*Anais do XII Encontro Mineiro de Estatística - MGEST 2013. Uberlândia - 05 e 06 de setembro de 2013.*
*Revista Matemática e Estatística em Foco - ISSN: 2318-0552*

# Estatística scan espacial multi-objetivo: uma nova estatística para a detecção de clusters espaciais

**Flávio dos Reis Moura** [1]
**Luiz Henrique Duczmal** [2]
**Ricardo Tavares**[1]

**Resumo**:

Este trabalho apresenta uma nova estatística para a detecção de clusters espaciais. Considere um mapa com $m$ regiões em que se conhece os casos observados de um certo evento de interesse (por exemplo, infecção ou óbito por alguma doença) e a população de cada área. A metodologia proposta analisa o teste da razão de verossimilhança (LLR) em alguns conjuntos de incidência destacada no mapa, e considera dois objetivos: o LLR e a ocupação circular (OC), que é uma medida de regularidade do formato do cluster. As regiões são ordenadas de forma decrescente conforme os valores da estatística scan, que são os conjuntos seletivos. A proposta considera clusters de formato regular e/ou irregular, além de generalizar o conceito de zona para abranger conjuntos não-conexos, além daqueles conexos. Um procedimento Monte Carlo é usado para a avaliação da significância da nova estatística. Uma aplicação a dados de homicídios no estado de Minas Gerais também foi realizada. Os resultados mostraram que o método scan espacial multi-objetivo é um procedimento eficiente na detecção de clusters com formato regular e irregular, conexo e desconexo, que permite a visualização da estrutura de clusterização do mapa. A presença de "joelhos" nos conjuntos Pareto indicou transições repentinas na estrutura dos clusters. Além disso, o cálculo da significância bi-objetivo de cada solução se mostrou uma ferramenta eficaz para a escolha do cluster mais significativo.

**Palavras-chave**: *Estatística Scan Espacial, Ocupação Circular, Estatística Scan Multi-Objetivo*.

## 1 Introdução

Algoritmos para a detecção e avaliação da significância estatística de clusters espaciais são importantes ferramentas geográficas em Epidemiologia, vigilância sindrômica, monitoramento de crimes e ciências ambientais. A elucidação da etiologia das doenças, a disponibilidade de alarmes confiáveis para

[1]DEEST - Universidade Federal de Ouro Preto. Email: *prof.flaviomoura@gmail.com*
[2]EST - Universidade Federal de Minas Gerais. Email: *duczmal@gmail.com*
[1]DEEST - Universidade Federal de Ouro Preto. Email: *prof.ricardotavares@gmail.com*

detectar surtos intencionais ou não de uma certa doença, o estudo de padrões espaciais de atividades criminais e monitoramento geográfico de mudanças ambientais são tópicos recentes de intensa pesquisa.

Atualmente, o método mais popular para encontrar clusters espaciais é a estatística espacial scan de [3]. A significância do cluster mais provável é estimada através de simulações de Monte Carlo. A Estatística Scan Circular varre todos os possíveis conjuntos de regiões conectadas cujos centros estejam dentro de um círculo com raio variando conforme o percentual de população dentro deste círculo. Muitas propostas têm sido sugeridas para encontrar clusters espaciais de formato arbitrário, mas a maioria delas sendo uma extensão da estatística scan de Kulldorff. Métodos para encontrar clusters espaciais e suas aplicações foram revisados em [2]. Em muitas situações necessitamos reconhecer clusters espaciais em uma classe geométrica mais geral.

Este trabalho tem como objetivo propor uma extensão da estatística scan espacial de Kulldorff utilizando técnicas de otimização multi-objetivo para levar em consideração o formato arbitrário das zonas candidatas a clusters e também os vários níveis de clusterização presentes em um mapa. O método desta proposta foi implementado em Dev C++, e teve o auxílio do R para confeccionar os mapas e gráficos.

## 2 Estatística Scan Multi-Objetivo

Muitos problemas do mundo real apresentam vários objetivos que podem ser conflitantes entre si. Neste contexto, a *otimização* é a tarefa de encontrar uma ou mais soluções que atendam a esses objetivos. Na otimização multiobjetivo, geralmente, não há uma única solução ótima, mas um conjunto de alternativas com *compromissos* (*trade-offs*) diferentes, chamado de *soluções ótimas de Pareto*. O *conjunto Pareto-ótimo* é formado por todas as soluções ótimas de Pareto.

A Estatística Scan Espacial é a medida mais utilizada para quantificar a intensidade de um cluster. [1] trabalharam com técnicas multiobjetivo usando o algoritmo genético para localizar clusters espaciais levando em conta, além da intensidade do cluster, a sua regularidade geométrica. A nova estatística proposta aqui foi apresentada em [4, 5]. O método foi proposto para analisar mais precisamente os vários níveis de clusterização que surgem naturalmente em um mapa dividido em $m$ regiões. As duas definições a seguir são os dois principais pontos da nossa proposta.

**Definição 2.1 (Conjuntos Seletivos)** *Os conjuntos seletivos são obtidos a partir das regiões ordenadas segundo suas verossimilhanças. Suponha um mapa com $m$ regiões $\{r_1, r_2, \cdots, r_m\}$. Defina $L_i = LLR(r_i)$ como sendo o valor do logaritmo da razão de verossimilhança (da estatística scan espacial) da zona contendo apenas a região $r_i$. Ordene as $m$ regiões do mapa de modo que $L_1 > L_2 > \cdots > L_m$. O subconjunto $R_j = \{r_1, r_2, \cdots, r_j\}$ é o conjunto seletivo de tamanho $j$ ($j = 1, 2, \cdots, m$).*

**Definição 2.2 (Ocupação Circular)** *A ocupação circular de uma zona $z$ é definida como a razão de sua*

*população pela população do menor círculo que a contém. Dado um conjunto seletivo $S$ e um círculo $C$, seja $z$ a zona formada pelas regiões de $S$ cujos centróides estão dentro de $C$. Seja $P(z)$ a população de $z$ e seja $P(C)$ a população de todas as regiões do mapa original cujos centróides estão dentro de $C$. A ocupação circular da zona $z$, $OC(z)$, é dada por $OC(z) = max_{r_j \in z}\{P(z)/P(C_{r_j})\}$ em que $C_{r_j}$ é o menor círculo centrado na região $r_j$ que contém a zona $z$.*

Desta forma, para cada zona $z$ desejamos maximizar dois objetivos: a $LLR(z)$ e a $OC(z)$. Num conjunto $C$ formado por $n$ zonas $z_1, z_2, \cdots, z_n$, considere os pares ordenados

$$(LLR(z_i), OC(z_i)).$$

O uso de técnicas multiobjetivo é necessário para avaliar possíveis clusters em níveis de regularidade e conexidade diferentes. O valor de $LLR$ é calculado para cada uma das $m$ regiões do mapa e ordenados decrescentemente. Seja $R(j)$ o conjunto contendo as j primeiras regiões. A extensão multiobjetivo do algoritmo scan circular é aplicada sucessivamente para cada conjunto $R(j)$. Em cada círculo, a zona candidata a ser um cluster é composta pelas regiões pertencentes a $R(j)$ e que estão dentro do círculo. Na prática, escolhe-se apenas alguns poucos valores de $j$ tais como $\lceil m \rceil$, $\lceil m/2 \rceil$, $\lceil m/4 \rceil$, $\cdots$, $\lceil 1 \rceil$. Para cada valor de $j$, constrói-se um conjunto Pareto-ótimo $P(j)$. Com base em todos esses conjuntos de Pareto, obtém-se o conjunto de Pareto-ótimo Global $P(0)$. Por fim, um procedimento de Monte Carlo é implementado para avaliar a significância dos clusters desse conjunto $P(0)$.

## 3 Resultados

Para avaliar o desempenho do método, foram construídos cinco clusters artificiais de formas e com tamanhos populacionais diferentes. Os clusters simulados foram: *A* (cluster circular), *B* (cluster em formato de tira), *C* (cluster com população pequena), *D* (cluster com população grande) e *E* (cluster desconexo). O cluster *E* é desconexo e foi construído através de um processo de difusão de doenças transmitidas por aves contaminadas que migram de um município à outro próximo, porém não vizinho ao município de partida, e representa a propagação da gripe aviária em patos selvagens. Algumas características desses cinco clusters estão na Tabela 1.

Conforme mostra a Tabela 2, para os clusters com formato mais regular (*A* e *D*), percebe-se que existe pelo menos uma faixa de valores da ocupação circular em que a Estatística Scan Multi-Objetivo tem poder muito próximo da Estatística Scan Circular, porém, com a vantagem da não existência de subestimação e superestimação do cluster. Para os clusters com geometria mais irregular (*B* e *C*), a Estatística Scan Multi-Objetivo tem poder de teste superior à Scan Circular em pelo menos uma faixa cuja ocupação circular seja menor que 1,0. Para o cluster desconexo *E*, observamos que para valores

Tabela 1: *Características dos quatro clusters artificiais conexos.*

| Características | *A* | *B* | *C* | *D* | *E* |
|---|---|---|---|---|---|
| Número de municípios | 22 | 5 | 11 | 7 | 9 |
| Número de habitantes | 383.488 | 89.010 | 57.521 | 3.725.982 | 143.340 |
| Risco Relativo | 1,506 | 2,152 | 2,496 | 1,171 | 1,873 |
| Ocupação Circular | 0,880 | 0,434 | 0,099 | 0,953 | 0,372 |

Tabela 2: *Poder da Estatística Multi-Objetivo para os cinco clusters artificiais avaliados.*

| Ocupação Circular | *A* | *B* | *C* | *D* | *E* |
|---|---|---|---|---|---|
| 1 (Est. Scan) | 0,7870 | 0,4380 | 0,0481 | 0,4632 | 0,4380 |
| 0,90 - 1,00 | 0,7807 | 0,4300 | 0,0482 | 0,4631 | 0,4300 |
| 0,80 - 1,00 | 0,7799 | 0,4224 | 0,0524 | 0,4628 | 0,4224 |
| 0,70 - 1,00 | 0,7910 | 0,4284 | 0,0645 | 0,4624 | 0,4284 |
| 0,60 - 1,00 | 0,7868 | 0,4771 | 0,0899 | 0,4623 | 0,4771 |
| 0,50 - 1,00 | 0,7497 | 0,4812 | 0,1076 | 0,4611 | 0,4812 |
| 0,40 - 1,00 | 0,7178 | 0,4702 | 0,1219 | 0,4591 | 0,4702 |
| 0,30 - 1,00 | 0,6613 | 0,4444 | 0,1368 | 0,4553 | 0,4444 |
| 0,20 - 1,00 | 0,5744 | 0,3935 | 0,1697 | 0,4490 | 0,3935 |
| 0,10 - 1,00 | 0,4221 | 0,3150 | 0,2095 | 0,4252 | 0,3150 |
| 0,00 - 1,00 | 0,1609 | 0,1985 | 0,2258 | 0,1726 | 0,1985 |

decrescentes de $b$ a Scan Multi-Objetivo tem seu poder aumentado em comparação com o scan circular usual.

O novo método foi aplicado à dados reais de homicídios, da última década, para o mapa dos 853 municípios mineiros. Cada um dos conjuntos seletivos com $a$ igual a $0.002$, $0.004$, $0.008$, $0.016$, $0.032$, $0.032$, $0.064$, $0.125$, $0.250$, $0.500$ e $1.000$, indicados com diferentes símbolos. Alguns clusters do conjunto de Pareto Global são indicados pelas setas, com seus respectivos mapas e apresentados na Figura 1.

Pode-se ver que nessa aplicação o cluster de maior significância dentre os clusters do Pareto Global é o cluster do segundo mapa incluso na parte superior da Figura 1 que é justamente o cluster conexo de maior LLR.

## 4 Conclusões

A Estatística Scan Espacial Multi-Objetivo mostrou-se muito eficiente na detecção de clusters de formato regular e irregular, conexo e desconexo, controlando ou evitando as superestimações e subestimações. Observou-se também que saltos significativos no conjunto de Pareto resultam em mudanças significativas no formato do cluster, inclusive modificando o grau de conexidade da solução.

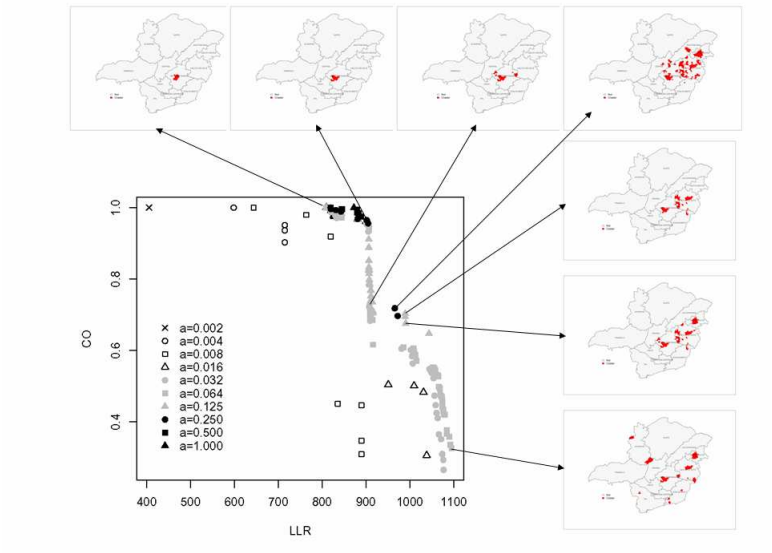


Figura 1: Conjunto de Pareto com a visualização de algumas soluções de clusters para o mapa.

## Referências

[1] DUCZMAL, L. H.; CANÇADO, A. L. F.; TAKAHASHI, R. H. C.. Delineation of Irregularly Shaped Disease Clusters through Multi-Objective Optimization. *Journal of Computational & Graphical Statistics*, v. 17, p. 243-262, 2008.

[2] DUCZMAL, L. H.; DUARTE, A. R.; TAVARES, R.. *Extensions of the scan statistic for the detection and inference of spatial clusters*. In.: Scan Statistics - Methods and Applications. Hamilton: Springer, 2009.

[3] KULLDORFF, Martin. A Spatial Scan Statistic. *Communications in Statistics. Theory and Methods*, v. 26 (6), p. 1481-1496, 1997.

[4] MOURA, Flávio dos Reis. *Detecção de clusters espaciais via algoritmo scan circular seletivo*. Dissertação de Mestrado, Universidade Federal de Minas Gerais, Departamento de Estatística, Belo Horizonte-MG, 2006.

[5] TAVARES, Ricardo. *Extensões da Estatística Scan Espacial utilizando Técnicas de Otimização Multiobjetivo*. Tese de Doutorado, Universidade Federal de Minas Gerais, Departamento de Estatística, Belo Horizonte-MG, 2009.