UTILIZAÇÃO DO PROCEDIMENTO INFERÊNCIA DATA-DRIVEN PARA A ESTATÍSTICA ESPACIAL SCAN EM CASOS DO DIABETES NO ESTADO DE MINAS GERAIS

Gilberto de Andrade[1], **Anderson Ribeiro Duarte**[1]

Resumo: *O método de detecção e inferência de conglomerados* (clusters) *Scan Circular para mapas da dados agregados, procura por clusters de casos sem especificar o tamanho (número de áreas) ou localização geográfica antecipadamente. Existe ainda, uma proposta de modificação para o teste inferencial usual da estatística Scan, denominada inferência Data-Driven, incorporando informações adicionais sobre o tamanho do cluster mais provável encontrado. Será apresentada a estrutura das duas técnicas inferenciais, e ainda, será proposta uma avaliação através do procedimento clássico e também do novo procedimento Data-Driven avaliando um conjunto de dados reais para ocorrência de casos do Diabetes no estado de Minas Gerais. As conclusões mostram que realmente o novo procedimento pode propiciar novas conclusões acerca da significância de eventuais conglomerados existentes nos conjuntos de dados em estudo.*

Palavras-chave: Scan Circular; clusters; inferência Data Driven; significância estatística; Diabetes.

Introdução

Observa-se, recentemente, um crescente número de trabalhos sobre metodologias para detecção e avaliação de clusters espaciais e temporais. No enfoque deste texto, um cluster é um conjunto conexo de regiões onde existe a ocorrência discrepante de casos localizados para algum fenômeno de interesse. O processo de detecção pode ser realizado em intervalos de tempo *(cluster temporal)* ou então, para localizações no espaço *(cluster espacial)*, ou em ambos *(cluster espaço-temporal)*.

O problema de detecção de clusters espaciais encontra-se presente em diversas situações, tais como problemas associados à saúde pública (epidemiologia e vigilância sindrômica), criminologia, pesquisa de mercados, entre outros. É importante determinar modelos satisfatórios para a execução de procedimentos para detecção e avaliação destes clusters. Este trabalho se restringe à detecção de clusters espaciais, entretanto as propostas aqui discutidas podem ser estendidas para a busca de clusters temporais e espaço-temporais.

Uma metodologia baseada em um teste de razão de verossimilhança [6] propõe a estatística de teste Scan Espacial. Um caso particular de vasta utilização da aplicação de tal metodologia denominado Scan Circular [7] constroi um teste que encontra o cluster mais verossímil dentre todas as zonas circunscritas por círculos de raios variados centrados em cada região do mapa.

O Scan Circular é utilizado para detectar e avaliar clusters com uma formação temporal, espacial e espaço-temporal. Isto é feito através de uma janela que gradualmente varre uma

[1]Departamento de Matemática, ICEB, UFOP,
gilberto_est09@yahoo.com.br, anderson@iceb.ufop.br

região para um determinado intervalo de tempo e/ou até alcançar um raio máximo de varredura pré-determinado.

Para aplicação do Scan Circular, considere um mapa dividido em m regiões, com uma população total P, e um número total de casos C para algum fenômeno de interesse a ser estudado. Defina um ponto arbitrário no interior de cada região, tal ponto será denominado centróide. Seja z, um conjunto conexo qualquer de regiões no mapa em estudo, definiremos este conjunto por zona. Assim, cada possível zona no mapa em estudo, tem uma população e um total de casos, $P(z)$ e $C(z)$ respectivamente. Estas zonas (candidatos a cluster) são definidas por círculos de raio arbitrário r centrados em cada um dos m centróides das regiões do mapa. Variando o valor r entre 0 e algum valor R pré-estabelecido, podemos definir zonas determinadas por tais janelas circulares. A zona cujos centroídes são interiores à janela circular em avaliação é dita zona definida por tal janela circular.

Cada zona será avaliada através do logarítimo da função de verossimilhança para a distribuição dos casos do fenômeno de interesse. Um modelo comumente utilizado assume que o número de casos em cada área segue distribuição Poisson com taxa proporcional à sua população.

Agrupamentos são então identificados para diferentes raios de varredura. Contudo, apenas alguns agrupamentos podem ser considerados de importância. Para identificar estes, para cada agrupamento, é testada a hipótese de o mesmo ter ocorrido ao acaso. O teste utilizado para esta finalidade é o da razão da verossimilhança. O agrupamento mais relevante é aquele que apresenta maior razão de verossimilhança.

Dado o conjunto de todas as zonas em avaliação para os diferentes raios de varredura, ora denominado conjunto Z. Busca-se determinar as zonas que podem ser considerados de maior relevância quanto ao valor do logaritmo da função de verossimilhança. É importante salientar que as zonas mais verossímeis, não são necessariamente clusters. Uma zona será dita cluster quando o valor do logaritmo da função de verossimilhança for considerado significativo do ponto de vista estatístico. Para tal avaliação, executa-se um teste de hipóteses com a Hipótese Nula de que não existe cluster no mapa em estudo, contra a Hipótese Alternativa de que existe pelo menos um cluster no mapa em estudo.

A estatística de teste Scan será definida então como a razão de verossimilhanças. Sob a validade da Hipótese Nula e assumindo o modelo Poisson, o número de casos esperados em uma possível zona z é dado por $\mu(z) = C\frac{P(z)}{P}$. Desta forma, temos o risco relativo na zona z dado por $I(z) = \frac{C(z)}{\mu(z)}$. Já o risco relativo fora da zona z é dado por $O(z) = \frac{C-C(z)}{C-\mu(z)}$. Para L_0 sendo a função de verossímilhança sob a Hipótese Nula e $L(z)$ sendo a função de verossímilhança sob a Hipótese Alternativa. O logaritmo da razão de verossimilança assumindo o modelo Poisson é dada por:

$$LLR(z) = \begin{cases} C(z)\log\left(I(z)\right) + \left(C - C(z)\right)\log\left(O(z)\right) & \text{se } I(z) > 1 \\ 0 & \text{caso contrário} \end{cases}$$

O logaritmo da razão de verossimilhança é então maximizado no conjunto Z. O formato de definição aqui exposto para o conjunto Z é o que define o método Scan Circular. Existem outros critérios para a definição do conjunto Z, como por exemplo, janelas elípticas, ou até mesmo uma busca exaustiva sobre todas as possíveis zonas conexas no mapa em estudo. No caso de considerarmos Z como o conjunto de todas as zonas conexas, o problema se tornaria impraticável para mapas com m da ordem de algumas centenas.

Para concluir o teste de hipóteses, a significância estatística de uma possível solução, obtida através da distribuição dos casos observados, em geral, é verificada através de simulações de Monte Carlo, dado o desconhecimento da distribuição exata da estatística de teste. No procedimento de Monte Carlo, casos simulados são distribuídos aleatoriamente no mapa em estudo, de forma que cada região recebe, em média, um número de casos proporcional à sua população. A significância estatística, de uma solução obtida através da técnica Scan Circular, é considerada sem pré-especificação do número de regiões e/ou da localização do clusters mais verossímel. O processo inferêncial usual compara a solução mais verossímil obtida dos dados observados com

as soluções mais verossímeis obtidas de cada distribuição de casos simulada. Esta comparação é feita através da distribuição empírica para a estatística de teste construída através dos dados da simulação de Monte Carlo.

Tendo como base a metodologia da Scan Circular, o procedimento inferencial original foi discutido em [2] e uma nova técnica inferencial, mais robusta para o problema foi apresentada. A nova técnica denominada inferência Data-driven, leva em consideração o conhecimento da quantidade de regiões que compõem a solução mais verossímel obtida dos dados observados. Neste caso a construção da distribuição empírica é condicionada ao conhecimento da quantidade de regiões na solução mais verossímel. Este procedimento será melhor discutido na seção que trata dos aspectos metodológicos deste trabalho.

Objetivo

O procedimento de inferência Data-Driven foi apresentado e discutido através de resultados baseados em dados simulados. Foi observado que efetivamente existe uma diferença significativa na análise de resultados quanto a efetiva significância dos clusters detectados. Por outro lado, ainda não existem estudos envolvendo esta técnica na avaliação de dados reais.

O objetivo central deste trabalho é verificar as modificações de tomada de decisão que podem ocorrer dada a mudança do procedimento inferencial quando analisando um conjunto de dados reais. Foram obtidos dados referentes a casos do Diabetes no estado de Minas Gerais para a realização deste estudo.

O Diabetes é uma doença crônica que ocorre quando o pâncreas não produz insulina suficiente, ou quando o corpo não pode utilizar eficazmente a insulina que produz. Hiperglicemia, ou açúcar no sangue elevado, é um efeito comum do Diabetes descontrolado e ao longo do tempo produz sérios danos a muitos dos sistemas do corpo humano, especialmente os nervos e vasos sanguíneos. No banco de dados de casos reais para o Diabetes no Estado de Minas Gerais, consideramos como população de risco, homens e mulheres com mais de 45 anos de idade. Totalizando uma população de risco de 7033712, entre o período de janeiro de 2002 até maio de 2011. Para esse estudo, no banco de dados utilizado consideramos conjuntamente Diabetes do Tipo 1 e Tipo 2, totalizando neste período, a ocorrência de 28039 casos. A população de risco e também os casos ocorridos estão distribuídos ao longo dos municípios nos quais ocorreu a identificação da doença. O mapa em estudo portanto será dividido por municípios, totalizando 853 regiões de estudo.

A população total usada nesse estudo é referente ao Censo Demográfico de 2010, segundo informações do Instituto Brasileiro de Geografia e Estatística, IBGE, [5]. Os conjuntos de dados referentes às populações de risco foram obtidos no site do Ministério da Saúde [4] e também podem ser encontrados em [3].

Metodologia

A utilização do procedimento inferencial clássico para o Scan Circular compara a zona em análise obtida dos dados observados com a distribuição empírica obtida através de simulações de Monte Carlo. Uma preocupação imediata seria questionar se a zona em análise obtida dos dados observados está sendo comparada com uma distribuição empírica produzida sob a validade da hipótese nula, mas através de zonas que se assemelham o máximo possível da obtida através dos dados observados.

O estudo em [2] vem exatamente contradizer este fato, os estudos desse trabalho mostram que a grande maioria das zonas produzidas através do procedimento de simulação não se assemelham muito à zona obtido dos dados observados, especiamente quando está é composta por um número elevado de regiões.

Em sua formulação original, o procedimento inferencial clássico calcula a importância do

cluster mais verossímil baseado na seguinte pergunta: "Dado que o cluster candidato encontrado tem estatística de teste igual ao valor x, qual é a probabilidade de encontrar uma zona mais verossímil sob a hipótese nula com a estatística de teste maior que o valor x?"

O procedimanto inferencial Data-Driven propoe utilizar as informações sobre a quantidade de regiões da zona em análise obtida dos dados observados. Neste caso, a seguinte pergunta é formulada: "Dado que o cluster candidato encontrado tem estatística de teste igual ao valor x e contém k regiões na sua composição, qual é a probabilidade de encontrar uma zona mais verossímil sob a hipótese nula com a estatística de teste maior que o valor x composta por exatamente k regiões?"

Dado que o Scan Circular encontrou uma zona nos dados observados com k regiões, então a sua significância estatística ainda será obtida através de simulações de Monte Carlo, mas selecionando apenas os réplicas em que as soluções têm exatamente k regiões. A distribuição empírica que considera soluções de qualquer tamanho é substituída então pela distribuição empírica Scan_k que é obtida considerando apenas as soluções de tamanho exatamente igual a k.

Através de extensos testes numéricos [1] foi mostrado que, sob a validade da hipótese nula, a distribuição empírica para a Estatística Scan é aproximada pela bem conhecida distribuição de Gumbel

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} e^{-e^{-\frac{x-\mu}{\beta}}}$$

com parâmetros μ (locação) e β (escala). Usando uma aproximação semi-paramêtrica, a distribuição para a estatística de teste pode ser estimada usando um número bem menor de replicações de Monte Carlo. Por exemplo, os valores críticos obtidos através da distribuição Gumbel ajustada, através de 100 réplicas de Monte Carlo, são tão precisos quanto valores críticos obtidos através da distribuição empírica obtida através de 10000 réplicas de Monte Carlo.

Da mesma forma que é possível executar o ajuste para a distribuição Gumbel no procedimento inferencial clássico para a avaliação da significância dos clusters, o ajuste é também bastante satisfatório para as distribuições empíricas Scan_k definindo então a distribuição ajustada Gumbel_k [2].

Avaliações Numéricas

Considerando o banco de dados para os casos do Diabetes (já mencionado anteriormente), o procedimento Scan Circular foi executado para a detecção da possível existência de clusters no mapa em estudo. Através do procedimento inferencial clássico foi verificada a existência de soluções significativas.

Neste momento, uma nova definição se faz importante, não buscamos apenas avaliar a zona mais verossímil encontrada no mapa de dados observados. Buscamos avaliar todas as soluções significativas, do ponto de vista estatístico, para os dados em análise. Para tanto, definiremos os conceitos de cluster primário, secundário, terciário e assim sucessivamente.

Definiremos como cluster primário, a solução mais significativa obtida no mapa em estudo. Já o cluster secundário, é a solução mais significativa obtida no mapa em estudo, que não intercepta o cluster primário. O cluster terciário é a solução mais significativa obtida no mapa em estudo, que não intercepta os clusters primário e o secundário e assim sucessivamente para as demais soluções.

De posse do conjunto de dados, 10000 simulações de Monte Carlo sob a validade da hipótese nula foram executadas, com o intuíto de produzir a distribuição empírica usual do procedimento inferencial clássico. Visando apenas construir uma estratégia de verificação, de forma arbitrária foi escolhido o nível de significância $\alpha = 0.01$ para a determinação de quais soluções obtidas eram efetivamente significativas do ponto de vista estatístico.

Através desta análise, verificou-se a presença de 36 soluções significativas no mapa, todas eles disjuntas, ou seja, considerando o conceito de clusters primários, secundários e assim por diante.

Dadas estas soluções significativas, a tomada de decisão do ponto de vista de saúde pública seria por estabelecer algum tipo de medida associada às regiões que pertencentes à união destas 36 soluções significativas.

A informação de maior interesse neste momento seria considerar se a troca do procedimento inferencial levaria a uma conclusão diferente, que poderia por exemplo, modificar a alocação de recursos na tomada de decisão acerca das medidas preventivas sobre a disseminação da doença em estudo.

O procedimento inferencial Data-Driven foi então considerado para cada uma quantidade distinta de regiões encontrando 39 soluções significativas. Para as soluções muito significativas (p-valor muito baixo), como já era previsto, as conclusões foram iguais para os dois procedimentos inferenciais, entretanto, para soluções cujo p-valor se aproximava do limiar de 1% uma análise mais cuidados deve ser observada. Os resultados são apresentados na Tabela 1.

Tabela 1: Soluções avaliadas para os casos do Diabetes através da distribuição empírica

Solução observada	Quant. de regiões k	Estatística de teste	Valor crítico Scan	Conclusão do teste	Valor crítico Scan_k	Conclusão do teste
1	171	2523.94019	11.34048	significativo	–	inconclusivo
2	65	316.16461	11.34048	significativo	–	inconclusivo
3	1	160.63046	11.34048	significativo	10.27693	significativo
⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	1	43.76292	11.34048	significativo	10.27693	significativo
12	5	43.35877	11.34048	significativo	11.21412	significativo
13	1	41.69544	11.34048	significativo	10.27693	significativo
⋮	⋮	⋮	⋮	⋮	⋮	⋮
36	1	12.43443	11.34048	significativo	10.27693	significativo
37	1	11.29505	11.34048	**não significativo**	10.27693	**significativo**
38	1	11.24991	11.34048	**não significativo**	10.27693	**significativo**
39	1	11.08631	11.34048	**não significativo**	10.27693	**significativo**
40	1	10.20864	11.34048	não significativo	10.27693	não significativo

As duas primeiras soluções tiveram a decisão do teste classificada como inconclusiva, quando utilizando o procedimento Data-Driven. Isto se deve ao pequeno número de componentes amostrais obtidos com estas quantidades de regiões no procedimento de Monte Carlo. Entretanto, parece óbvio considerar que tais soluções seriam ditas significativas ao considerarmos um maior número de simulações. Esta conclusão se deve ao número de soluções significativas que são obtidas mesmo com estatística de teste de valor bastante inferior às duas primeiras soluções, mesmo considerando os dois procedimentos inferenciais. As soluções 4 até 10 e 14 a 35 são significativas nos dois procedimentos e todas compostas por apenas uma região, em virtude disto, não foram expostas na Tabela 1. O procedimento inferencial clássico e o procedimento Data-Driven foi também considerado para todas as soluções através do ajuste da distribuição Gumbel. As conclusões obtidas foram as mesmas e podem ser observadas na Tabela 2.

Também para a distribuição Gumbel ajustada, as duas primeiras soluções tiveram a decisão do teste classificada como inconclusiva, quando utlizando o procedimento Data-Driven. Novamente, isto se deve ao pequeno número de componentes amostrais obtidos com estas quantidades de regiões no procedimento de Monte Carlo para a estimação de parâmteros para a distribuição ajustada. Entretanto, também parece óbvio considerar que tais soluções seriam ditas significativas ao considerarmos um maior número de simulações. Esta conclusão se deve ao número de soluções significativas que são obtidas mesmo com estatística de teste de valor bastante inferior às duas primeiras soluções, mesmo considerando os dois procedimentos inferenciais. Considerando a análise através da distribuição Gumbel, novamente as soluções 4 até 10 e 14 a 35 são significativas nos dois procedimentos e todas compostas por apenas uma região, em virtude disto, não foram expostas na Tabela 2.

Tabela 2: Soluções avaliadas para os casos do Diabetes através da distribuição Gumbel

Solução observada	Quant. de regiões k	Estatística de teste	Valor crítico Gumbel	Conclusão do teste	Valor crítico Gumbel$_k$	Conclusão do teste
1	171	2523.94019	11.35769	significativo	–	inconclusivo
2	65	316.16461	11.35769	significativo	–	inconclusivo
3	1	160.63046	11.35769	significativo	10.21647	significativo
⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	1	43.76292	11.35769	significativo	10.21647	significativo
12	5	43.35877	11.35769	significativo	11.16788	significativo
13	1	41.69544	11.35769	significativo	10.21647	significativo
⋮	⋮	⋮	⋮	⋮	⋮	⋮
36	1	12.43443	11.35769	significativo	10.21647	significativo
37	1	11.29505	11.35769	**não significativo**	10.21647	**significativo**
38	1	11.24991	11.35769	**não significativo**	10.21647	**significativo**
39	1	11.08631	11.35769	**não significativo**	10.21647	**significativo**
40	1	10.20864	11.35769	não significativo	10.21647	não significativo

Verificou-se que três soluções compostas cada uma por um município (Urucânia, Araporã e Patos de Minas) foram consideradas não significativas no procedimento inferencial clássico, mas significativas quando utilizado o procedimento inferencial Data-Driven. Esta conclusão é a mesma tanto usando a distribuição empírica, quanto utilizando a distribuição ajustada Gumbel. Em algum momento, uma análise descuidada pode considerar que se trata de um diferença pequena quanto às conclusões. Entretanto, vale ressaltar que a retirada de três municipios da região dita significativa para a ocorrência de casos do Diabetes, pode levar a uma distribuição de recursos na tomada de decisão que se torne mais eficiente quanto a prevenção para a ocorrência de futuros casos do Diabetes, inclusive considerando, em particular, o caso de um município de população significativa como Patos de Minas.

A figura 1 ilustra a diferença entre o conjunto de soluções significativas utilizando o procedimento clássico e o procedimento Data-Driven.

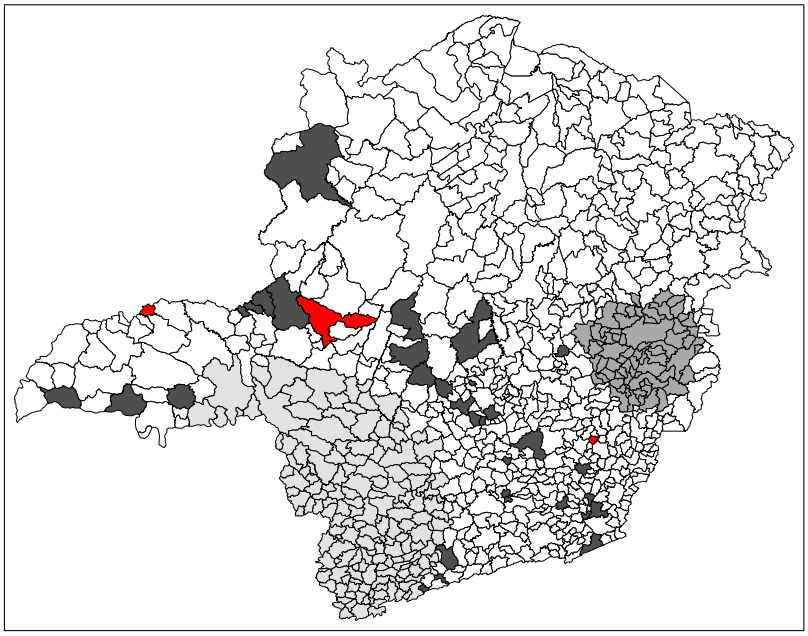

Figura 1: Mapa de Minas Gerais dividido em municípios; em tons de cinza do mais claro ao mais escuro, solução primária, secundária, demais soluções significativas nos dois métodos,respectivamente; em vermelho as soluções significativas apenas no Data-Driven.

As soluções de conclusão conflitante revelaram estatística de teste com valores 11.29505, 11.24991 e 11.08631 respectivamente. O valor crítico para $\alpha = 0.01$ utilizando o procedimento clássico através da distribuição empírica foi de 11.34048, enquanto o valor crítico para zonas compostas por uma região através do procedimento Data-Driven com a distribuição empírica foi de 10.27693, que mostra a modificação na conclusão quanto a significância das soluções observadas. Analogamente para as mesmas soluções considerando nível de signficância $\alpha = 0.01$ e utilizando o procedimento clássico através da distribuição ajustada Gumbel, o valor crítico foi de 11.35769, enquanto o valor crítico para zonas compostas por três regiões através do procedimento Data-Driven com a distribuição ajustada Gumbel foi de 10.21647, que mostra novamente, a modificação na conclusão quanto a significância das soluções observadas.

Conclusões

O processo de Inferência para procedimentos de detecção e avaliação de clusters através da estatística Scan foi avaliado na sua forma clássica e também através da estratégia inferencial data-Driven. O procedimento clássico considera todos os clusters mais verossímeis encontrado em em diversas simulações de Monte Carlo sob a validade da hipótese nula de não existência de clusters no mapa em estudo, a fim de construir a distribuição empírica para a estatística de teste, independentemente do tamanho do cluster e sua localização.

A abordagem usual apresenta uma desvantagem, é assumida implicitamente a independência entre a estatística de teste e as diversas possíveis quantidades de regiões que podem compor o cluster mais verossímil. Experimentos numéricos atestam que em alguns caso esta hipótese pode não ser verdadeira. Considerando que o cluster observado mais verossímil seja composto por k regiões, o processo de inferência clássica avalia a sua significância estatística com base no comportamento de uma maioria de clusters mais verossímeis, ao longo das simulações de Monte Carlo, cujos tamanhos (quantidade de regiões) são diferentes do valor k. Este trabalho utiliza uma outra forma de avaliação, denominada Inferência Data-Driven, que leva em conta apenas os grupos de soluções mais verossíeis encontradas ao longo das simulações de Monte Carlo, cuja quantidade de regiões seja exatamente a mesma do cluster mais verossímil obtido dos dados observados para o mapa em estudo. A inferência Data-Driven pode ainda ser aplicada a dados pontuais do tipo Caso/Controle considerando o número de casos e a população (pontos) dentro do clusters.

Nesta análise de caso, considerando os dados do Diabetes no estado de Minas Gerais, podemos concluir que em três soluções compostas cada uma por um município (Urucânia, Araporã e Patos de Minas) foram consideradas não significativas no procedimento inferencial clássico, mas significativas quando utilizado o procedimento inferencial Data-Driven, também tendo a mesma conclusão para a distribuíção empírica e para a distribuição Gumbel ajustada. Novamente ressaltamos que a retirada de três municipios da região dita significativa para a ocorrência de casos do Diabetes, pode levar a uma distribuição de recursos na tomada de decisão que se torne mais eficiente quanto a prevenção para a ocorrência de futuros casos do Diabetes, inclusive considerando, em particular, o caso de um município de população significativa como Patos de Minas.

Referências

[1] Abrams, A. M., Kleinman, K. e Kulldorff, M., Gumbel based p-value approximations for spatial scan statistics, *International Journal of Health Geographics* **9** (2010) 61 (online version).

[2] Almeida, A. C. L., Duarte, A. R., Duczmal, L. H., Oliveira, F. L. P., Takahashi, R. H. C., Data-driven inference for the spatial scan statistic, *International Journal of Health Geographics* **10** (2010) 47 (online version).

[3] Almeida, C. P., Aplicação da Função Intensidade no Delineamento de Clusters de Doença e uma Proposta da Função Intensidade Ponderada, *Dissertação de Mestrado, UFMG-Brasil, Departamento de Estatística* (2010).

[4] Sistema Único de Saúde-SUS. Disponível em: <http://www.datasus.gov.br>. Acesso em: 01 nov. 2011.

[5] IBGE - Instituto Brasileiro de Geografia e Estatística. Disponível em: <http://www.ibge.gov.br>. Acesso em: 01 nov. 2011.

[6] Kulldorff, M., A Spatial Scan Statistic, *Communications in Statistics: Theory and Methods* **26(6)** (1997) 1481-1496.

[7] Kulldorff, M. e Nagarwalla, N., Spatial disease clusters: detection and inference, *Statistics in Medicine* **14** (1995) 799-810.

Agradecimentos

Os autores agradecem o apoio recebido através do projeto de iniciação científica PROBIC apoiado pela FAPEMIG.