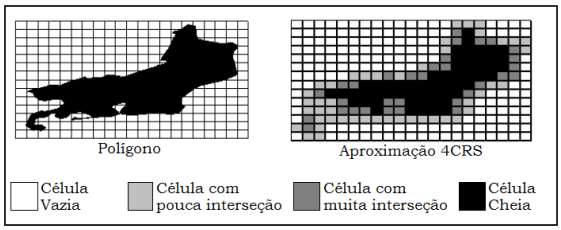
Uma das formas de executar processamento aproximado de consultas é utilizar aproximações dos dados ao invés de utilizar dados reais. [Zimbrão e Souza, 1998] propuseram a Assinatura Raster de Quatro Cores (4CRS – Four-Colour Raster Signature) para aproximar objetos espaciais descritos como polígonos.

Este trabalho propõe um algoritmo para calcular a similaridade entre polígonos através de suas assinaturas 4CRS. O algoritmo retorna um valor entre 0 e 1, indicando a semelhança aproximada dos objetos. Para computar a similaridade, o algoritmo utiliza outros algoritmos já propostos na literatura: algoritmo para calcular a área aproximada de polígonos [Azevedo *et al*., 2004] e algoritmo para calcular a área aproximada de interseção entre polígonos [Azevedo *et al*., 2005]. Além destes algoritmos, também é necessário calcular a assinatura correspondente à união de duas assinaturas 4CRS. Este algoritmo foi proposto neste trabalho e corresponde a uma contribuição do mesmo. Outra contribuição é a implementação destes algoritmos no SECONDO, o qual é um SGBD extensível que suporta principalmente tipos de dados não-convencionais como, por exemplo, dados espaciais [Güting *et al*., 2005]. A fim de avaliar a precisão do algoritmo, foram realizados testes experimentais utilizando conjuntos de dados compostos por polígonos de municípios da região norte do Brasil.

Este trabalho está dividido da seguinte forma. A Seção 1 corresponde a presente introdução. A Seção 2 apresenta os conceitos relacionados à Assinatura 4CRS. A Seção 3 apresenta a proposta de algoritmo para cálculo da similaridade entre polígonos, os algoritmos utilizados pela proposta para cálculo da área aproximada de polígonos e para cálculo da área de interseção aproximada de polígonos, além da nova proposta de algoritmo para cálculo da união de assinaturas 4CRS. A Seção 4 apresenta as implementações realizadas no SECONDO, os testes experimentais realizados e as análises dos resultados. Finalmente, a Seção 5 apresenta a conclusão do trabalho.

2. Assinatura Raster de Quatro Cores (4CRS)

A Assinatura Raster de Quatro Cores (Four-Colour Raster Signature - 4CRS) foi proposta por Zimbrao e Souza [1998]. A 4CRS armazena as principais características dos dados em uma representação aproximada e compacta que pode ser acessada e processada mais rapidamente do que os dados reais. A assinatura corresponde a uma grade de células (Figura 1) onde cada célula armazena informação relevante do objeto utilizando poucos bits. A escala da grade pode ser modificada a fim de obter uma representação mais compacta (menor escala) ou com maior precisão (maior escala).



Tipo	Descrição
Vazio (00)	A célula não tem interseção com o polígono.
Pouco (01)	A célula tem uma interseção de 50% ou menos com o polígono.
Muito (10)	A célula tem uma interseção de mais de 50% e menos de 100% com o polígono.
Cheio (11)	A célula é totalmente ocupada pelo polígono.

Figura 1 – Exemplo de assinatura 4CRS [Azevedo *et al*., 2004] seus tipos de célula.

A assinatura 4CRS apresenta bons resultados para aproximação de polígonos em relação ao processamento exato de consultas, como apresentado por [Zimbrao e Souza, 1998]. Isto motivou a avaliação do uso da 4CRS para processamento aproximado de consultas e um conjunto de algoritmos foram propostos e implementados por Azevedo *et al*. [2004, 2005 e 2006] e demonstraram bons resultados.

No processamento aproximado de consultas utilizando a 4CRS, ao invés de utilizar o objeto real, utiliza-se a assinatura 4CRS do objeto para realizar a consulta. Como resultado, uma resposta aproximada é retornada, juntamente com um intervalo de confiança. Como um exemplo, no cálculo da área aproximada de um polígono *p* [Azevedo *et al*., 2004], o algoritmo calcula um valor aproximado de área *v*, e um intervalo de confiança *i*, considerando um percentual de confiança *c*. Portanto, a resposta final é que a área real é um valor entre *v-i* e *v+i*, com confiança igual a *c*.

3. Algoritmos para cálculo da similaridade raster

Uma função de similaridade, de forma intuitiva, retorna a similaridade de objetos em termos de tamanho, forma e localização. Um exemplo, para objetos espaciais que possuem área, é calcular a similaridade a partir da razão entre as suas áreas de interseção e de união (Equação 1). Esta equação é uma medida intuitiva e conhecida como índice de Jaccard [Jaccard, 1912], como apontado por Hemert e Baldock [2007]. Yanchi *et al.* [2009] também apresentam esta equação com o intuito de definir uma medida de distância espacial.

$S\ (o1, o2) = \frac{A_{\cap}\ (o1,o2)}{A_{\cup}\ (o1,o2)}$, *onde:*

o1 e o2: objetos espaciais que possuem área

$A_{\cap}$: *área de interseção aproximada entre polígonos*

$A_{\cup}$: *área aproximada da união entre polígonos*

Equação 1 – Equação para calcular a similaridade entre polígonos

Este trabalho propõe o cálculo da similaridade de polígonos a partir da razão entre a área de interseção e a área de união de suas assinaturas 4CRS. Dessa forma, na Equação 1, os objetos *o1* e *o2* são substituídos pelas assinaturas raster dos polígonos.

O algoritmo para calcular a similaridade (Figura 2) recebe como parâmetros duas assinaturas 4CRS de dois polígonos e retorna um valor no intervalo fechado [0,1], que indica o quanto eles são similares. Quanto mais este valor se aproxima de 1, mais similares são os objetos. O algoritmo usa algoritmos para calcular: área de interseção aproximada entre polígonos, união de assinaturas 4CRS e área aproximada de polígono.

```
1. REAL similar(assinat4CRS1, assinat4CRS2)
2.  REAL areaInter = areaIntersecaoAproximada(assinat4CRS1, assinat4CRS2);
3.  SE(areaInter == 0) /* Não existe área de interseção */
4.   RETORNA 0;
5.  SENAO /* Existe área de interseção */
6.   ASSINAT4CRS assinatUniao = uniaoAssinat4CRS(signat4CRS1, signat4CRS2);
7.   REAL areaUniao = areaAproximada(assinatUniao);
8.  RETORNA areaInter / areaUniao;
```

Figura 2 – Similaridade entre polígonos.

O algoritmo para calcular a área aproximada de um polígono a partir de sua assinatura 4CRS foi proposto por Azevedo *et al.* [2004] e é apresentado na Figura 3. O algoritmo soma a área estimada do polígono dentro de cada célula da grade, contando o número de células de cada tipo (Vazio, Pouco, Muito ou Cheio) na assinatura e multiplicando pela média da área do polígono de acordo com cada tipo de célula.

```
1. REAL areaAproximada(assinat4CRS)
2.   numCelulasPouco = numCelulasMuito = numCelulasCheio = 0;
3.   areaCelula = assinat4CRS.tamanhoAresta * assinat4CRS.tamanhoAresta;
4.   PARA CADA celula EM assinat4CRS.celulas FACA
5.      SE (celula.tipo == POUCO)
6.         numCelulasPouco++;
7.      SENAO SE (celula.tipo == MUITO)
8.         numCelulasMuito++;
9.      SENAO SE (celula.tipo == CHEIO)
10.        numCelualasCheio++;
11.  RETORNA (numCelulasPouco * pesoCelulasPouco + numCelulasMuito *
        pesoCelulasMuito + numCelulasCheio * pesoCelulasCheio) * areaCelula;
```

Figura 3 - Área aproximada de polígono [Azevedo *et al.*, 2004].

O algoritmo para calcular a área de interseção aproximada entre dois polígonos a partir de suas assinaturas 4CRS foi proposto por Azevedo *et al.* [2005] e é apresentado na Figura 4. Ele computa a soma das áreas estimadas das células das assinaturas [Azevedo *et al.*, 2004, 2005] que se sobrepõem, e multiplica o resultado pela área da célula. Como existem quatro tipos diferentes de células (Vazio, Pouco, Muito ou Cheio), as possibilidades de sobreposição são dezesseis.

```
1. REAL areaIntersecaoAproximada(assinat4CRS1, assinat4CRS2)
2.   areaAproximada = 0;
3.   MBRIntersecao = calculaMBRIntersecao(assinat4CRS1, assinat4CRS2);
4.   SE (assinat4CRS1.tamanhoAresta == assinat4CRS2.tamanhoAresta)
5.      menor4CRS = assinat4CRS1;
6.      maior4CRS = assinat4CRS2;
7.   SENAO
9.      menor4CRS = assMenorTamanho(assinat4CRS1, assinat4CRS2);
10.     maior4CRS = assMaiorTamanho(assinat4CRS1, assinat4CRS2);
11.  areaAproximada = 0;
12.  PARA CADA celula b de maior4CRS que está contida em MBRIntersecao FACA
13.     PARA CADA celula s de menor4CRS que está contida em celula b FACA
14.        areaAproximada += areaEstimada[s.tipo,b.tipo];
15.  areaCelula = menor4CRS.tamanhoAresta * menor4CRS.tamanhoAresta;
16.  RETORNA areaAproximada * areaCelula;
```

Figura 4 - Área de interseção aproximada de polígonos [Azevedo *et al.*, 2005].

O algoritmo para calcular a assinatura resultante da união de duas assinaturas 4CRS também corresponde a uma contribuição deste trabalho e é apresentado na Figura 5. O cálculo é realizado da seguinte forma: se não existe MBR (Minimum Bounding Rectangle – menor retângulo que envolve objeto) de interseção entre as assinaturas, então é retornado NULL. Por outro lado, quando existe MBR de interseção, uma nova assinatura 4CRS é criada considerando o MBR de união dos MBRs das assinaturas.

```
1. ASSINAT4CRS uniaoAssinat4CRS(assinat4CRS1, assinat4CRS2)
2.   SE existeIntersecao(assinat4CRS1, assinat4CRS2)
3.     SE (assinat4CRS1.tamanhoAresta > assinat4CRS2.tamanhoAresta)
4.       maior4CRS = assinatat4CRS1;
5.       menor4CRS = mudarResolucao(assinat4CRS2,
                     assinat4CRS1.tamanhoAresta);
6.     SENAO
7.       maior4CRS = assinat4CRS2;
8.       menor4CRS = mudarResolucao(assinat4CRS1,
                     assinat4CRS2.tamanhoAresta);
9.       MBRUniao = calculaMBRUniao(menor4CRS.MBR, maior4CRS.MBR)
10.      /* Cria assinatura 4CRS com células Vazio */
11.      n4CRS = criaAssinatura(MBRUniao, maior4CRS.tamanhoAresta, VAZIO);
12.      PARA CADA b em maior4CRS.cels interceptando n em n4CRS.cels FACA
13.        n.tipo = b.tipo;
14.        PARA CADA s em menor4CRS.cels interceptando n em n4CRS.cels FACA
15.          SE n.tipo == VAZIO OU s.tipo == CHEIO
16.            n.tipo = s.tipo;
17.          SENAO SE n.tipo == POUCO E s.tipo == MUITO
18.            n.tipo = s.tipo;
20.       RETORNA n4CRS;
21.  SENAO
22.     RETORNA NULL;
```

Figura 5 – União entre duas assinaturas 4CRS.

Ao executar uma consulta que retorna resultados aproximados, é importante mostrar para o usuário um intervalo de confiança para a resposta obtida a fim de que ele possa decidir se a precisão é suficiente. A partir dos cálculos de intervalo de confiança propostos por Azevedo *et al.* [2004, 2005], neste trabalho, propomos uma forma de cálculo do intervalo de confiança para o algoritmo de similaridade (Equação 2).

$$IC_{Similaridade} = \left[\frac{A_{\cap} - \Delta_{IC_{\cap}}}{A_{\cup} + \Delta_{IC_{\cup}}}, \frac{A_{\cap} + \Delta_{IC_{\cap}}}{A_{\cup} - \Delta_{IC_{\cup}}}\right], \textit{onde:}$$

$A_{\cap}$: *área de interseção aproximada entre polígonos*
$A_{\cup}$: *área aproximada da união entre polígonos*
$\Delta_{IC_{\cap}}$: *variância do intervalo de confiança do algoritmo que calcula a área de interseção aproximada*
$\Delta_{IC_{\cup}}$: *variância do intervalo de confiança do algoritmo a área de união aproximada*

Equação 2 – Cálculo do intervalo de confiança para similaridade raster.

Um exemplo de cálculo é demonstrado a seguir. Considere que o algoritmo seja executado em dois polígonos quaisquer e obtenha os seguintes dados: $A_{\cap}$: $4{,}95 \times 10^6$; $A_{\cup}$: $1{,}27 \times 10^7$; $\Delta_{IC_{\cap}}$: $2{,}37 \times 10^5$; e, $\Delta_{IC_{\cup}}$: $2{,}45 \times 10^5$. Aplicando a fórmula para cálculo do intervalo de confiança na resposta do algoritmo que calcula a similaridade entre polígonos, vamos obter o seguinte resultado:

$$IC_{Similaridade} = \left[\frac{4{,}95 \times 10^6 - 2{,}37 \times 10^5}{1{,}27 \times 10^7 + 2{,}45 \times 10^5}; \frac{4{,}95 \times 10^6 + 2{,}37 \times 10^5}{1{,}27 \times 10^7 - 2{,}45 \times 10^5}\right]$$

$$IC_{Similaridade} = \left[\frac{4{,}72 \times 10^6}{1{,}30 \times 10^7}; \frac{5{,}19 \times 10^6}{1{,}25 \times 10^7}\right] = [0{,}364\ ;\ 0{,}416]$$

Figura 6 – Exemplo de cálculo de intervalo de confiança.

4. Avaliação experimental

4.1. Implementação dos algoritmos

Os algoritmos apresentados neste trabalho foram implementados no SECONDO. O SECONDO [Dieker e Güting, 2000; Güting *et al.*, 2005] é um ambiente genérico que suporta a implementação de sistemas de banco de dados para grande número de modelos de dados e linguagens de consulta. Tipos de dados, construtores desses tipos de dados e operadores são implementados em álgebras [Güting, 1993]. Os operadores de similaridade, área aproximada, área de interseção aproximada e união de assinaturas 4CRS foram implementados na álgebra Raster, correspondente à assinatura 4CRS.

4.2. Testes experimentais

Foram realizados testes experimentais com o objetivo de avaliar a precisão da resposta do algoritmo para cálculo aproximado da similaridade em relação à resposta exata.

Nos testes experimentais, foram utilizados polígonos que representam municípios da região norte do Brasil (*BRNorte*). Com o objetivo de obter outro conjunto de dados com sobreposição com *BRNorte,* para avaliarmos o operador de similaridade, os polígonos originais de *BRNorte* foram deslocados aleatoriamente nas coordenadas *x* e *y*, seguindo a proposta de Brinkhoff *et al.* [1994], gerando o conjunto de dados *BRNorteT*. A partir destes dois conjuntos foram geradas as assinaturas 4CRS.

Em seguida, foram calculadas a similaridade a partir dos objetos reais (utilizando os valores exatos da área de interseção e área de união), e a similaridade a partir de suas respectivas assinaturas 4CRS (utilizando os valores aproximados da área de interseção e da área de união). Os resultados obtidos são apresentados considerando erros abaixo de 10% e erros acima de 10%. Devido a limitações de espaço para colocar todos os resultados em uma única tabela, os resultados correspondentes a cada caso são apresentados parcialmente. A Tabela 1 e a Tabela 2 apresentam os resultados com erros acima de 10%, enquanto que a Tabela 3 e a Tabela 4 apresentam os resultados com erros abaixo de 10%. Os rótulos das colunas da Tabela 1, Tabela 2, Tabela 3 e Tabela 4 são: (a) ***ID1, ID2***: identificadores dos objetos pertencentes à *BRNorte* e *BRNorteT*; (b) ***SB1, SB2***: tamanhos da célula das assinaturas; (c) ***AIR, AIA***: área de interseção real e de interseção aproximada; (d) ***%EAIA***: porcentagem de erro da área de interseção aproximada; (e) ***MIN, MAX***: valor inferior e superior fornecido pelo intervalo de confiança da área de interseção aproximada; (f) ***NC***: número de células com interseção entre as assinaturas em potência de dois (ou seja, 2^n), desconsiderando células *Vazio*; (g) ***AUR, AUA***: área de união real e aproximada; (h) ***MIN2, MAX2***: valor inferior e superior fornecido pelo intervalo de confiança da área de união aproximada; (i) ***%EAUA***: porcentagem de erro da área de união aproximada; (j) ***SRaster***: similaridade raster (percentual); (l) ***MIN3, MAX3***: valor inferior e superior fornecido pelo intervalo de confiança da similaridade raster; (m) ***SReal***: similaridade real (percentual); (n) ***%ES***: Porcentagem de erro da similaridade raster (calculada de acordo com a Equação 3); (o) ***%DE***: porcentagem referente à diferença absoluta entre a similaridade real e a similaridade raster (*SRaster – SReal*).

$$\%ES = \frac{|\ SRaster - SReal\ |}{SReal}$$

Equação 3 – Equação para calcular o erro de cálculo da similaridade.

Tabela 1 – Tabela de resultados com erros acima de 10% (primeira parte)

	ID1	ID2	SB1	SB2	AIR	AIA	%EAIA	MIN	MAX	NC
1	**234**	**397**	**256**	**1024**	**124,011**	**312895**	**252212,29%**	**283097**	**342693**	**2**
2	**10**	**25**	**1024**	**256**	**486,325**	**1,05E+06**	**214822,12%**	**137623**	**1,95E+06**	**4**
3	274	164	512	256	1315,56	21286,1	1518,03%	-18765,4	61337,6	2
4	28	32	512	512	28746,6	248775	765,41%	6657,73	490892	10
8	188	167	1024	256	75920,8	199020	162,14%	-152650	550689	2
9	139	38	256	512	9513,49	21286,1	123,75%	11273,2	31299	2
10	**2**	**5**	**1024**	**512**	**5,14E+06**	**1,07E+07**	**108,93%**	**7,90E+06**	**1,36E+07**	**26**
11	**140**	**146**	**1024**	**1024**	**613326**	**1,26E+06**	**105,12%**	**325925**	**2,19E+06**	**9**
12	301	414	512	1024	441836	774898	75,38%	655461	894335	3

Tabela 2 – Tabela de resultados com erros acima de 10% (segunda parte)

	ID1	ID2	AUR	AUA	MIN2	MAX2	%EAUA	SRaster	MIN3	MAX3	SReal	%ES	%DE
1	**234**	**397**	**5,53E+07**	**5,43E+07**	**5,09E+07**	**5,76E+07**	**1,96%**	**0,58%**	**0,49%**	**0,67%**	**0,00%**	**257252,94%**	**0,58%**
2	**10**	**25**	**4,21E+07**	**4,17E+07**	**3,82E+07**	**4,52E+07**	**1,04%**	**2,51%**	**0,30%**	**5,11%**	**0,00%**	**217074,82%**	**2,51%**
3	274	164	2,82E+07	2,82E+07	2,72E+07	2,92E+07	0,01%	0,08%	-0,06%	0,23%	0,00%	1518,24%	0,07%
4	28	32	3,50E+07	3,53E+07	3,42E+07	3,63E+07	0,70%	0,71%	0,02%	1,43%	0,08%	759,35%	0,62%
8	188	167	1,79E+08	1,81E+08	1,77E+08	1,86E+08	1,33%	0,11%	-0,08%	0,31%	0,04%	158,70%	0,07%
9	139	38	1,87E+07	1,80E+07	1,69E+07	1,90E+07	3,89%	0,12%	0,06%	0,19%	0,05%	132,80%	0,07%
10	**2**	**5**	**6,87E+07**	**6,87E+07**	**6,51E+07**	**7,22E+07**	**0,04%**	**15,63%**	**10,93%**	**20,84%**	**7,48%**	**108,86%**	**8,15%**
11	**140**	**146**	**1,33E+08**	**1,32E+08**	**1,27E+08**	**1,36E+08**	**1,32%**	**0,96%**	**0,24%**	**1,72%**	**0,46%**	**107,88%**	**0,50%**
12	301	414	9,02E+07	9,10E+07	8,69E+07	9,50E+07	0,83%	0,85%	0,69%	1,03%	0,49%	73,94%	0,36%

Tabela 3 – Tabela de resultados com erros abaixo de 10% (primeira parte)

	ID	IDT	SB	SBT	EIA	AIA	%EAIA	MIN	MAX	NC
51	68	158	512	512	2820430	3092300	9,64%	2425870	3758730	29
52	**188**	**379**	**1024**	**256**	**1,1E+07**	**12765200**	**11,24%**	**11089900**	**1,4E+07**	**22**
53	155	38	256	512	843964	882482	4,56%	771151	993812	12
54	4	3	512	512	457746	498283	8,86%	44088,3	952478	9
56	163	38	256	512	1074720	921305	14,27%	805981	1036630	12
57	79	40	128	256	389058	424844	9,20%	394579	455109	16
58	**153**	**97**	**512**	**512**	**4785020**	**4490630**	**6,15%**	**3831560**	**5149710**	**35**
59	7	2	512	1024	1770680	1837840	3,79%	1436140	2239540	9
89	280	209	128	256	936733	940258	0,38%	904778	975739	31
90	**24**	**14**	**1024**	**1024**	**5,3E+07**	**52254500**	**1,96%**	**49210300**	**5,5E+07**	**80**
91	14	14	1024	1024	3,7E+07	36984400	0,87%	33747200	4E+07	61
92	75	32	256	512	1146480	1181850	3,09%	1077500	1286200	13
95	9	9	1024	1024	4E+07	39669600	0,36%	36794700	4,3E+07	60
96	434	419	1024	512	1,5E+07	14089500	3,29%	12210000	1,6E+07	28
97	**188**	**188**	**1024**	**1024**	**1,2E+08**	**124048000**	**0,24%**	**1,2E+08**	**1,3E+08**	**163**
98	129	228	256	128	2080950	2136570	2,67%	1982730	2290410	52

Tabela 4 – Tabela de resultados com erros abaixo de 10% (segunda parte)

	ID1	ID2	AUR	AUA	MIN2	MAX2	%EAUA	SRaster	MIN3	MAX3	SReal	%ES	%DE
51	68	158	3,45E+07	3,44E+07	3,34E+07	3,54E+07	0,15%	8,99%	6,85%	11,25%	8,19%	9,80%	0,80%
52	**188**	**379**	**1,72E+08**	**1,74E+08**	**1,70E+08**	**1,78E+08**	**1,46%**	**7,33%**	**6,23%**	**8,49%**	**6,69%**	**9,64%**	**0,64%**
53	155	38	1,46E+07	1,40E+07	1,31E+07	1,49E+07	4,52%	6,32%	5,19%	7,61%	5,77%	9,52%	0,55%
54	4	3	3,84E+07	3,83E+07	3,71E+07	3,94E+07	0,45%	1,30%	0,11%	2,56%	1,19%	9,35%	0,11%
56	163	38	1,74E+07	1,63E+07	1,53E+07	1,72E+07	6,38%	5,67%	4,69%	6,77%	6,19%	8,43%	0,52%
57	79	40	8,63E+06	8,72E+06	8,48E+06	8,96E+06	1,05%	4,87%	4,41%	5,37%	4,51%	8,06%	0,36%
58	**153**	**97**	**3,35E+07**	**3,41E+07**	**3,31E+07**	**3,52E+07**	**1,97%**	**13,15%**	**10,89%**	**15,56%**	**14,29%**	**7,97%**	**1,14%**
59	7	2	6,99E+07	6,74E+07	6,37E+07	7,10E+07	3,56%	2,73%	2,02%	3,51%	2,53%	7,62%	0,19%
89	280	209	7,49E+06	7,63E+06	7,43E+06	7,84E+06	1,97%	12,32%	11,54%	13,13%	12,51%	1,56%	0,20%
90	**24**	**14**	**1,42E+08**	**1,41E+08**	**1,38E+08**	**1,45E+08**	**0,45%**	**36,98%**	**33,94%**	**40,18%**	**37,55%**	**1,51%**	**0,57%**
91	14	14	2,21E+08	2,22E+08	2,17E+08	2,26E+08	0,31%	16,70%	14,96%	18,50%	16,90%	1,18%	0,20%
95	9	9	1,87E+08	1,89E+08	1,85E+08	1,93E+08	0,88%	21,02%	19,09%	23,03%	21,13%	0,52%	0,11%
96	434	419	9,08E+07	8,81E+07	8,47E+07	9,15E+07	2,99%	16,00%	13,35%	18,86%	16,05%	0,30%	0,05%
97	**188**	**188**	**2,17E+08**	**2,18E+08**	**2,14E+08**	**2,22E+08**	**0,51%**	**56,88%**	**54,01%**	**59,85%**	**57,03%**	**0,28%**	**0,16%**
98	129	228	6,63E+06	6,80E+06	6,60E+06	7,00E+06	2,54%	31,42%	28,34%	34,70%	31,38%	0,13%	0,04%

4.3. Análises dos resultados

Como o cálculo da similaridade raster é feito pela razão das áreas aproximadas de interseção e de união, é importante analisar quanto estes valores influenciam na precisão do resultado. Na Figura 7, o eixo Y apresenta os percentuais de erro, enquanto que o eixo X apresenta os objetos ordenados do maior erro para o menor erro. O erro da área de união é relativamente baixo, enquanto que o erro da similaridade raster acompanha o erro da área de interseção. Quanto menor o erro da área de interseção, menor o erro da similaridade raster.

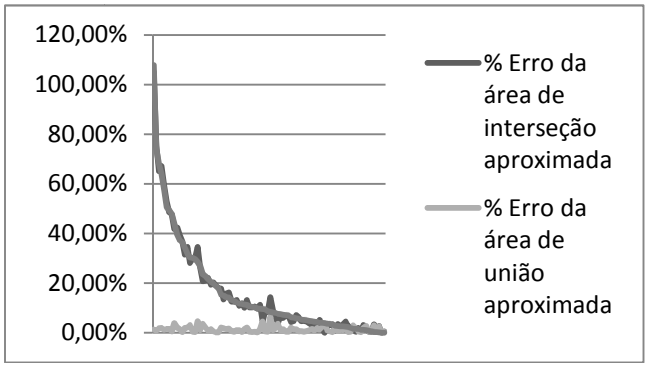


Figura 7 – Percentuais de erro da área de interseção, área de união e similaridade.

Analisando o pior resultado do algoritmo, destacado na linha 1 da Tabela 1 e da Tabela 2 e representado pelos objetos apresentados na Figura 8.a, podemos observar que a interseção entre os objetos é muito pequena. Somente duas células têm interseção, como apresentado na coluna *NC*. Neste caso, a similaridade raster é de 0,58% e a similaridade real é aproximadamente igual a 0% (Tabela 2). Logo, o erro na estimativa de similaridade (%ES) foi muito grande (257.252,94%). Conseqüentemente, usar a aproximação para estimar a similaridade neste caso se torna muito ruim. Quando a similaridade é muito pequena, pode ser mais vantajoso realizar a consulta exata. O mesmo acontece no segundo caso, destacado na linha 2 da Tabela 1 e da Tabela 2 e representado pela Figura 8.b.

É interessante notar também que nos dois exemplos ocorre uma mudança de escala de 2^8 para 2^{10}, ocasionando uma perda de informação. A mudança de escala é feita agrupando-se um conjunto de células de tamanho menor para representar uma célula de tamanho maior, somando pesos para cada tipo de célula de maneira pessimista. Isto leva a maior perda na precisão do algoritmo.

(a) Obj. 234 x Obj. 397 (b) Obj. 10 x Obj. 25 (c) Obj. 2 x Obj. 5 (d) Obj. 140 x Obj. 146

Figura 8 – Exemplos de interseções de objetos das Tabela 2 e Tabela 3.

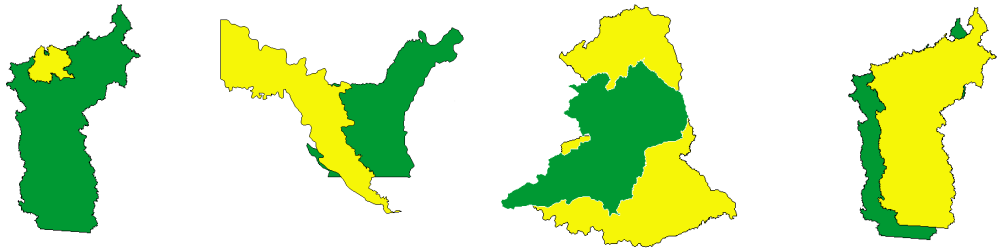
Analisando outro resultado ruim, destacado na linha 10 das Tabela 1 e Tabela 2 e representado pelos objetos da Figura 8.c, o percentual de erro (%ES) foi de 108,86%. A similaridade raster é igual a 15,63% e a similaridade real é igual a 7,48% (Tabela 2).

Neste caso, o número de células com interseção (*NC*) é muito pequeno e houve mudança de escala do fator de 2^9 para 2^{10}. Analisando a quantidade de células que têm interseção de acordo com seus tipos (Tabela 5 – linha 1), observamos que não existem interseções de células do tipo *Cheio* × *Cheio*, cuja interseção tem 100% de precisão e que, caso ocorressem, poderiam levar a uma maior precisão.

Por outro lado, analisando o caso destacado na linha 11 da Tabela 1 e da Tabela 2 e representado pelos objetos da Figura 8.d, podemos observar que não há mudança de escala, mas, em compensação, o número de células que têm interseção é muito pequeno (nove células) e a similaridade raster e a similaridade real também são muito pequenas (0,98% e 0,46%, respectivamente - Tabela 2). Neste caso, o erro da similaridade raster é igual a 107,88%.

Dessa forma, observamos que são três os principais fatores que influenciaram no resultado: (i) número pequeno de células com interseção entre as assinaturas e, conseqüentemente, similaridade com valor muito baixo; (ii) maioria das células com interseção ser do tipo onde é feita uma aproximação considerando a média (*Pouco* × *Pouco*, *Pouco* × *Muito*, *Pouco* × *Cheio*, *Muito* × *Muito* ou *Muito* × *Cheio*); (iii) mudança de escala para comparação.

Por outro lado, o algoritmo retorna bons resultados em casos onde estes fatores são minimizados. Por exemplo, no caso destacado na linha 52 da Tabela 3 e da Tabela 4 e representado pelos objetos da Figura 9.a, o erro na estimativa é 9,64%. A similaridade raster é igual a 7,33% e a similaridade real é igual a 6,69%. Neste caso, o número de células com interseção também é pequeno (22 células), porém, agora existem interseções do tipo *Cheio* × *Cheio*, onde a precisão é de 100% (Tabela 5 - linha 2).



(a) Obj. 188 x Obj. 379 (b) Obj. 153 x Obj. 97 (c) Obj. 24 x Obj. 14 (d) Obj. 188 x Obj. 188

Figura 9 – Exemplos de interseções de objetos das Tabela 3 e da Tabela 4

Tabela 5 – Número de células que têm interseção (Pouco × Pouco, Pouco × Muito, Pouco × Cheio, Muito × Muito, Muito × Cheio e Cheio × Cheio)

	Objetos	P × P	P × M	P × C	M × M	M × C	C × C
1	Obj. 2 x Obj. 5	2	7	6	6	5	0
2	Obj. 188 x Obj. 379	2	1	6	3	6	4
3	Obj. 153 x Obj. 97	3	6	8	2	5	2
4	Obj. 24 x Obj. 14	5	7	18	5	16	29
5	Obj. 188 x Obj. 188	10	9	23	7	33	81

Analisando outro caso, com resultado ainda melhor, destacado na linha 58 da Tabela 3 e da Tabela 4 e representado pelos objetos da Figura 9.b, o erro na estimativa é de 7,97%, a similaridade raster é igual a 13,15% e a similaridade real é igual a 14,29%. Neste caso, observamos que existem células com interseção do tipo *Cheio* × *Cheio* (Tabela 5 - linha 3) e que o número total de células com interseção é maior (35 células).

Outro resultado muito bom é destacado na linha 90 da Tabela 3 e da Tabela 4 e representado pelos objetos da Figura 9.c. O erro na estimativa é de 1,51%, a

similaridade raster é igual a 36,98% e a similaridade real é igual a 37,55%. Neste caso, o número de células com interseção é grande (80 células) e boa parte dessas células é do tipo *Cheio* × *Cheio* (Tabela 5 - linha 4).

Um dos melhores resultados, onde um objeto é comparado com ele mesmo deslocado, é destacado na linha 97 das Tabela 3 e Tabela 4, e representado pelos objetos da Figura 9.d. O erro de estimativa nesse caso é de apenas 0,28%, a similaridade raster é igual a 56,88% e a similaridade real é igual a 57,03%. Neste exemplo, o número de células com interseção é ainda maior que no exemplo anterior (153 células) e boa parte dessas interseções de células é do tipo *Cheio* × *Cheio* (Tabela 5 - linha 5).

O erro no algoritmo apresentado na Tabela 1 e na Tabela 2 pode ser explicado pela aproximação da área de interseção entre células dos tipos *Pouco* × *Pouco*, *Pouco* × *Muito*, *Pouco* × *Cheio*, *Muito* × *Muito* e *Muito* × *Cheio.* Quando este número é muito pequeno, não podemos supor a distribuição normal. Logo, dois casos podem ocorrer. No primeiro, quando a área de interseção entre a maioria das células se aproxima da média, o resultado aproximado será bem próximo do real. Porém, no segundo caso, quando esta área se distancia da média, o resultado aproximado será bem distante do resultado exato. Observamos ainda que todos os resultados com erros acima de 10% tinham como interseção a borda dos objetos, enquanto que nos resultados com erros abaixo de 10%, os objetos tinham boa parte de suas áreas sobrepostas.

Logo, temos fortes indícios de que o que determinou a precisão do algoritmo foi a aproximação da área de interseção entre células dos tipos *Pouco* × *Pouco*, *Pouco* × *Muito*, *Muito* × *Muito* e *Muito* × *Cheio.* Dessa forma, é necessário um estudo para melhorar o algoritmo que calcula a área de interseção a partir de assinaturas raster.

Nos casos onde temos um percentual grande de sobreposição dos objetos, os resultados foram satisfatórios. Quanto maior o valor da similaridade raster, mais próximo ele está da similaridade real.

Quanto ao intervalo de confiança calculado, para os resultados com erros acima de 10%, em 70% dos casos, o valor da similaridade real (calculada a partir dos objetos reais) ficou dentro do intervalo. Em contrapartida, para os resultados com erros abaixo de 10%, em 96% dos casos, o valor da similaridade real ficou dentro do intervalo. Este intervalo pode ser consultado através dos campos *MIN3* e *MAX3* da Tabela 2 (resultados com erros acima de 10%) e da Tabela 4 (resultados com erros abaixo de 10%).

5. Conclusões

A similaridade entre objetos é um conceito fundamental. Uma métrica de similaridade fornece uma medida de semelhança entre pares de coisas que permite identificar a que classes elas pertencem. Este trabalho tem como principal contribuição a proposta de um algoritmo para calcular a similaridade entre polígonos a partir de suas assinaturas 4CRS. Outras contribuições são a proposta de algoritmo para calcular a união de duas assinaturas 4CRS e a implementação no Secondo [Güting *et al.*, 2005] destes algoritmos e dos algoritmos propostos por Azevedo *et al.* [2004, 2005] para calcular a área aproximada de um polígono e a área de interseção aproximada entre dois polígonos.

Testes experimentais foram realizados sobre dados reais de municípios do norte do Brasil a fim de avaliar a precisão do algoritmo. Os resultados obtidos demonstraram que existe uma variação da precisão do algoritmo. Os casos em que há um erro superior

a 10% têm como principais características: número pequeno de células com interseção entre as assinaturas e, conseqüentemente, similaridade com valor muito baixo; maioria das células com interseção ser do tipo onde é feita uma aproximação considerando a média ($P \times P$, $P \times M$, $P \times C$, $M \times M$ ou $M \times C$) quando a interseção real ocorre nas bordas dos objetos; e, ocorrência de mudança de escala para executar os algoritmos de interseção aproximada e união de assinaturas raster.

Os resultados para casos onde temos um percentual grande de sobreposição dos objetos foram satisfatórios. Logo, podemos afirmar que, quanto maior o valor da similaridade raster, mais próximo ele está da similaridade real. Por outro lado, existe um erro percentual grande quando o valor da similaridade raster é pequeno. Logo, se a aplicação do algoritmo está buscando encontrar objetos com similaridade alta, indicamos que assinaturas com similaridade baixa não sejam consideradas no resultado. Por outro lado, se há interesse em ter uma noção de similaridade baixa ou alta, então para os casos de similaridade baixa, indicamos que o cálculo exato seja realizado.

Em relação à fórmula para cálculo do intervalo de confiança, também proposta neste trabalho, o valor exato ficou dentro do intervalo de confiança em 96% dos casos, onde o erro na estimativa da similaridade raster era menor que 10%, o que demonstra que, para estes casos, a similaridade raster tem um bom resultado. Por outro lado, para os casos em que o erro da similaridade raster em relação à similaridade real foi superior a 10%, em 70% dos casos o valor da similaridade real ficou dentro do intervalo de confiança da similaridade raster. Apesar do valor de 70% ser razoável, consideramos que é necessário analisar o algoritmo para melhorá-lo.

Como trabalhos futuros, propomos: melhorar o algoritmo para calcular a área aproximada de interseção, pois há uma relação direta entre o erro resultante deste algoritmo em relação ao algoritmo que calcula a similaridade raster, como demonstrado no gráfico apresentado na Figura 7; realizar testes de desempenho comparando o algoritmo de similaridade raster em relação ao cálculo da similaridade real; evoluir o algoritmo para poder ser utilizado em outras situações como, por exemplo, o algoritmo permitir comparar objetos de acordo com sua forma, independente do tamanho dos mesmos e sem a necessidade de mudança de escala (por exemplo, comparação de uma maquete de um objeto, em tamanho reduzido, com um objeto original, em tamanho maior); implementar a visualização de assinaturas Raster no Secondo, o que pode auxiliar muito nas análises dos resultados obtidos.

Referências

Azevedo, L. G., Monteiro, R. S., Zimbrão, G.; Souza, J. M. (2004) “Approximate Spatial Query Processing Using Raster Signatures”. In: VI Simpósio Brasileiro de GeoInformática (GeoInfo 2004), Campos do Jordão, Brasil.

Azevedo, L. G., Zimbrão, G., Souza, J. M., Güting, R. H. (2005) “Estimating the Overlapping Area of Polygon Join”. In: International Symposium on Advances in Spatial and Temporal Databases, v. 1. p. 91-108, Angra dos Reis, Brasil.

Azevedo, L. G., Zimbrão, G., Souza, J. M. (2006) “Approximate Query Processing in Spatial Databases Using Raster Signatures”. Advances in Geoinformatics. 1 ed. Heidelberg: Springer, 2006, v. 1, p. 69-85.

Brinkhoff, T., Kriegel, H. P., Schneider, R., Seeger, B. (1994) “Multi-step Processing of Spatial Joins”. *ACM SIGMOD Record*, v. 23, n.2 (Jun), pp. 197-208.

Cakmakov, D., Celakoska, E. (2004) “Estimation of Curve Similarity Using Turning Function**”.** In: International Journal of Applied Math., vol. 15, no. 4, pp. 403-416.

Gibbons, P. B., Matias, Y., Poosala, V. (1997) “Aqua project white paper”. Technical Report, Bell Laboratories, Murray Hill, New Jersey, USA.

Güting, R. H. (1994) “An Introduction to Spatial Database Systems”. In: The International Journal on Very Large Data Bases, v. 3, n. 4 (Oct), pp. 357-399.

Güting, R. H., Almeida, V., Ansorge, D., Behr T., Ding, Z., Höse, T., Hoffmann F., Spiekermann, M. (2005) “SECONDO: An Extensible DBMS Platform for Research Prototyping and Teaching”. Demo-Paper, 21st International Conference on Data Engineering (ICDE, Tokyo, Japan), 1115-1116.

Jaccard, P. (1912) “The distribution of flora in the alpine zone”. In: The New Phytologist, vol. 11(2), pp. 37–50.

Hellerstein, J. M., Haas, P. J., Wang, H. J., 1997, “Online aggregation”. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 171-182, Tucson, Arizona, USA, May.

Hemert, J. V., Baldock, R. (2007) “Mining Spatial Gene Expression Data for Association Rules”. BIRD 2007, LNBI 4414, pp 66-76, Springer, 2007.

Holt, A. (2003) “Spatial similarity”. In: 15th Annual Colloquium of the Spatial Information Research Centre (SIRC 2003: Land, Place and Space), 1-2 December, Dunedin, New Zealand, pp. 77-80.

Papadias, D., Mamoulis, N., Theodoridis, Y. (1999) “Processing and optimization of multiway spatial joins using R-Trees”. In: Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 189-200, Philadelphia, Pennsylvania, USA, May-Jun.

Quine, W. V. (1969) Ontological Relativity and Other Essays. Columbia University Press, New York.

Sako Y., Fujimura K. (2000) Shape Similarity by Homotopic Deformation. *The Visual Computer*, *16*(1), pp. 47-61.

Samet, H. (1990) The Design and Analysis of Spatial Data Structure. Addison-Wesley Publishing Company, 1a ed., Boston, Massachusetts.

Yanchi L., Hongwei G., Xuedong G. (2009) “Analysis of Blast Furnace Cross Thermometric Based on Spatial Data Mining”**.** In: International Conference onCyber-Enabled Distributed Computing and Knowledge Discovery (Cyber-C), p.33-36.

Zimbrao, G., Souza, J. M. (1998) “A Raster Approximation for Processing of Spatial Joins”. In: Proceedings of 24rd International Conference on Very Large Data Bases, pp. 558-569.