VII Simpósio Brasileiro de Geoinformática, Campos do Jordão, Brasil, 20-23 novembro 2005, INPE, p. 213-222.

Detecção de Conglomerados Espaciais com Geometria Arbitrária

Marcelo A. Costa, Luciano R. Schrerrer, Renato M. Assunção

Departamento de Estatística – Universidade Federal de Minas Gerais (UFMG)
Caixa Postal 702 – 31270-901 – Belo Horizonte – MG – Brasil

{azevedo,assuncao}@est.ufmg.br, luscherrer@hotmail.com

Abstract. *This paper evaluates a variation of the spatial scan statistic that aggregates the neighbor structure into the cluster growing process. This proposal allows the detection of arbitrarily shaped clusters and it is an alternative to the original circular shaped scan geometry. The spatial scan statistic is widely used and presents consistent results with real and simulated data sets. In order to test the method, star shaped simulations were performed as well as with circular clusters and with real data set. Two heuristics were also evaluated to avoid oversized clusters and decrease computational cost.*

Resumo. *Este artigo propõe uma análise de uma variação do método scan de detecção de conglomerados espaciais na qual é agregada a estrutura de vizinhança espacial ao processo de crescimento e busca de conglomerados. Este procedimento possibilita a detecção de conglomerados de geometria arbitrária uma vez que o método scan foi originalmente proposto para a detecção de conglomerados de geometria circular e apresenta alto poder neste contexto. Uma avaliação do poder de detecção do método para conglomerados com geometria arbitrária é realizada via dados simulados e reais. Restrições durante o processo de crescimento são sugeridas para evitar conglomerados de tamanho excessivo e geometria muito irregular.*

1. Introdução

Estudos de detecção de conglomerados espaciais são procedimentos importantes na área de saúde pública. O diagnóstico preciso sobre a característica aleatória ou não de um determinado evento espacial como, por exemplo, uma doença contagiosa, e a delimitação da região geográfica de ocorrência possibilitam aos órgãos competentes a elaboração de políticas eficientes de controle e combate. Como resultado, procura-se identificar áreas geográficas com um risco significativamente elevado sem o conhecimento, a princípio, de quais e quantas áreas são, caracterizando um teste genérico de conglomerado.

Um conglomerado pode ser definido como um conjunto de áreas que apresentam um risco significativamente elevado quando considerada a hipótese nula (H_0) de que os eventos são gerados aleatoriamente sobre a região de estudo. Um conglomerado pode ser caracterizado como temporal, espacial ou espaço-temporal, dependendo da variável (espaço e/ou tempo) sobre a qual é realizada a análise de contagem dos eventos. Especificamente, o presente trabalho abrange a detecção de conglomerados espaciais.

Diversas abordagens são apresentadas para a delimitação de conglomerados. Métodos puramente gráficos identificam regiões críticas a partir de sobreposição de círculos, mas não fornecem uma medida de significância da região delineada [Besag 1991, Openshaw et al. 1988]. Neste contexto, o Método de varredura espacial *scan* proposto por Kulldorff (1997) tem sido amplamente utilizado em virtude do poder de detecção [Kulldorff et al. 2003, Costa and Assunção 2005] e da capacidade de atribuir um nível de significância à estatística de teste via simulação Monte Carlo, reduzindo o erro do tipo I. Entretanto, em sua formulação original, o método é condicionado à busca de conglomerados que apresentam geometria circular. Tal característica reduz substancialmente o custo computacional do método uma vez que uma busca exaustiva sobre todos os possíveis candidatos a cluster em uma área subdividida em *n* sub-áreas acarretaria uma varredura sobre 2^n candidatos. Apesar da vantagem da eficiência computacional, o método apresenta limitações quando o conglomerado real passa a apresentar uma geometria irregular, detectando nenhuma ou pequenas áreas do mesmo. O tratamento da irregularidade do conglomerado tem sido abordado a partir de heurísticas computacionais, como o método de Simulated Annealing [Duczmal and Assunção 2004] ou delimitando uma região circular de tamanho fixo, menor que a região de estudo, e realizando uma busca exaustiva nas áreas contidas em seu interior [Tango and Takahashi 2005]. Sob suposição de que as regiões que definem o conglomerado compartilham fronteira geográfica, foi proposto o método de árvore dinâmica [Assunção et al. 2006] que promove o crescimento de conglomerados agregando as áreas vizinhas que favorecem a maximização da verossimilhança do conglomerado.

Neste trabalho, o método de crescimento dinâmico de conglomerados (*dMST – dynamic Minimum Spanning Tree*) é avaliado. A generalização da geometria resulta na identificação de conglomerados de tamanho elevado e muito irregulares quando comparado com dados simulados. Com a finalidade de minimizar esses efeitos, são propostas restrições sob a verossimilhança das vizinhanças bem como sobre o tamanho mínimo do conglomerado a ser detectado.

2. O método de varredura *SCAN*

Seja uma região geográfica delimitada, subdividida em *n* sub-áreas, sendo associada a cada sub-área o número observado de casos y_i e o número total de pessoas em risco na área, N_i. Sob a hipótese nula de aleatoriedade ou ausência de conglomerados, o número esperado de casos na *i*-ésima área pode ser modelado por uma variável aleatória de Poisson e é independente das demais áreas: $H_0 : y_i \sim Poisson(E_i = \lambda N_i)$, onde a taxa estimada de ocorrência de casos é calculada como: $\hat{\lambda} = C/M$, onde $C = \sum_i y_i$ e $M = \sum_i N_i$.

Seja Z o conjunto das áreas z candidatas a formarem um conglomerado. Se não for imposta nenhuma restrição o conjunto Z possui 2^n elementos. O vetor de parâmetros do método de máxima verossimilhança para o método *scan* é definido pela área candidata z, a probabilidade de que um indivíduo em z seja um caso (p), e a probabilidade de um indivíduo fora de z seja um caso (r). Sob a hipótese nula: $p=r$ e sob

a hipótese alternativa: $p>r$. Definindo n_z como a população em z, c_z o número de casos em z, $\hat{p} = c_z / n_z$ e $\hat{r} = (C - c_z)/(M - n_z)$, o candidato a conglomerado é definido por:

$$L(z, p_z, r_z) = \sup_{z \in Z, p>r} p^{c_z} (1-p)^{(n_z - c_z)} r^{(C-c_z)} (1-r)^{(M-n_z-C+c_z)} \quad (1)$$

referente ao modelo de Bernoulli, para todo $z \in Z$. Ao conglomerado verossímil é atribuída uma estatística baseada na razão de verossimilhança: $\kappa = L(\hat{z}, \hat{p}_z, \hat{r}_z) / L_0$, onde $L_0 = C^C (M-C)^{M-C} / M^M$.

A distribuição empírica da estatística κ condicionada ao número total de casos é obtida via simulação Monte Carlo a partir dos seguintes passos:

1. Gera-se *S* conjuntos independentes de vetores de casos, cuja soma dos elementos de cada vetor seja *C*, a partir de realizações de uma distribuição multinomial proporcional a população de cada área. Calcula-se a estatística κ para cada conjunto: (κ_1, …, κ_S);
2. Ordena-se os valores de κ. Se o valor obtido com o conjunto de dados original estiver entre os maiores 100(1-α)%, rejeita-se H_0 ao nível de significância α.
3. Caso H_0 tenha sido rejeitada, a área $\hat{z}$ associada é o conglomerado mais verossímil.

Em sua proposta inicial [Kulldorff 1997], o conjunto *Z* representa círculos de raio *r* arbitrário centrados em cada um dos *n* centróides das sub-áreas. Tal restrição reduz significativamente o número de candidatos a conglomerados e conseqüentemente, o custo computacional.

3. O Algoritmo de Construção de Conglomerados com Geometria Irregular

Seja a região de interesse definida por sub-áreas que compartilham fronteira geográfica de forma que, para uma particular sub-área *i*, exista pelo menos uma outra sub-área *j* que possui fronteira comum. Pode-se expressar essa informação sob a forma de um grafo interconectando os centróides das sub-áreas aos seus vizinhos, conforme ilustra a Figura 1-A.

Uma árvore geradora mínima de um grafo representa um sub-grafo interconectando todas as arestas, mas cujo caminho entre dois nodos *i* e *j* seja único, de tal forma que se uma aresta da árvore geradora mínima for removida obtêm-se dois sub-grafos não conectados. A Figura 1-B ilustra uma árvore geradora mínima.

O algoritmo para a construção de geometrias arbitrárias tem como objetivo a construção de árvores geradoras mínimas na qual o custo de agregação de uma área à árvore está associada à verossimilhança da árvore resultante. O algoritmo de crescimento da árvore geradora mínima utilizando a equação de verossimilhança é descrito a seguir:

1. Partindo de cada sub-área, calcule a verossimilhança de acordo com a Equação 1 considerando cada um de seus vizinhos como possíveis candidatos a incorporar o conglomerado;
2. Inclua na árvore o vizinho que resulta na maior verossimilhança;

3. Defina os vizinhos da nova árvore;
4. Retorne à etapa 2 e repita o procedimento até que todas as sub-áreas estejam incluídas na árvore geradora mínima ou até que a árvore alcance um tamanho máximo pré-definido.

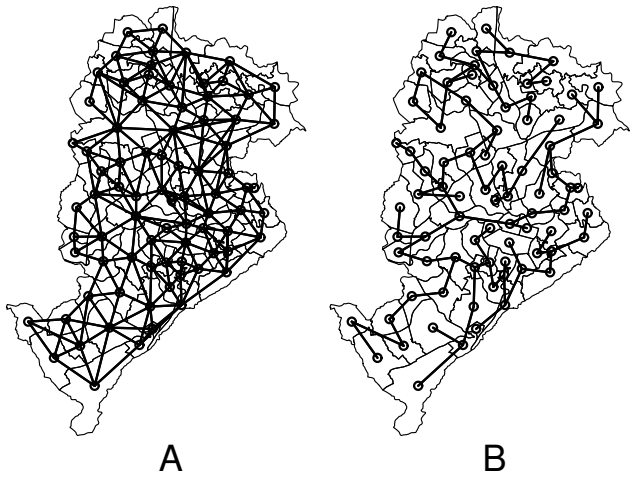


Figure 1. Mapa de Belo Horizonte subdividido em unidades administrativas interconectadas por um grafo de arestas (A) e o grafo da árvore geradora mínima (B)

Em sua proposta original, o critério de parada do algoritmo de construção de árvores é o tamanho máximo especificado pelo usuário e, para uma região com *n* sub-áreas, são geradas *n* árvores. Durante o crescimento das árvores, o método armazena a estrutura (sub-árvore) de máxima verossimilhança. Em seqüência, o método de simulação de Monte Carlo é utilizado para o cálculo do nível descritivo associado à estatística da razão de verossimilhança κ, sob H_0.

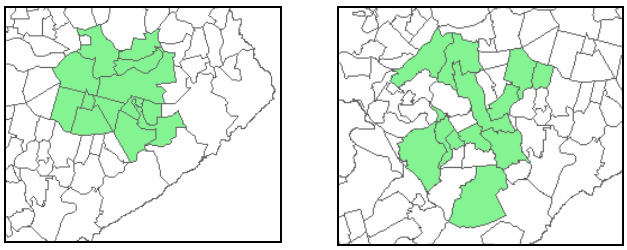
4. Metodologia

Uma análise de desempenho do método *dMST* é proposta a partir de dados simulados e dados reais. A região de interesse é representada pela região metropolitana de Belo Horizonte subdividida em bairros.

Os dados simulados foram gerados a partir de dois cenários distintos para os conglomerados, apresentados na Figura 2. A população de interesse em cada bairro foi obtida a partir do censo do ano de 2000. No primeiro cenário, especificou-se um conglomerado com geometria circular constituído por 13 bairros. No segundo cenário, especificou-se um conglomerado com geometria *estrela* constituído por 12 bairros. O número de casos especificado para a região é de 420, sendo distribuídos de acordo com uma distribuição multinomial na qual as probabilidades referentes aos bairros do conglomerado foram ajustadas a partir da especificação de um risco relativo, favorecendo a rejeição da hipótese nula com probabilidade 0.999 [Kulldorff et al.,

2003]. Uma vez definidos os parâmetros de simulação, foram geradas 10.000 simulações para cada cenário, onde para cada simulação, foram distribuídos 420 casos entre os bairros. Em seguida avaliou-se o poder de detecção do método *scan* circular, *dMST* e mais duas variações propostas:

1. $dMST_2$: Método *dMST* com parada prematura. Nesta abordagem a árvore irá crescer enquanto existir algum vizinho que, ao ser acrescentado, resulta em uma árvore com verossimilhança maior que a árvore anterior, caso contrário o crescimento é interrompido e uma nova árvore é gerada a partir das demais áreas;
2. $dMST_3$: Método *dMST* com parada prematura, busca suavizada e tamanho mínimo. Semelhante à abordagem anterior, inicialmente a árvore cresce até atingir um tamanho mínimo. Em seqüência, é agregado à árvore o vizinho que proporciona o menor crescimento da verossimilhança dentre todos os vizinhos capazes de maximizar a verossimilhança em relação à árvore anterior. Caso nenhum vizinho proporcione o aumento da verossimilhança, o método é interrompido e novas árvores são geradas a partir das demais áreas.



(a) circular (b) estrela

Figura 2. Cenários de conglomerados para dados simulados

Para avaliar o desempenho das metodologias em situações reais, aplicou-se a análise de conglomerados para os dados de homicídios do ano de 2000.

5. Resultados

A partir das 10.000 simulações avaliou-se o poder de detecção dos métodos scan, *dMST*, $dMST_2$ e $dMST_3$. Os resultados obtidos para o método *scan* são mostrados na Figura 3, na qual a escala dos eixos foram previamente padronizada: 0-60 (eixo x) e 0-5.000 (eixo y) para os gráficos de contagem de tamanho detectado e 0-3.000 (eixo y) para os gráficos de interseção sendo a escala do eixo x definido em função do cenário: 0-13 (*circular*) e 0-12 (*estrela*). Para o cenário circular, o método scan apresentou um bom desempenho, mas com um número elevado de conglomerados detectados sem interseção com o conglomerado real (2.225). Para o cenário estrela, ocorreu uma queda de desempenho e uma maior irregularidade na distribuição das interseções. Em ambos os cenários, o método detectou com maior freqüência conglomerados de tamanho unitário.

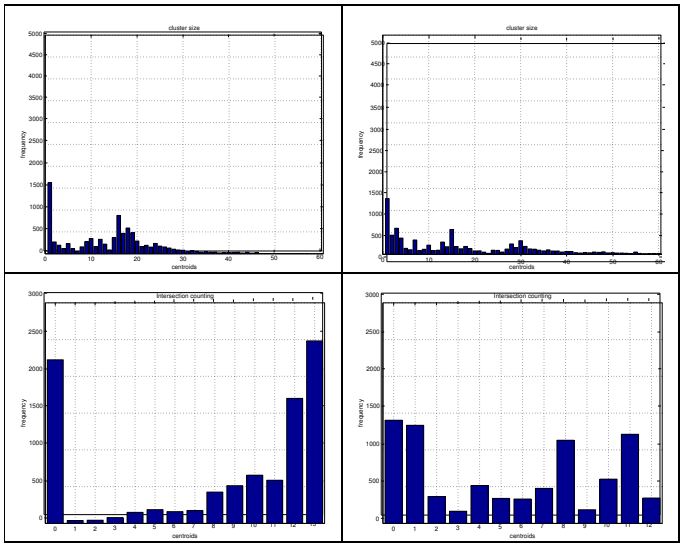


Figura 3. Distribuição do tamanho do conglomerado encontrado (linha 1) e distribuição da contagem da interseção entre o conglomerado encontrado e o conglomerados real (linha 2), pelo método *scan* para os cenários circular (coluna 1) e estrela (coluna 2)

As Figuras 4 e 5 apresentam os resultados para os métodos *dMST*, $dMST_2$ e $dMST_3$. O método *dMST* detectou conglomerados com dimensões elevadas, concentrados próximas do limite de tamanho máximo (60 bairros) em ambos os cenários. As distribuições das interseções são bem regulares com médias inferiores ao tamanho real do conglomerado mas, próximas do mesmo. Os métodos $dMST_2$ e $dMST_3$ possibilitaram uma redução da média da distribuição do tamanho detectado. Em especial, o método $dMST_2$ gerou uma contagem maior de elementos sem interseção com o conglomerado real, característica que foi minimizada pelo método $dMST_3$. Por outro lado, o método $dMST_3$ detectou uma quantidade maior de conglomerados com interseções de 1 e 2 bairros em relação ao conglomerado real. A Tabela 1 apresenta a contagem entre as 10.000 simulações realizadas, que foram efetivamente consideradas, uma vez que eram computadas apenas as simulações nas quais a hipótese nula era rejeitada. Destas, também são apresentadas as contagens de interseção nula, ou seja, os casos nos quais o conglomerado detectado não apresenta nenhuma interseção com o conglomerado real. A partir desta Tabela pode-se compara os métodos em relação à detecção efetiva, por exemplo, apesar do método *scan* não rejeitar 9.423 simulações (circular) ocorreram 2.225 casos sem interseção, por outro lado o método $dMST_3$ não rejeitou 8.002 simulações, mas ocorreu um número menor de simulações sem interseção, 517.

A Figura 6 apresenta os resultados para os dados de Homicídio em Belo Horizonte durante o ano de 2000. O conglomerado detectado pelo método *scan* abrange 16 bairros sendo que, em alguns bairros não são observados casos. O método *dMST* obteve o maior conglomerado (48 bairros) e, visualmente, o conglomerado resultante é uma interligação de vários sub-conglomerados. O método $dMST_2$ identificou um conglomerado de tamanho 1 e o método $dMST_3$ obteve um conglomerado de tamanho 8 constituído pelo conglomerado detectado pelo método $dMST_2$, parte do conglomerado detectado pelo método *dMST* e um bairro sem contagem interligando os mesmos.

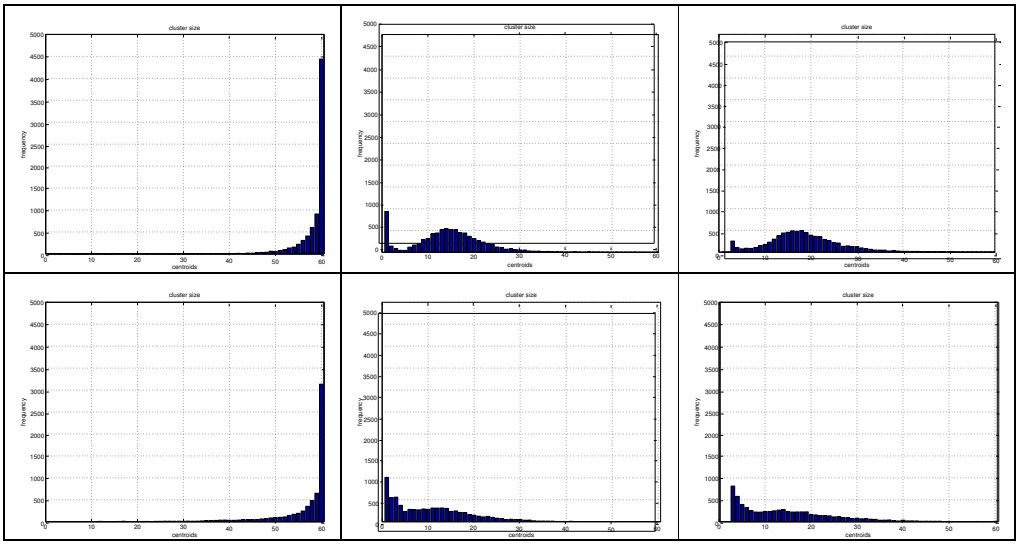


Figura 4. Distribuição do tamanho do conglomerado encontrado pelo método *dMST* (coluna 1), $dMST_2$ (coluna 2) e $dMST_3$ (coluna 3) para os cenários circular (linha 1) e estrela (linha 2)

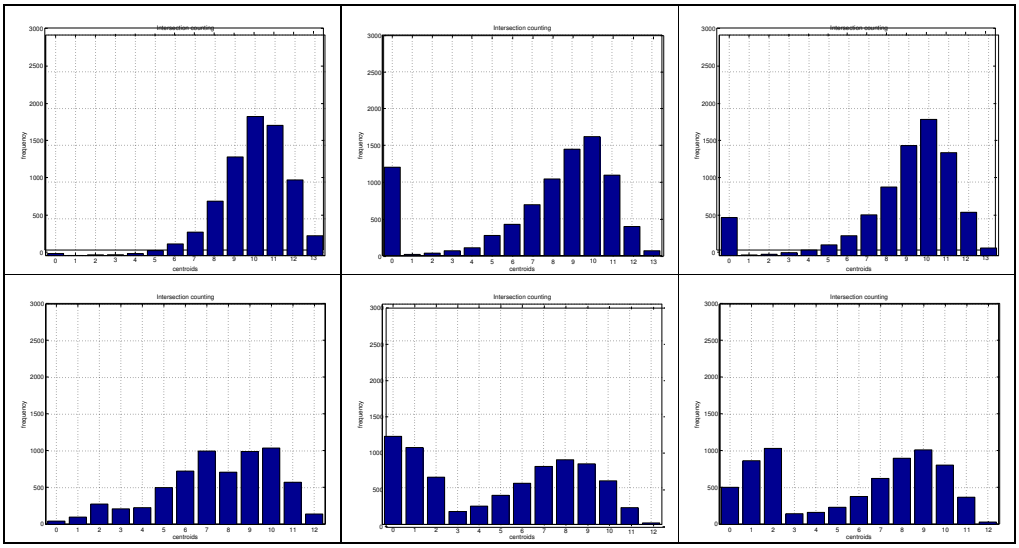
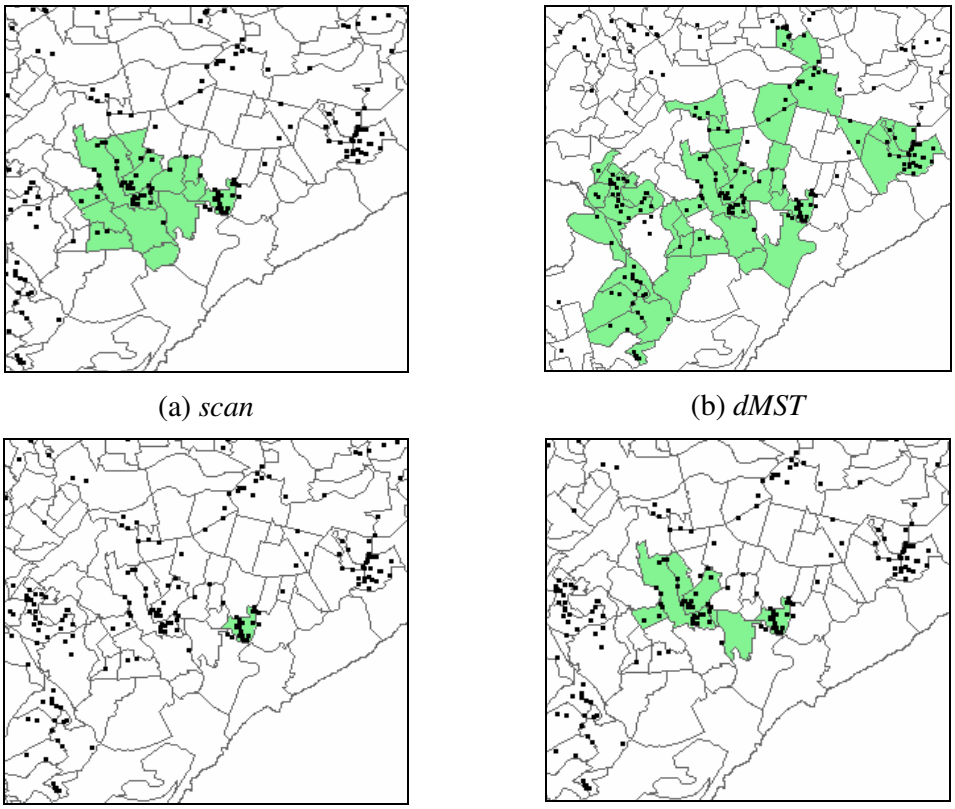


Figura 5. Distribuição da contagem da interseção entre o conglomerado encontrado e o conglomerado real obtido pelo método *dMST* (coluna 1), $dMST_2$ (coluna 2) e $dMST_3$ (coluna 3) para os cenários circular (linha 1) e estrela (linha 2).

Tabela 1. Contagem do número de simulações, em 10.000, nas quais a hipótese nula foi rejeitada e , entre essas, comparação com a contagem de simulações sem interseção entre o conglomerado encontrado e o conglomerado real para os cenários: circular e estrela.

Método	Característica	Circular	Estrela
scan	*p*-valor < 0.5	9423	8405
	sem interseção	2225	1403
dMST	*p*-valor < 0.5	7669	6549
	sem interseção	29	45
$dMST_2$	*p*-valor < 0.5	8476	7605
	sem interseção	1202	1199
$dMST_3$	*p*-valor < 0.5	8002	7082
	sem interseção	517	505



(a) *scan* (b) *dMST*

(c) $dMST_2$ (d) $dMST_3$

Figura 6. Conglomerados identificados para os dados de Homicídios em Belo Horizonte durante o ano de 2000.

8. Discussões e Conclusões

A partir dos dados simulados foram levantados alguns aspectos em relação às características dos conglomerados detectados por cada método. Para os cenários simulados, o método *scan* detecta conglomerados com elevada intensidade de bairros sem casos, provavelmente devido à própria restrição da geometria. Este fato também é observado no método *dMST* com geometrias arbitrárias. Neste caso, há evidência de que esta característica está relacionada ao crescimento exagerado da árvore uma vez que o método $dMST_3$ minimiza este efeito e o método $dMST_2$ praticamente o elimina.

Do ponto de vista de freqüência de detecção do conglomerado real nas simulações, seja na totalidade ou na parcialidade, ambos os métodos apresentam um desempenho próximo. O método scan apresenta uma baixa taxa de rejeição da hipótese nula mas uma alta taxa de conglomerados sem interseção. Os métodos *dMST*, $dMST_2$ e $dMST_3$ apresentam alta taxa de rejeição da hipótese nula mas, baixa taxa de não-interseção.

Na aplicação aos dados reais, a análise da interseção de todos os métodos indica uma concentração anormal de casos em um único bairro. Por outro lado, um segundo conglomerado pode ser identificado a partir das interseções dos bairros detectados pelos métodos *scan*, *dMST* e $dMST_3$.

A partir dos resultados obtidos pode-se concluir que, para as bases de dados estudadas, incorporar a estrutura de vizinhança tornou o método *scan* mais focado permitindo a redução do número de interseções nulas e do número de regiões sem casos no conglomerado final. A metodologia também permite obter, além de um valor para a estatística de teste e um p-valor associado, a geometria do conglomerado.

Agradecimentos

Os autores agradecem à FAPEMIG e à Pró-Reitoria de Pesquisa da UFMG pelo apoio financeiro.

Bibliografia

Assunção, R., Costa, M., Tavares, A., Ferreira, S. (2006), Fast detection of arbitrary shaped disease clusters. *Statistics in Medicine*. Forthcoming

Besag, J. and Newell, J. (1991), The detection of clusters in rare diseases. *Journal of the Royal Statistic Society A*, vol. 154, pages 143-155.

Costa, M. A., Assunção, R. M. (2005), A fair comparison between the spatial scan and the Besag-Newell disease clustering tests. *Environmental and Ecological Statistics*, vol. 12, pages 297-315.

Duczmal, L. and Assunção, R. (2004), A simulated annealing strategy for the detection or arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, vol. 45, pages 269-286.

Kulldorff, M., Tango, T., Park, P. J., (2003), Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, vol. 42, pages 665-684.

Kulldorff, M. (1997), A Spatial Scan Statistics. *Commun. Statist. – Theory and Methods*., vol. 26, pages 1481-1496.

Openshaw, S., Craft, A. W., Charlton, M. and Birch, J. M. (1988), Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet* **i**, pages 272-273

Tango, T., Takahashi, K. (2005), A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics,* **4**:11.