EXTENSÃO DA INFERÊNCIA DATA-DRIVEN AO SCAN ELÍPTICO PARA AVALIAÇÃO DE CLUSTERS IRREGULARES

Gabriel Juliano Camêlo[1,2], Gilberto de Andrade[1,2]
Henrique José de Paula Alves[1,2], Telma de Souza Lobo[1,2]
Anderson Ribeiro Duarte[1,2], Spencer Barbosa da Silva[1,2]

Resumo: *A estatística Scan Espacial é comumente usada para detecção de clusters geográficos, vigilância sindrômica e monitoramento de doenças. A forma de utilização mais difundida é o Scan Circular, entretanto se mostra ineficaz para problemas envolvendo clusters de formato irregular. Uma das soluções propostas neste cenário é o Scan Elíptico. Nos dois formatos (circular ou elíptico), o procedimento inferencial para determinar a significância estatística de um possível cluster se baseia em simulações de Monte Carlo. Uma vasta discussão existe sobre a validade do procedimento inferencial usual para o Scan Circular, tal discussão levou a proposição da metodologia de inferência denominada Data-driven. Não existem estudos sobre a aplicabilidade dessa metodologia para o Scan Elíptico. Neste trabalho explora-se a versão elíptica do Scan associada ao procedimento de inferência Data-driven com o intuito de verificar se existem diferenças evidentes entre as duas técnicas inferenciais. As avaliações são realizadas através de um benchmark de dados reais de casos de câncer no nordeste dos Estados Unidos.*

Palavras-chave: *Scan Elíptico, Data-driven, cluster irregulares, Simulação de Monte Carlo.*

1 Introdução

Os estudos de procedimentos para detecção de conglomerados espaciais são de grande interesse para epidemiologistas e profissionais afins. Tais estudos são úteis para detectar e monitorar riscos potenciais para a saúde pública. Define-se por conglomerado (*cluster*) uma área do mapa em estudo cuja incidência do fenômeno de interesse seja alta ou baixa demais a ponto de ser considerada estatisticamente significativa. Considerando um mapa em estudo subdividido em m regiões em que a população p_i e o número de casos do fenômeno de interesse c_i de cada uma das regiões são conhecidos, tem-se o interesse em avaliar se existe ou não um cluster.

A maioria dos procedimentos para esse fim se baseia na estatística Scan Espacial [2]. Esta estatística de teste foi construída com base em um teste de razão de verossimilhanças. O referido teste considera em sua hipótese nula a não existência de clusters no mapa em estudo, ao passo que a hipótese alternativa considera a existência de pelo menos um cluster estatisticamente significativo no mapa em estudo. Considerando o número de casos para cada uma das regiões como uma variável aleatória Poisson com taxa proporcional ao seu tamanho populacional, pode-se obter uma expressão fechada para a estatística de teste da razão de verossimilhanças. Para obter esse formato fechado, defina inicialmente o termo zona (z) representando qualquer subconjunto de regiões no mapa em estudo e seja ainda Z o conjunto de todas as possíveis zonas. Defina

[1]DEMAT-Universidade Federal de Ouro Preto, gabrieljulianocamelo@yahoo.com.br, gilberto_est09@yahoo.com.br, jpahenrique@gmail.com, tsouzalb@yahoo.com.br, anderson@iceb.ufop.br, spencerbars@gmail.com

[2]Agradecimento à FAPEMIG pelo apoio financeiro.

também C e P como os números totais de casos observados e população no mapa em estudo, seja ainda p_z, c_z e μ_z a população, o número de casos observados e esperados respectivamente para qualquer zona z em avaliação. Considerando que o número de casos observados em uma zona seja Poisson com taxa proporcional à seu tamanho populacional, μ_z é dado por $C(\frac{p_z}{P})$.

Considerando a notação definida anteriormente a razão de verossimilhanças fica expressa por $LR(z) = \left(\frac{c_z}{\mu_z}\right)^{c_z} \left(\frac{C-c_z}{C-\mu_z}\right)^{C-c_z}$ quando $c_z > \mu_z$ e $LR(z) = 1$ quando $c_z \leq \mu_z$ [2]. A estatística de teste será então $\max_{z \in Z} LR(z)$. Tal estatística de teste apresenta duas questões centrais em sua utilização, a primeira delas diz respeito ao fato da cardinalidade do conjunto Z ser bastante elevada para mapas com o número de regiões da ordem de algumas centenas dificultando sobremaneira a avaliação de todos os possíveis candidatos a cluster; a segunda está no fato do desconhecimento da distribuição de probabilidades para a estatística de teste em uso.

A solução da primeira questão, em geral, é feita através da utilização de heurísticas visando não avaliar todos os candidatos, mas somente um número restrito de candidatos em potencial. Dentre estas heurísticas, podemos citar o Scan Circular [1] e o Scan Elíptico [3]. Já para a segunda questão, usualmente são consideradas simulações de Monte Carlo e/ou o ajuste da distribuição Gumbel que adere bem a distribuições para máximos. Serão descritos de forma sucinta os métodos Circular e Elíptico que visam reduzir o espaço dos candidatos em avaliação.

Considere um ponto arbitrário escolhido no interior de cada uma das regiões que subdividem o mapa em estudo e denomine tal ponto por centróide. O Scan Circular [1] avalia janelas circulares centradas em cada um dos centróides do mapa com raios variando de zero até um raio máximo pré-fixado. Cada possível janela circular define uma zona z a ser avaliada considerando as regiões cujos centróides são interiores à janela circular. O raio máximo usual considera o aumento de cada uma das janelas circulares até que a zona definida por tal janela atinja uma população de no máximo 50% da população total do mapa em estudo. Dessa forma, o número de zonas z a serem avaliadas é reduzido substancialmente, por outro lado, as zonas avaliadas tendem a possuir uma forma geométrica bastante regular (quase circular) fazendo com que possíveis clusters de forma muito irregular não sejam avaliados.

O Scan Elíptico [3] se baseia em premissas similares, entretanto avalia janelas elípticas considerando variações entre as diferenças de comprimento entre o eixo maior e menor da elipse e também rotacionando por diversos ângulos as elipses em avaliação. O restante do procedimento é análogo, mas acaba por permitir a avaliação de zonas com uma maior grau de irregularidade em sua forma geométrica.

2 Objetivo

O trabalho de Almeida et al. [4] questiona a validade do procedimento inferencial utilizado no Scan Circular [1] e apresenta uma proposta eficiente para melhoria da técnica inferencial através da metodologia Data-driven. Entretanto, a conhecida deficiência do Scan Circular na avaliação de possíveis clusters de formato irregular leva a um caminho quase óbvio, verificar a existência de possíveis diferenças ao considerar a inferência usual ou a inferência Data-driven quando utilizando o Scan Elíptico, sabidamente mais robusto para detecção de clusters de forma irregular. Disto surge o objetivo de explorar a versão elíptica do Scan associada ao procedimento de inferência Data-driven visando verificar se existem diferenças evidentes entre as duas técnicas inferenciais quando utilizando o Scan Elíptico.

3 Metodologia

A utilização do procedimento inferencial clássico para o Scan Circular compara a zona obtida dos dados observados com a distribuição empírica obtida através de simulações de Monte Carlo. Uma preocupação 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 daquela obtida dos dados observados.

Os estudos envolvendo a inferência Data-driven [4] exatamente contradizem este fato. Mostra-se que a grande maioria das zonas produzidas através do procedimento de simulação não se assemelham muito à zona obtida dos dados observados, especiamente quando esta é composta por um número elevado de regiões. Em sua formulação original, o procedimento inferencial clássico se baseia 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 utiliza 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, qual é a probabilidade de encontrar uma zona mais verossímil sob a hipótese nula com a estatística de teste maior que o valor x com 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 as 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.

Analogamente será buscado aqui, avaliar as diferenças obtidas quando considera-se o mesmo procedimento na utilização do Scan Elíptico.

4 Resultados e Discussões

A versão original de utilização da inferência Data-driven [4] foi apresentada em associação ao Scan Circular utilizando um benchmark de dados reais em um mapa composto por 245 condados em 10 estados e no Distrito de Columbia, no Nordeste dos EUA, com 58.943 casos de câncer de mama no período de 1988 a 1992, para uma população de risco de 29.535.210 mulheres em 1990.

No estudo aqui apresentado, novamente utilizou-se este benchmark de dados reais, entretanto com o procedimento de detecção baseado no Scan Elíptico. Foram executadas 100.000 replicações de Monte Carlo sob a validade da hipótese nula de não existência de clusters no mapa em estudo. O conglomerado mais verossímil de cada replicação e sua quantidade de regiões foram obtidos.

O primeiro interesse era verificar a estrutura da distribuição da variável tamanho (quantidade de regiões) do cluster mais verossímil obtido em cada uma das replicações de Monte Carlo. Já era conhecida esta informação para o mapa em estudo quando utilizando o Scan Circular, nesse caso foi verificada uma grande concentração de soluções com um pequeno número de regiões (de uma a cinco regiões), ao passo que soluções com um número de regiões elevado apareciam com uma freqüência bem inferior. Antes mesmo da realização do experimento, já era esperado que este resultado fosse confirmado para a utilização do Scan Elíptico. Dessa forma, o interesse central da pesquisa aqui apresentada está em confirmar este fato e também verificar as diferenças obtidas no processo decisório (conclusão do teste de hipóteses) partindo da utilização do procedimento inferencial Data-driven.

As soluções foram classificadas de acordo com seus tamanhos visando observar através da simulação a estrutura da distribuição da variável aleatória Tamanho da zona mais verossímil como pode ser observado na Figura 1.

Assim como verificado através do Scan Circular, quando da utilização do Scan Elíptico nota-se a não existência de uniformidade na quantidade de regiões para o conglomerado identificado. Observa-se uma grande frequência de soluções compostas por um número pequeno de regiões, confirmando que o processo de decisão fica um tanto comprometido quando a solução mais verossímil obtida dos dados observados possui um número de regiões elevado.

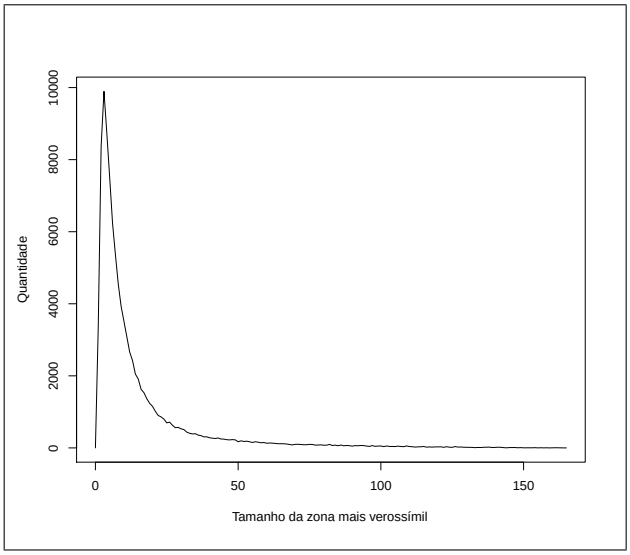

Figura 1: Frequência de soluções observados para cada tamanho de zona mais verossímil.

É importante salientar que quando o valor-p para a solução obtida dos dados observados é muito menor ou muito maior do que o nível de significância α pré-estabelecido, não há mudança na decisão de aceitar ou rejeitar a hipótese nula. Assim, apenas o processo de decisão que realmente poderia ser deteriorado ao utilizar o procedimento inferencial usual se encontra nos casos cujo valor-p se encontra próximo do valor α pré-estabelecido.

Os experimentos numéricos realizados atestam diferenças significativas nos valores críticos para um nível de significância fixo. Foram obtidos os valores críticos para os níveis de significância $\alpha = 5\%$ e $\alpha = 1\%$ considerando o procedimento inferencial usual e também o procedimento Data-driven para diversas quantidades de regiões compondo o cluster mais verossímil. Esses resultados podem ser observados através da Figura 2 para $\alpha = 5\%$ e $\alpha = 1\%$, a linha contínua representa os valores críticos para cada possível tamanho de zona enquanto a linha tracejada representa o valor crítico obtido através do procedimento usual.

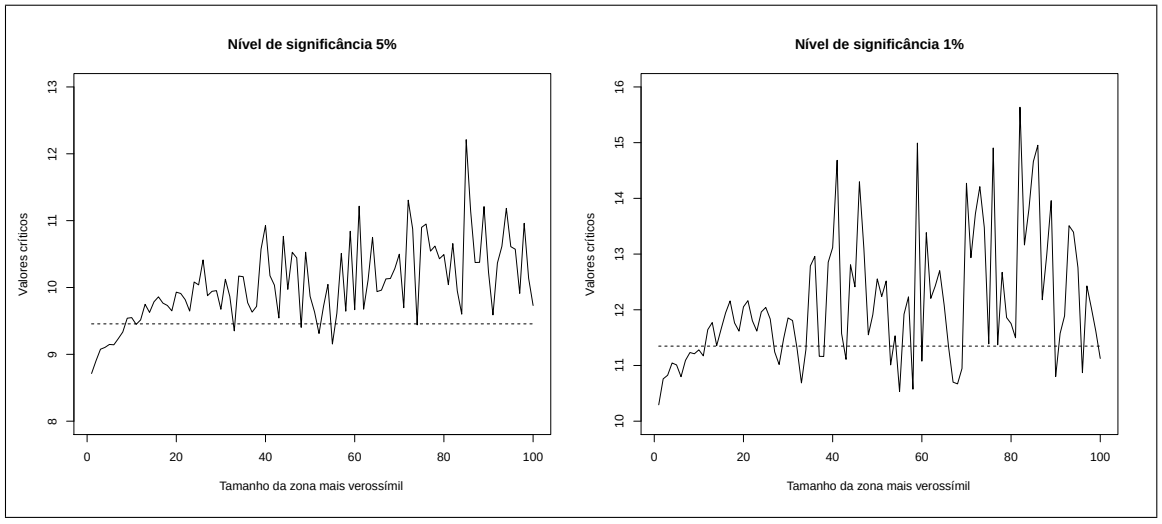

Figura 2: Valores críticos para $\alpha = 5\%$ e $\alpha = 1\%$.

Esses valores críticos foram obtidos para zonas de até 100 regiões, esta escolha se deve ao fato da amostra obtida não possuir um volume representativo de soluções para zonas com um número maior de regiões.

Observa-se que para zonas com um maior número de regiões, os valores críticos para $\alpha = 5\%$ se apresentam constantemente superiores ao valor crítico obtido através do procedimento usual.

As avaliações para o nível de significância $\alpha = 1\%$ são semelhantes, entretanto se observa uma variabilidade maior quando utilizando o procedimento Data-driven.

É possível observar que mudanças significativas são detectadas quando comparados os valores críticos obtidos através do procedimento inferencial usual e do procedimento Data-driven. É importante ressaltar que para soluções obtidas dos dados observados cujos valores-p sejam muito diferentes do nível de significância α as conclusões não sofrem alteração, entretanto para soluções cujos valores-p estejam suficientemente próximos de α as conclusões podem sofrer alterações. Desta forma, soluções anteriormente vistas como não significativas estatisticamente, podem ser agora avaliadas como soluções significativas e vice-versa.

5 Conclusões

O processo inferencial clássico utilizado na inferência de clusters espaciais empregando a Estatística Scan considera todos os clusters mais verossímeis obtidos através das replicações de Monte Carlo sob a hipótese nula, a fim de construir a distribuição empírica para a estatística de teste, independentemente do tamanho do cluster e de sua localização. Uma desvantagem potencial desta abordagem é a independência implicitamente assumida da distribuição da estatística de teste em relação à quantidade de regiões do cluster detectado. Os experimentos numéricos confirmam que esta hipótese de independência não é verdadeira. Dado que o cluster mais verossímil observado tem um tamanho k, a inferência clássica avalia sua significância com base no comportamento de todos os clusters detectados ao longo das simulações de Monte Carlo, que na maioria são compostas por uma diferente quantidade de regiões e logo populações de risco bastante diferentes do cluster mais verossímil obtido dos dados observados.

Neste trabalho foi discutida a proposta alternativa, Inferência Data-driven, que leva em conta apenas os clusters mais verossímeis encontrado cuja quantidade de regiões é idêntica à quantidade de regiões do cluster mais verossímil obtido dos dados observados. Esta abordagem utiliza uma comparação mais específica, evitando assim que o comportamento de clusters de tamanho muito pequenos seja utilizado no processo decisório.

Os resultados levam a acreditar que não apenas para o Scan Elíptico, mas na verdade para qualquer estratégia de detecção de clusters baseada na estatística Scan o procedimento Data-driven seja mais robusto. Portanto se torna a ferramenta a ser utilizada na concepção do processo de estimativa para os valores-p associados à soluções obtidas dos dados observados.

Referências

[1] Kulldorff, M., Nargarwalla, N., Spatial disease clusters: detection e inference, *Statistics in Medicine*, **14**, 799-810, 1995.

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

[3] Kulldorff, M., Huang, L., Pickle, L., Duczmal, L., An elliptic spatial scan statistic, *Statistics in Medicine*, **25**, 3929-3943, 2006.

[4] Almeida, A. C. L., Duarte, A. R., Duczmal, L., Oliveira, F. L. P., Takahashi, R. H. C., Data-driven inference for the spatial scan statistic, *International Journal of Health Geographics*, **10**, 47, 2011.