Ronaldo Lage Figueiredo

# Detecção de clusters usando a Estatística Scan Espacial Circular em conjuntos seletivos e um fator de penalização: a ocupação circular

Dissertação de Mestrado apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof. Sabino Ferreira
Co-orientador: Prof. Ricardo Tavares

Universidade Federal de Minas Gerais
Belo Horizonte, Dezembro de 2010

# Dedicatória

Dedico este trabalho à minha família e amigos.

*No meio do Caminho*

No meio do caminho tinha uma pedra
tinha uma pedra no meio do caminho
tinha uma pedra
no meio do caminho tinha uma pedra.
Nunca me esquecerei desse acontecimento
na vida de minhas retinas tão fatigadas.
Nunca me esquecerei que no meio do caminho
tinha uma pedra
Tinha uma pedra no meio do caminho
no meio do caminho tinha uma pedra.

Carlos Drummond de Andrade (1928)

# Agradecimentos

Apesar de tantas “pedras no meio do caminho”, venço uma importante etapa da minha vida, a conclusão desta pesquisa. Trabalho conquistado com a benção de Deus, e ao incondicional apoio dos meus queridos e amados Pais, Geraldo e Maria de Fátima, por sempre acompanharem a minha jornada e torcerem para o meu sucesso.

Aos meus irmãos Marcelo, Robson, Douglas e minhas queridas sobrinhas, Thaís e Giovanna, pelo apoio familiar.

Aos meu queridos amigos Adilson, Angélica, Bruno, Cleide, Dayanne, Edson, Gerson, João, José Luiz, Leonardo, Luciana, Luís, Reginaldo, Sérgio e Spencer, amigos estes que me ajudaram nos momentos difíceis e compartilharam as alegrias das conquistas.

Aos professores do Departamento de Estatística da UFMG, em especial ao meu estimado orientador, Prof. Sabino Ferreira e ao Prof. Luiz Duczmal, por sempre estarem dispostos a esclarecer quaisquer dúvidas para o desenvolvimento do nosso trabalho. A eles tenho um enorme respeito e a admiração pelos excelentes mestres que são.

Ao Prof. Fabrício Goecking Avelar do Departamento de Matemática da UNIFAL-MG.

Às amigas Rogéria, Rose e Marcinha do Departamento de Estatística da UFMG, por sempre torcerem pelo meu sucesso e me ajudarem pelas energias positivas.

Como parte dessa conquista, agradeço também aos meus queridos professores da UFOP, em especial o meu co-orientador Prof Ricardo Tavares, o Prof. Flávio Moura e a Prof[a] Maria Cláudia, pelos incentivos e sugestões.

A todos que contribuíram de maneira positiva para este trabalho.

Com tanto apoio consegui remover algumas “pedras do meio do caminho” e alcançar a mais uma de várias conquistas de minha vida.

A todos, o meu Muito Obrigado.

# Resumo

A Estatística Scan Espacial Circular proposta por Kulldorff tem sido bastante utilizada em algoritmos para detecção e inferência de clusters em mapas onde ocorrências (doenças, homicídios, etc.) estão distribuídos aleatoriamente entre as regiões que compõe o mapa. estudada nesses últimos anos. A importância desses estudos acerca da Estatística Scan Espacial é propor um algoritmo que consiga detectar clusters em situações reais.

Este trabalho propõe selecionar um conjunto de regiões que possuem os maiores valores da razão de verossimilhança (LR ou o logaritmo da razão de verossimilhança LLR) no mapa inicial. A essa seleção denomina-se conjuntos seletivos das *a%* regiões com maiores LR, proposto por Moura (2006). A partir dessas regiões selecionadas, aplicaremos o algoritmo Scan Espacial Circular com um fator de penalização: a Ocupação Circular. A estatística de teste é definida como sendo o produto da LLR com a Ocupação Circular nas regiões selecionadas pelo conjunto seletivo. A esse algoritmo de detecção de clusters dá-se o nome de Estatística Scan Seletivo.

Como o algoritmo Scan Espacial Circular busca maximizar a estatística de teste, ou seja, encontrar o cluster com o maior valor da LR, o conjunto seletivo encontra o cluster que maximiza o valor do produto da LLR com a Ocupação Circular nas regiões selecionadas por cada conjunto seletivo. Devido à natureza da formação dos conjuntos seletivos, os clusters detectados pelo Scan Seletivo não são necessariamente formados de regiões conexas. Neste trabalho, mostramos que as partes conexas de um cluster não conexo detectado tem uma correpondência com a hierarquia de clusters primário, secundário,

terciário, etc, identificados pelo algortimo Scan Circular de Kulldorff.

Utilizamos simulações de Monte Carlo, para determinar o poder de detecção, a sensibilidade e o poder preditivo do algoritmo Scan Seletivo, e verificamos que houve uma melhoria em relação à detecção de clusters irregulares quando comparado à Estatística Scan Circular de Kulldorff. Apresentamos uma comparação do desempenho do Scan Seletivo com outros algoritmos eficientes para a detecção de clusters irregulares. Finalmente mostramos uma aplicação do Scan Seletivo na detecção de clusters em casos reais de homicídios no estado de Minas Gerais.

Para a escolha do melhor parâmetro de seleção da Estatística Scan Seletivo, utilizamos a distribuição generalizada de valores extremos.

**Palavras-chave**: cluster espacial; estatística Espacial Scan Circular; conjuntos seletivos; algoritmo scan seletivo; ocupação circular; função generalizada de valores extremos.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

A Estatística Scan Espacial proposta por Kulldorff (1997) tem sido extensivamente utilizada por algoritmos de detecção de clusters espaço-temporais. Estes algoritmos são ferramentas importantes em estudos epidemiológicos e de vigilância de doenças orientando órgãos governamentais de saúde pública para a aplicação prioritária de recursos no conjunto das regiões onde a incidência de doenças é maior.

Esta dissertação propõe apresentar uma discussão do uso da Estatística Scan Espacial Circular e o respectivo algoritmo na detecção de clusters de regiões conexas em um mapa de ocorrências aleatórias. O algoritmo Scan Circular apresenta suas soluções na forma de uma hierarquia que as classifica em primárias, secundárias, terciárias, etc de acordo com a hierarquia dessas soluções relativo aos respectivos valores das estatística de teste. Apresentaremos o algoritmo para detecção de clusters que denomimamos Scan Seletivo que se baseia nos conceitos de conjuntos seletivos e ocupação circular. O algoritmo é uma extenção da Estatística Scan Espacial Circular com uma nova estatística de teste denominada Estatística Scan (Mono) Seletivo que permite a detecção de clusters formados por conjuntos não necessariamente conexos de regiões. Nesse texto, denotaremos Estatística Scan (Mono) Seletivo por Estatística Scan Seletivo.

## 1.1 Objetivos

Um dos entuito deste trabalho é mostrar que o algoritmo Scan Seletivo com seu novo conceito de conjuntos seletivos combinado com uma nova função de penalização, a Ocupação Circular, é um algoritmo eficiente (na sua versão mono-objetivo) na detecção de clusters irregulares ou não.

Outro objetivo desse trabalho é investigar se um possível cluster detectado pelo Scan Seletivo e formado por um conjunto não conexo de regiões corresponde à hierarquia de soluções encontradas pelo Scan Circular. Ou seja, será que cada uma das diversas partes conexas de um possível cluster detectado pelo Scan Seletivo corresponde as soluções primárias, secundárias, terciárias do Scan Circular?

Aplicaremos a metodologia em um banco de dados reais da população com casos de câncer de mama no Nordeste do Estados Unidos. Esse banco, consiste em 245 condados em 10 estados mais o distrito de Columbia com uma população total de risco de 29.535.210 mulheres com um total 58.943 casos no período de 1988 a 1992. Também faremos um estudo de casos reais de homicídios no estado de Minas Gerais de 1998 a 2002.

## 1.2 Estrutura do trabalho

Este trabalho apresenta no Capítulo 2 a Estatística Scan Espacial de Kulldorff, em seguida, no Capítulo 3 a Estatística Scan Seletivo. Os resultados de nossas análises são apresentados no Capítulo 4 e no Capítulo 5 as conclusões deste trabalho.

# Capítulo 2

# Estatística Scan Circular

## 2.1 A Estatística de Kulldorff

Para descrever a Estatística Scan Circular proposta por Kulldorff (1997), definiremos o conceito de zona e cluster. Considere um mapa dividido em $m$ regiões, onde cada região $R_i$ $(i = 1, ..., m)$ tem uma população $N_i$ e uma contagem $C_i$ que representa o número de casos de um determinado tipo de ocorrência (crime, doenças, etc.) que estão distribuídos aleatoriamente pelas regiões do mapa. Denotamos por $N = \sum_{i=1}^{m} N_i$ e $C = \sum_{i=1}^{m} C_i$ respectivamente a população e o número total de ocorrências no mapa. Uma zona $z$ é definida como um subconjunto conexo de regiões do mapa. O conjunto de todas as zonas $z$ será denotado por $Z$. Um cluster representa um subconjunto de regiões do mapa em que a ocorrência de casos é discrepante do restante do mapa, seja por ser a contagem dos casos alta ou baixa demais.

Usa-se o modelo de Poisson para descrever a distribuição dos casos entre as regiões do mapa. Consideramos que o número de casos $C_i$ é uma variável aleatória com distribuição de Poisson cujo parâmetro $\lambda_i$ que representa o número esperado de casos é tal que $\lambda_i = p_i N_i$, onde $p_i$ representa a probabilidade de um indivíduo da região $R_i$ ser um caso ou ocorrência. Escrevemos que $C_i \sim Poisson(\lambda_i = p_i N_i)$. Neste caso o número de casos

$C_z$ em uma zona $z$ será uma variável de Poisson com parâmetro $\lambda_z = pN_z$ sendo $N_z$ a população da zona $z$ e $p$ a probabilidade de um indivíduo qualquer da zona $z$ ser um caso.

A situação em que a probabilidade $p$ de um indivíduo ser um caso é igual em qualquer parte do mapa sendo estimada por $\frac{C}{N}$ é uma hipótese que supõe a não existência de cluster no mapa. Esta é denominada hipótese nula $H_0$. A hipótese alternativa supõe a existência de pelo menos uma zona $z_c \in Z$ onde a probabilidade $p$ de um indivíduo ser um caso seja maior ou menor do que a probabilidade $q$ de um indivíduo em uma região qualquer do mapa fora da zona $z_c$ ser um caso. O problema de detecção de clusters pode ser visto como um teste de hipóteses onde

$$H_0 : p = q$$

$$H_1 : p > q \; ou \; p < q$$

em particular, nosso interesse é o caso onde $p > q$.

Defina-se *L(z)* como sendo a verossimilhança da zona $z$ sob a hipótese alternativa $H_1$ e $L_0$ a verossimilhança sob $H_0$. Considere $\mu(z) = \frac{C}{N}N_{(z)}$ o número esperado de casos dentro da zona $z$ sob $H_0$, então definimos a razão de verossimilhança como sendo,

$$LR(z) = \begin{cases} \left(\frac{C_z}{\mu(z)}\right)^{C_z} \left(\frac{C-C_z}{C-\mu(z)}\right)^{C-C_z}, & se \; \frac{c(z)}{\mu(z)} > 1 \\ 1 & \text{c.c.} \end{cases} \tag{2.1}$$

Para cada zona $z$ teremos uma LR(z), Kulldorff (1997) definiu a Estatística Scan Espacial como

$$T = max_{z \in Z}\{LR(z)\} \tag{2.2}$$

onde $Z$ é o conjunto de todas as zonas $z$. A zona $z_c$ que maximiza $LR(z)$ será definida como zona mais *verossímil*. Outra maneira de definir a Estatística Scan Espacial é usar o logaritmo da razão de verossimilhança, $LLR(z) = log\{LR(z)\}$, no processo de maximização da equação 2.2.

Em particular a *Estatística Scan Circular* será definida como o máximo no conjunto de todas as janelas circulares com raios variáveis centradas nos centróides de cada região. Esses raios variam até que um percentual máximo especificado da população total esteja contida no círculo, por exemplo 30% da população total.

Cabe ressaltar, que a Estatística Scan Espacial de Kulldorff detecta clusters formados por *conjunto conexos de regiões*. Duas regiões são conectadas quando compartilham uma fronteira.

## 2.2 Significância Estatística

A princípio a zona $z_c$ que maximiza a razão de verossimilhança é uma candidata a cluster. Somente após a verificação de sua significância estatística a zona $z$ poderá ter seu status alterado para cluster detectado.

Para testar a significância da estatística de teste, utilizaremos as simulações de Monte Carlo como apresentado em Dwass (1957). Essa simulação consiste em construir milhares de réplicas do mapa original em que o número total de casos C está fixo e os casos em cada região são distribuídos aleatoriamente sob $H_0$. Para cada réplica teremos um valor da estatística T e o conjunto delas obtido pela simulação gerará uma distribuição empírica da estatística de teste. O p-valor da estatística de teste $LR(z_c)$ para o mapa dos casos observados pode ser estimado determinando o posto ocupado pelo seu valor no meio dos valores da distribuição empírica da estatística de teste sob $H_0$, ou seja, o p-valor do cluster mais provável do mapa original é estimado como sendo a proporção entre o número de valores empíricos de $LR(z)$ que ultrapassam o valor de $LR(z_c)$ pelo número total de mapas gerados aleatoriamente na simulação, conforme a Figura 2.1.

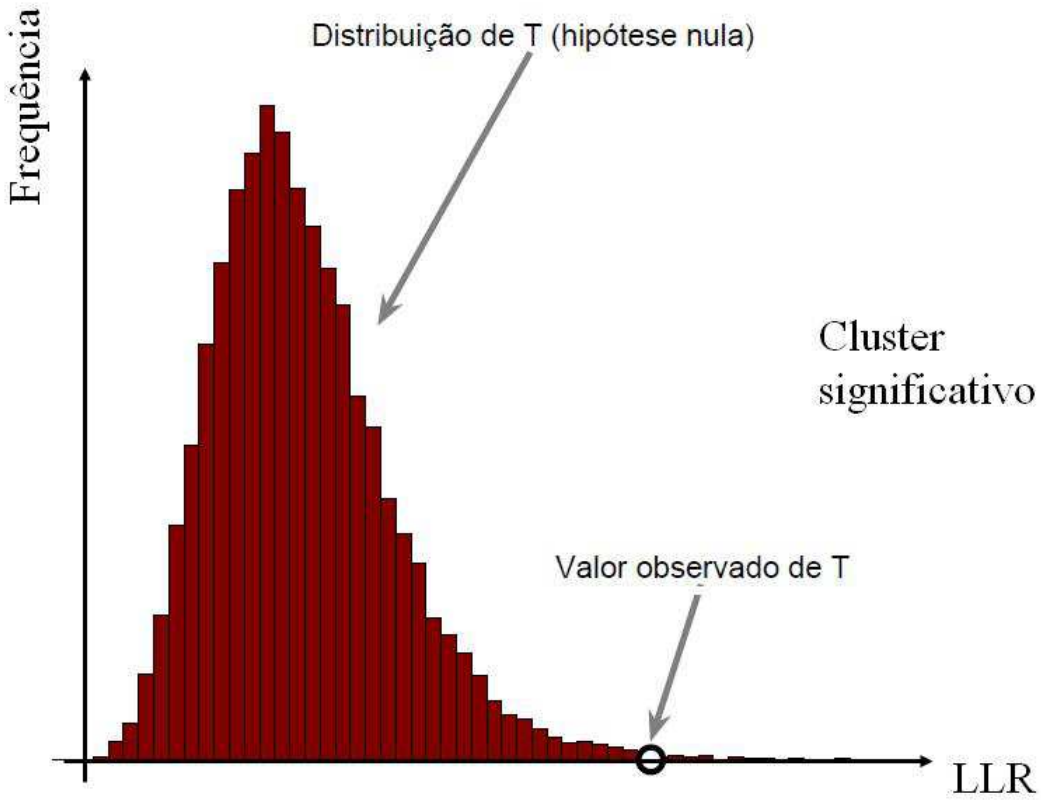



Figura 2.1: Distribuição empírica da estatística de teste

### 2.2.1 Cálculo do valor p através de um ajuste semi-paramétrico usando a distribuição generalizada de valores extremos (GEV)

Para encontrar o p-valor da estatística de teste em cada conjunto seletivo, utilizou-se um ajuste semi-paramétrico dos máximos para uma distribuição generalizada de valores extremos (GEV), alguns trabalhos podem ser visto em Abrams et al. (2006) e Duczmal et al. (2008). Com base nas 1000 réplicas Monte Carlo sob $H_0$ obtivemos valores da estatística de teste que foram usados para estimar os parâmetros de uma distribuição GEV. A partir desses parâmetros, foi possível calcular os quantis da GEV referentes ao valor da estatística de teste para o mapa de casos observados em cada conjunto seletivo. Esse ajuste foi feito com base no pacote *evd* (Stephenson, 2002) do ambiente computacional R (R Development Core Team, 2010).

Sejam $X_1, X_2, \cdots, X_n$ uma sequência de variáveis aleatórias. Então define-se o máximo como $M_n = max\{X_1, X_2, \cdots, X_n\}$.

A função de distribuição da GEV com os parâmetros $a$ (de locação), $b$ (de escala) e $s$ (de forma), no pacote evd do R, é parametrizada da seguinte maneira

$$G(z) = exp\left[-\{1+s(z-a)/b\}^{-1/s}\right]$$

para $1+s(z-a)/b > 0$, com $b > 0$.

Se $s = 0$ a distribuição é definida pela continuidade. Se $1+s(z-a)/b \leq 0$, o valor de $z$ é maior que o ponto de extremidade superior (se $s < 0$), ou menor que o ponto de extremidade inferior (se $s > 0$). A forma paramétrica da GEV engloba as distribuições Gumbel, Frechet e Weibull Inversa, que são obtidas para $s = 0$, $s > 0$ e $s < 0$, respectivamente. Maiores detalhes podem ser vistos em (Jenkinson, 1985).

A função responsável pela estimação dos parâmetros e de seus erros padrão utiliza uma aproximação numérica proposta por Smith (1985).

## 2.3 Algoritmo da Estatística Scan Circular de Kulldorff

A seguir apresentaremos o algoritmo de detecção de cluster proposto por Kulldorff (1997). Consideremos um mapa de ocorrências aleatórias dividido em *m* regiões. Define-se o centróide de cada região como um ponto arbitrariamente escolhido nesta região do mapa.

1. Escolher o centróide de uma das regiões em estudo;

2. Representamos o conjunto das distâncias entre dois centróides quaisquer em uma matriz simétrica denominada matriz de distâncias. Cada linha i da matriz representa as distâncias entre o centróide *i* e os demais centróides das regiões do mapa. Ou seja, cada linha *i* da matriz de distâncias representa um vetor contendo as distâncias do centróide da região *i* com os centróides das demais regiões.

3. Em seguida, para cada um desses vetores as distâncias são ordenadas em ordem crescente.

4. Centrada em cada uma das regiões construímos um círculo de raio variável, de tal forma que, o raio desse círculo aumente de acordo com as distâncias crescentes até que a população das regiões englobadas pelo círculo atinja um percentual máximo pré estabelecido da população total. Para cada círculo o número de casos e população são atualizadas e calcula-se a razão de verossimilhança apresentada na equação (2.1).

5. Calculamos o valor da estatística de teste (2.2).

6. Utilizar a simulação de Monte Carlo para avaliar a significância do teste, como descrito anteriormente.

7. Se a $H_0$ for rejeitada, a zona $z_c$ que maximiza T será o cluster mais verossímel ou provável.

O algoritmo Scan Circular descrito anteriormente dará como resultado o cluster mais verossímil. Esse cluster é classificado como *cluster primário*. O segundo maior valor que maximiza a estatística T, será classificado como *cluster secundário*. E de maneira análoga, temos os *clusters terciário* e *quaternário,* assim em diante. Um trabalho que compara esses clusters pode ser visto em Lima (2004).

A Estatística Scan Espacial de Kulldorff é mais indicada para detecção de um único cluster bem definido, pois apresenta grande poder de teste, ou seja, o teste baseado na Estatística de Kulldorff é uniformemente mais poderoso para detecção de clusters como mostra Kulldorff (1997). Esse poder diminui no caso do mapa em estudo apresentar mais de um cluster ou cluster de formato muito irregular como descrito em Kulldorff et al. (2003) e Duczmal et al. (2006).

A Estatística Scan Circular tem o seu poder reduzido ao tentar detectar cluster que não tenha o formato aproximadamente circular. Isto acarretaria uma redução do poder do teste que está quase sempre associada a superestimação (cluster detectado maior do que o cluster real), Figura 2.2 ou subestimação (cluster detectado menor do que o cluster real), Figura 2.3.

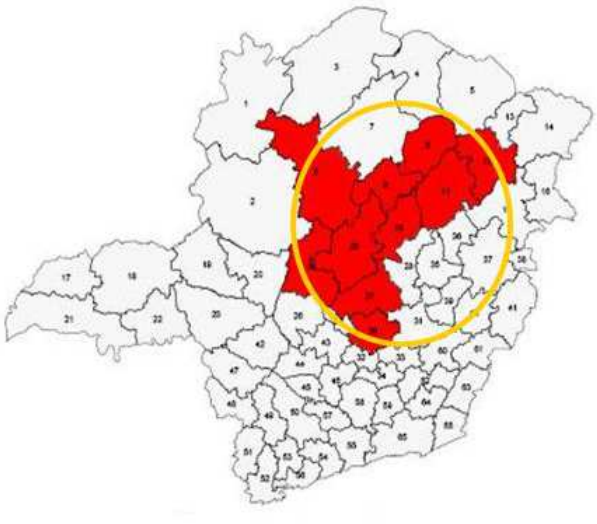


Figura 2.2: Superestimação de um cluster.

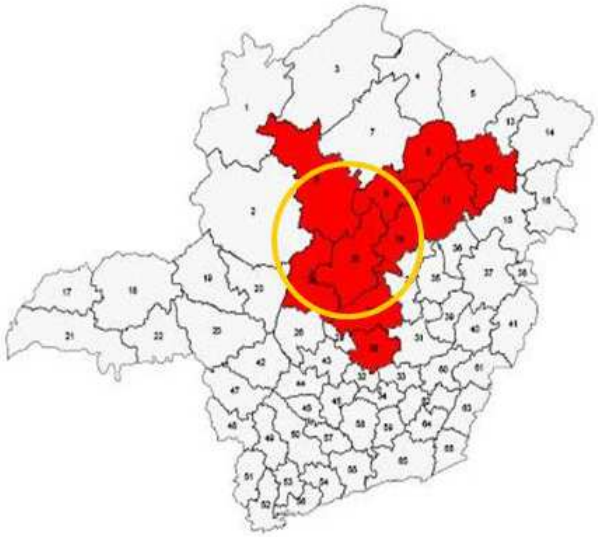


Figura 2.3: Subestimação de um cluster.

# Capítulo 3

# Estatística Scan SeletivO

A Estatística Scan SeletivO proposta por Moura (2006) é uma extensão da Estatística Scan Espacial de Kulldorff (1997) que generaliza a varredura circular de forma a englobar um conjunto não necessariamente conexo de regiões.

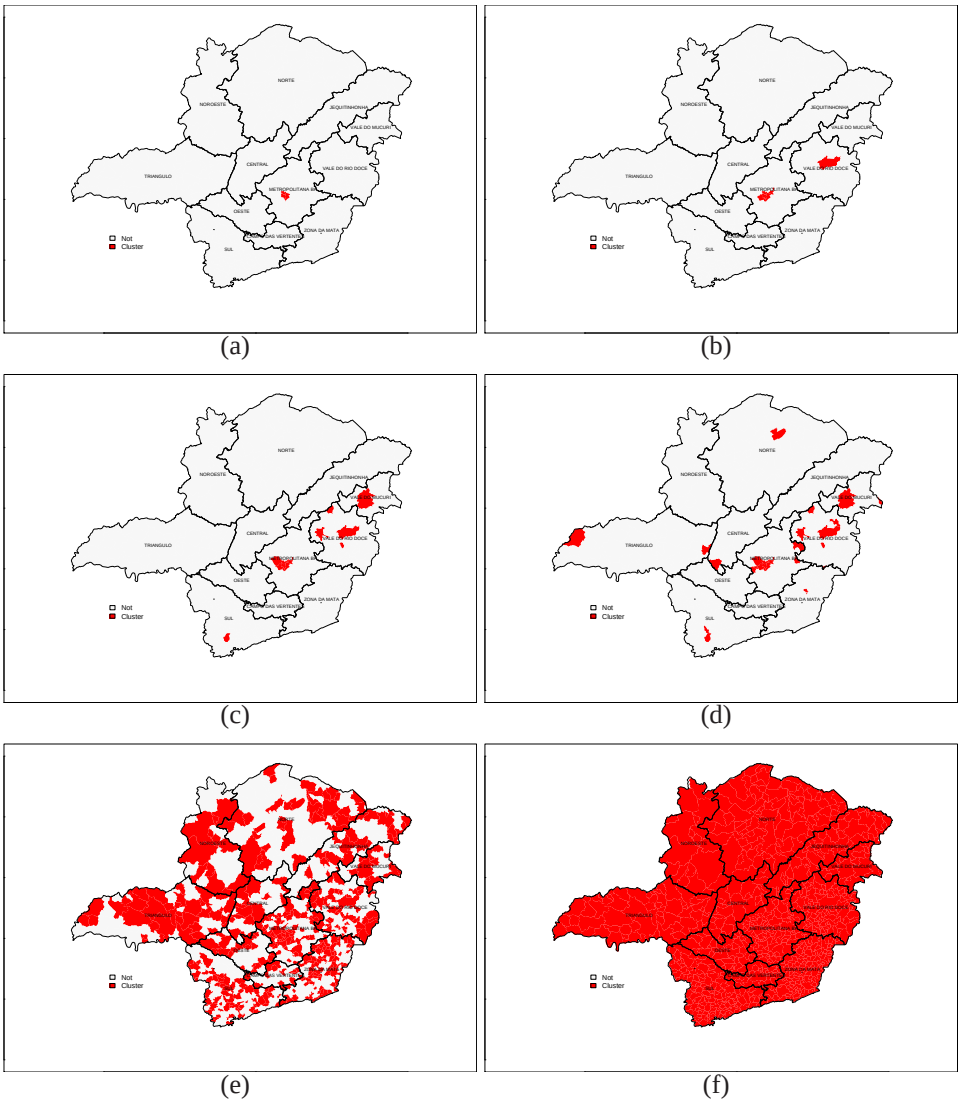
A seguir apresentaremos dois conceitos fundamentais: *conjuntos seletivos* e *ocupação circular*.

## 3.1 Conjuntos Seletivos

Os conjuntos seletivos foram propostos por Moura (2006) e são obtidos a partir das regiões ordenadas segundo as suas verossimilhanças. Considere um mapa com $m$ regiões $\{r_1, r_2, ..., r_m\}$ em que $r_i$ é a i-ésima região do mapa, sendo i=1,2,...,$m$. Seja $L_i = LLR_{r_i}$, com i=1,2,...,$m$, o logaritmo da verossimilhança na região $i$. Em seguida, ordena-se os $L_i$ das $m$ regiões do mapa, de modo que, $L_{(1)} \geq L_{(2)} \geq ... \geq L_{(m)}$ e defina o subconjunto $R_j = \{r_{(1)}, r_{(2)}, ..., r_{(j)}\}$, tal que, a região $r_{(i)}$ corresponde a região que tem o valor do logaritmo da razão de verossimilhança, $L_{(i)}$, de posto $i$, e $j$ é tal que j=1,2,...,$m$. É importante notar que as regiões que constituem cada conjunto $R_j$ não são necessariamente conexas.

As Figuras 3.1 apresenta uma ilustração sobre o conceito de conjuntos seletivos. A

Figura 3.1(a) representa o mapa de Minas Gerais em que as 0,4%(ou seja, 0,004*853=3) regiões com maiores LLR individuais (referentes aos óbitos por homicídio, 1998 a 2002) são destacadas em cinza escuro. Da mesma forma, a Figura 3.1(b) apresenta 0,8% (7) regiões; a Figura 3.1(c), 1,6%(14); a Figura 3.1(d), 3,2%(27), a Figura 3.1(e), 50%(427) e a Figura 3.1(f) que contém as 100% (853) regiões do mapa.




(a) (b) (c) (d) (e) (f)

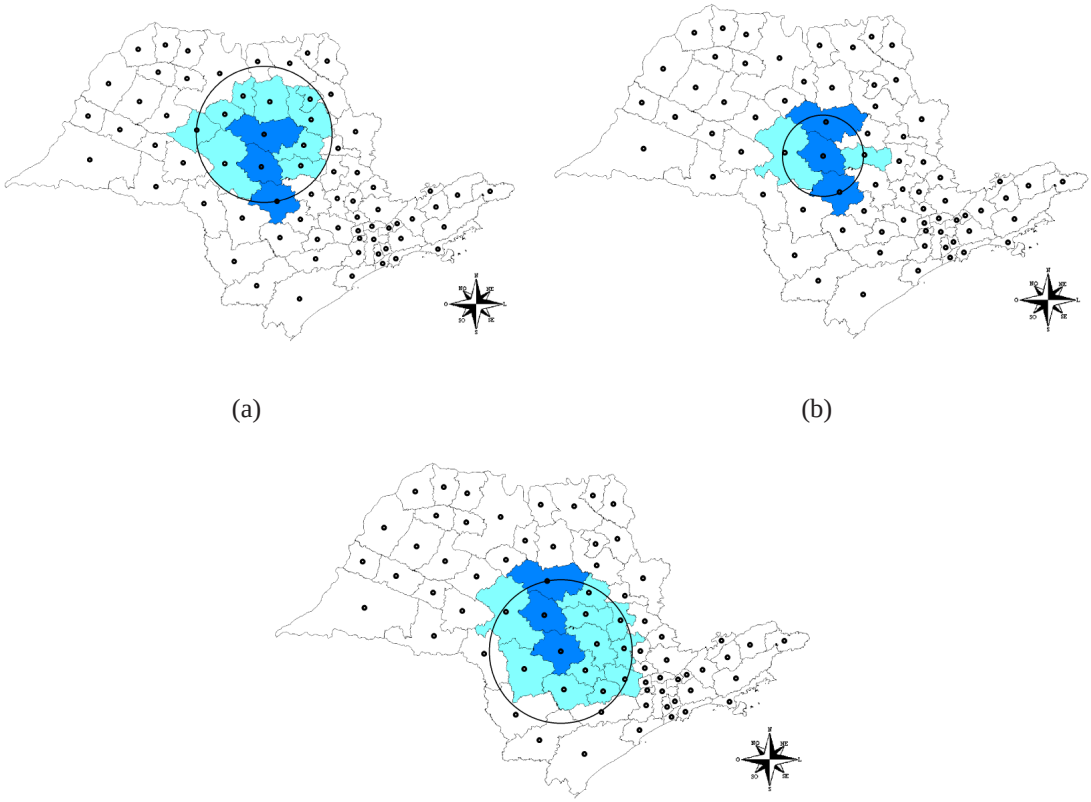
Figura 3.1: Ilustração dos conjuntos seletivos

## 3.2 Ocupação Circular

Moura (2006) propôs, o conceito de ocupação circular (OC) de uma zona z 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 contido no círculo C. Seja P($z$) a população da zona $z$ e seja P(C) a população formada por todas as regiões do mapa original cujos centróides estão contidos no círculo C. Logo a OC da zona $z$ é dada por:

$$OC(z) = \frac{P(z)}{P(C)} \tag{3.1}$$

Um pequeno problema que aparece com a definição acima é que, dado uma zona formada pelas regiões de S, a OC($z$) não seria única, pois com o centro da janela circular em cada centróide da zona teríamos um valor de OC($z$) e a cada diferente centróide o valor do denominador mudaria conforme a Figura 3.2.



(a) (b)

(c)

Figura 3.2: Diferentes valores da OC para a mesma zona formadas pelas regiões em cinza escuro.

Moura(2006) propõe que a OC($z$) seja definida como o máximo de todos os quocientes possíveis, ou seja,

$$OC(z) = max_{r_j \in z} \left\{ \frac{P(z)}{P(C_{r_j})} \right\} \tag{3.2}$$

em que $C_{r_j}$ é o menor círculo centrado na região $r_j$ que contém a zona $z$.

A Ocupação Circular entra na expressão da estatística de teste como um fator de penalização, ou seja, como o expoente da razão de verossimilhança, $LR(z)^{OC(z)}$, ou como um fator multiplicativo se considerarmos o logaritmo da razão de verossimilhança, $OC(z) \times LLR(z)$.

A OC($z$) varia entre 0 e 1, sendo que valores próximos de 1 representam clusters aproximadamente circulares e valores próximos de 0 representam clusters irregulares.

Como os conjuntos seletivos podem ser formados por regiões não conexas que podem estar espalhadas pelo mapa original porém com elevado valor de verossimilhança o baixo valor da Ocupação Circular destes conjuntos seletivos impõe limites ao aumento quase sem restrições da verossimilhança.

## 3.3 Algoritmo da Estatística Scan Seletivo

Baseado na versão multi-objetivo do Scan Seletivo apresentado por Moura (2006) apresentamos, a seguir, a versão mono-objetivo penalizada pela função Ocupação Circular do algoritmo Scan Seletivo.

1. Obtenha as verossimilhanças de cada região do mapa e em seguida ordene-as em ordem decrescente;

2. Construa a matriz de distâncias euclidianas entre os centróides das regiões ordenadas no passo 1;

3. Construa o vetor de população acumulada das j-ésimas regiões vizinhas para cada uma das regiões ordenadas;

4. Fixe um valor de $a$, $0 < a \leq 1$, de tal forma que esse parâmetro determine as $100a\%$ das regiões com maiores verossimilhanças. O número de regiões selecionadas é determinado como o maior número inteiro menor ou igual a $A = a.m$;

   (a) Após selecionar as $A$ regiões, construa a matriz de distâncias entre as regiões selecionadas, definida no passo 2.

   (b) Calcule a população acumulada do conjunto seletivo expresso por $PC_i = \sum_{i=1}^{A} Pop[r_i]$, onde $Pop[r_i]$ são as populações das $A$ regiões selecionadas;

   (c) A partir das regiões selecionadas no passo anterior, aplicaremos o algoritmo Scan Circular de Kulldorff. Considerando as regiões selecionadas fixaremos o centro do círculo no centróide de uma das regiões selecionadas. Fixado o centróide, construíremos a janela circular com o vizinho mais próximo entre as regiões selecionadas. O cálculo da LR nesse passo será feito da mesma forma como descrito no algoritmo Scan Circular, ou seja, o risco relativo da zona, formada por todas as regiões do mapa cujos centróides estão contidos no círculo, elevado ao número de casos da zona, vezes o risco relativo fora do círculo elevado ao número de casos fora do círculo. Para cada zona formada calcula-se a Ocupação Circular.

   (d) Em seguida, faz-se o produto da LLR com a Ocupação Circular e encontra-se para o maior produto, o cluster mais verossímil.

   (e) Identificam-se as regiões que formam o cluster mais verossímil.

5. Para determinar a significância estatística do cluster detectado, compare o valor da estatística encontrada para o cluster detectado com a distribuição da estatística sob $H_0$ obtida via simulação de Monte Carlo.

6. Se o algoritmo for aplicado em um mapa com cluster artificial avaliamos a qualidade do processo de detecção de cluster calculando o Poder de detecção, o PPV e a

Sensibilidade definidos no Capítulo 4.

# Capítulo 4

# Resultados

Nesse capítulo fazemos a avaliação numérica do algoritmo Scan Seletivo através do cálculo do poder do teste, da Sensibilidade e o PPV. Também apresentamos uma aplicação do Scan Seletivo para casos reais de homicídios no estado de Minas Gerais.

## 4.1 Avaliação Numérica

### 4.1.1 Poder

O poder de detecção do algoritmo é utilizado para avaliar a probabilidade do algoritmo detectar um cluster quando de fato ele existe.

O valor do poder do teste é estimado pelo seguinte procedimento:

1. Condicionado no número total de casos distribuímos os casos pelo mapa de acordo com $H_0$ e calculamos o valor de estatística de teste. Esse procedimento é repetido milhares de vezes.

2. Estes milhares de valores são ordenados em ordem crescente e fazemos $T_{crit}$ igual ao percentil 95 deste conjunto de dados, considerando um nível de significância $\alpha = 0,05$.

3. Para simular um cluster artificial distribuímos o número total de casos (fixos) pelas *m* regiões do mapa, onde as regiões que pertencem ao cluster tem um risco relativo elevado com valor obtido como é discutido na Seção 4.1.3 e as demais regiões do mapa tem risco relativo igual a um.

4. Para cada distribuição dos casos totais no mapa sob $H_1$, calcula-se o valor da estatística do teste, $T$.

5. Para cada $T$ calculado no item anterior, ele é comparado com o valor do $T_{crit}$. Se $T > T_{crit}$ consideramos que um cluster foi detectado.

6. Logo, o valor estimado do poder do teste é definido como sendo a razão entre o número de vezes que o algoritmo detecta o cluster, isto é, a quantidade de vezes que $T > T_{crit}$, e o número total de simulações.

O poder do teste é interpretado como sendo a proporção de vezes que o algoritmo detecta o cluster. O algoritmo Scan Seletivo, como descrito na secção 3.3, fornecerá para cada valor do parâmetro *a* um poder correspondente.

A seguir descreveremos à relação do cluster real com o cluster detectado.

### 4.1.2 Sensibilidade e PPV

A seguir serão descritas duas medidas bastante utilizadas para determinar a qualidade de um algoritmo de detecção de clusters, a *sensibilidade* e *PPV (valor preditivo positivo)*.

Define-se, *cluster detectado* como o cluster encontrado pelo algoritmo utilizado e *cluster real* como cluster artificialmente produzido no mapa de acordo com a hipótese alternativa, $H_1$.

Huang et al. (2007) adaptou os conceitos de sensibilidade e PPV para a estatística scan espacial. O cálculo dessas medidas são dadas por:

$$Sens = P(D|R) = \frac{\text{População}(\text{Cluster Detectado} \cap \text{Cluster Real})}{\text{População do Cluster Real}}$$

e

$$PPV = P(R|D) = \frac{\text{População}(\text{Cluster Detectado} \cap \text{Cluster Real})}{\text{População do Cluster Detectado}}$$

em que D representa o indivíduo escolhido aleatoriamente da população do mapa pertencente ao cluster detectado e R é o indivíduo escolhido aleatoriamente da população do mapa e que pertence ao cluster real.

O PPV representa a proporção de regiões do cluster detectado que pertencem ao cluster real, enquanto, a sensibilidade representa a proporção de regiões do cluster real que pertencem ao cluster detectado. Para os métodos de detecção de cluster, valores altos de PPV evidenciam ou que o cluster detectado se aproxima muito do cluster real ou subestima ele. No caso da sensibilidade, valores altos indicam que o cluster detectado ou se aproxima bastante do cluster real ou superestima ele.

### 4.1.3 Simulação de clusters artificiais

Conforme Kulldorff et al. (2003), apresentamos brevemente como o risco relativo de um cluster é calculado.

Seja $n_z$ a população em risco do cluster, e $N$ a população total do mapa. Dado o número total de casos $C$, o número de casos observados $c_z$, no cluster sob a hipótese nula $(H_0)$ de não existir cluster espacial no mapa, tem distribuição Binomial com parâmetros $(C, \tau_z)$ com $\tau_z = \frac{n_z}{N}$. A média e a variância desta distribuição são dadas, respectivamente, por:

$$m_0 = \frac{n_z C}{N} \quad e \quad v_0 = \frac{n_z C(N - n_z)}{N^2}$$

Usando a aproximação normal para a distribuição binomial, o número crítico de casos $k$ para que o teste unilateral rejeite a hipótese nula com o nível de significância $0 < \alpha < 1$

é tal que:

$$\Phi\left(\frac{k-m_0}{\sqrt{v_0}}\right) = 1-\alpha \quad \Longrightarrow \quad \frac{(k-m_0)}{\sqrt{v_0}} = \Phi^{-1}(1-\alpha)$$

em que $\Phi(.)$ é a função de distribuição acumulada da Normal padrão. Se $\alpha = 0,05$ e $\theta = 1-\alpha$ temos que $\Phi^{-1}(\theta) = 1,645$, daí o valor crítico $k$ é tal que $\frac{(k-m_0)}{\sqrt{v_0}} = 1,645$. Sob a hipótese alternativa, com o risco relativo $\rho_z$ para a região do cluster, o número de casos nesta região tem distribuição Binomial com média $m_a = \frac{n_z C \rho_z}{(N-n_z+n_z\rho_z)}$ e variância $v_a = \frac{n_z C \rho_z (N-n_z)}{(N-n_z+n_z\rho_z)^2}$. Observe, neste caso, que $\tau_z = \frac{n_z\rho_z}{(N-n_z+n_z\rho_z)}$. Usando novamente a aproximação Normal, selecionamos o risco relativo $\rho_z$ tal que $\frac{(k-m_a)}{\sqrt{v_a}} = \Phi^{-1}(\theta)$. Desta forma o risco relativo é escolhido de modo que o poder atingido por qualquer teste para cluster espacial tem um limite superior igual a $\theta$. Neste trabalho foi escolhido o valor de $\theta$ igual a $0,999$.

### 4.1.4 Câncer de Mama no Nordeste do Estados Unidos

Para avaliar o poder, sensibilidade e ppv do algoritmo Scan Seletivo, utilizamos o conjunto de dados de Câncer de Mama no Nordeste dos Estados Unidos. Este banco de dados (**??**) consiste de uma população de risco de 29.535.210 mulheres distribuídas por 250 condados em 10 estados do Nordeste dos Estados Unidos mais o distrito de Columbia. Os cluster artificais foram simulados distribuindo-se 600 casos de câncer de mama de acordo com o modelo de Poisson onde um cluster está presente. Escolhidas as regiões para formarem o cluster, atribuímos um risco relativo igual a 1 para todas as regiões fora do cluster e um risco relativo elevado e igual para todas as regiões pertencentes ao cluster. Conforme discutito na Secção **?** o valor do risco relativo elevado das regiões do cluster é tal que, se a localização é conhecida antecipadamente, o poder desse cluster ser detectado é 0.999.

As figuras 4.1, 4.2 e 4.3 mostram os clusters simulados que utilizamos para fazer o estudo de poder, sensibilidade e ppv.

A Tabela 4.1 apresenta as comparações do poder do Scan Seletivo (SS), do algoritmo genético sem penalização (NP), o genético com penalização geométrica (GC), o genético com penalização por não conectividade (NC) e o algoritmo dos nós desconectantes (DN). Os resultados dos algoritmos NP, GC, NC e DN aqui reproduzidos foram obtidos em Cançado et al. (2010).

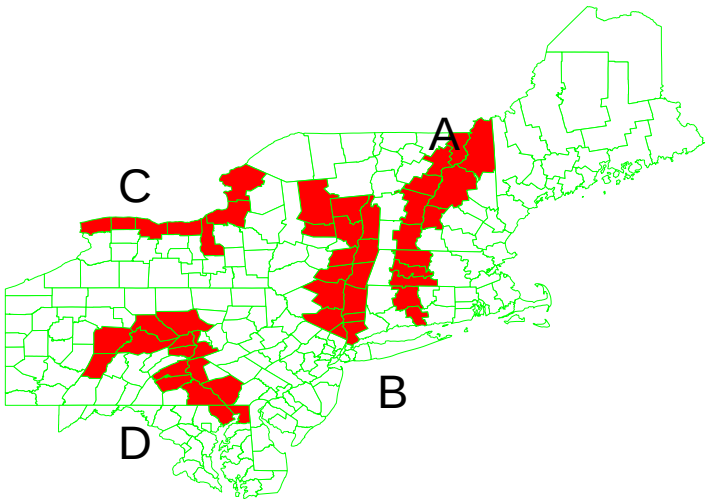



Figura 4.1: Clusters artificiais A, B, C e D para casos de câncer de mama no Nordeste dos Estados Unidos.

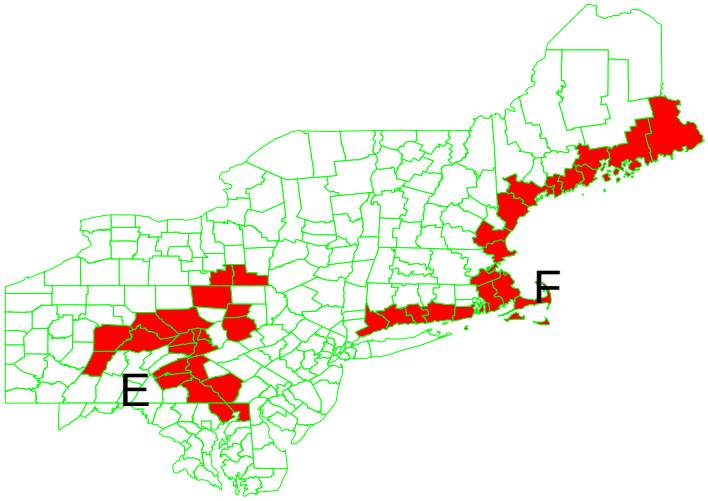



Figura 4.2: Clusters artificiais E e F para casos de câncer de mama no Nordeste dos Estados Unidos.

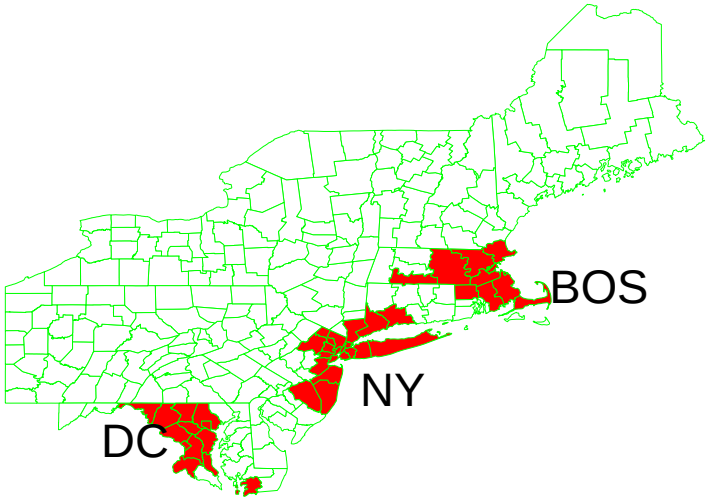



Figura 4.3: Clusters artificiais BOS, NY, DC para casos de câncer de mama no Nordeste dos Estados Unidos.

Tabela 4.1: Comparação do poder entre os algoritmos mono-objetivo.

| cluster | NP | GC | NC | DN | SS |
|---|---|---|---|---|---|
| A | 0.838 | 0.822 | 0.881 | 0.839 | 0.754 (a=0.80) |
| B | 0.882 | 0.843 | 0.926 | 0.898 | 0.730 (a=1.00) |
| C | 0.827 | 0.814 | 0.826 | 0.667 | 0.717 (a=1.00) |
| D | 0.896 | 0.840 | 0.922 | 0.877 | 0.794 (a=0.70) |
| E | 0.874 | 0.778 | 0.885 | 0.822 | 0.741 (a=0.70) |
| F | 0.629 | 0.433 | 0.585 | 0.510 | 0.662 (a=0.80) |
| NY | 0.759 | 0.747 | 0.819 | 0.868 | 0.828 (a=1.00) |
| BOS | 0.792 | 0.834 | 0.864 | 0.892 | 0.903 (a=0.80) |
| D.C. | 0.803 | 0.903 | 0.877 | 0.901 | 0.904 (a=0.85) |

Analisando a Tabela 4.1, notamos que o poder do algoritmo Scan Seletivo (SS) é bem próximo ou em alguns casos superior aos demais algoritmos.

O computador utilizado para as execuções do algoritmo Scan Seletivo com um processador Intel(R) Pentium Dual-Core 2 GB de memória RAM levou um tempo médio de 4 minutos para realizar as 1000 execuções do algoritmo Scan Seletivo.

Tabela 4.2: Comparação do valor preditivo positivo (PPV) entre os algoritmos mono-objetivo.

| cluster | NP | GC | NC | DN | SS |
|---|---|---|---|---|---|
| A | 0.624 | 0.578 | 0.665 | 0.619 | 0.730 (a=0.80) |
| B | 0.699 | 0.691 | 0.786 | 0.765 | 0.623 (a=1.00) |
| C | 0.625 | 0.344 | 0.659 | 0.582 | 0.683 (a=1.00) |
| D | 0.696 | 0.616 | 0.771 | 0.734 | 0.606 (a=0.70) |
| E | 0.719 | 0.633 | 0.762 | 0.704 | 0.586 (a=0.70) |
| F | 0.664 | 0.314 | 0.650 | 0.565 | 0.692 (a=0.80) |
| NY | 0.898 | 0.621 | 0.929 | 0.941 | 0.923 (a=1.00) |
| BOS | 0.781 | 0.389 | 0.827 | 0.861 | 0.827 (a=0.80) |
| D.C. | 0.788 | 0.518 | 0.865 | 0.887 | 0.870 (a=0.85) |

Tabela 4.3: Comparação da sensibilidade dos algoritmos mono-objetivo.

| cluster | NP | GC | NC | DN | SS |
|---|---|---|---|---|---|
| A | 0.796 | 0.551 | 0.792 | 0.767 | 0.658 (a=0.80) |
| B | 0.707 | 0.598 | 0.784 | 0.743 | 0.578 (a=1.00) |
| C | 0.851 | 0.360 | 0.796 | 0.607 | 0.682 (a=1.00) |
| D | 0.668 | 0.506 | 0.713 | 0.668 | 0.594 (a=0.70) |
| E | 0.534 | 0.414 | 0.544 | 0.508 | 0.526 (a=0.70) |
| F | 0.583 | 0.170 | 0.523 | 0.430 | 0.485 (a=0.80) |
| NY | 0.580 | 0.364 | 0.650 | 0.643 | 0.767 (a=1.00) |
| BOS | 0.747 | 0.295 | 0.806 | 0.841 | 0.812 (a=0.80) |
| D.C. | 0.725 | 0.426 | 0.791 | 0.802 | 0.864 (a=0.85) |

Quanto ao PPV (valor preditivo positivo), a Tabela 4.2 mostra que o Scan Seletivo teve um PPV maior quando comparado com o NP, GC, NC e DN para o clusters A, C e F. Para os clusters B e E o PPV identificado pelo Scan Seletivo foi menor do que os outros algoritmos e notamos que o cluster D apresenta menor PPV do que os demais e mais próximo do valor do algoritmo GC. O cluster de BOS teve um PPV maior quando comparado com os algoritmos NP, GC; igual ao do NC e menor do que o DN. O cluster de D.C. teve um PPV inferior apenas comparado com o algoritmo DN.

A Tabela 4.3 mostra a comparação da sensibilidade dos cinco algoritmos mono-objetivo, sendo que a sensibilidade identificada pelo Scan Seletivo é superior quando comparado com o algoritmo GC, exceto para o cluster B. A maior sensibilidade na detecção do cluster DC e NY foi obtida pelo Scan Seletivo.

## 4.2 Aplicações

### 4.2.1 Homicídios em Minas Gerais

Nesta seção, para um total de 11751 casos de homicídios distribuídos por 853 regiões municipais do estado de Minas Gerais ocorridos entre 1998 a 2002, analisaremos os clusters detectados pelo SaTScan e o Scan Seletivo.

A Figura 4.4 mostra um cluster artificialmente simulado contendo um conjunto não-conexo de regiões separado em 4 subconjuntos conexos de regiões. Partindo no sentido horário do conjunto conexo de regiões superior que contém os municípios de Montes Claros e Diamantina, o segundo conjunto conexo contém o município de Governador Valadares, o terceiro conjunto contém o município de Itabira e o quarto conjunto contém os municípios de Belo Horizonte, Betim e Contagem. Estes clusters artificiais foram criados atribuindo um risco relativo igual a 2.00 às regiões do cluster e igual a 1.00 às demais regiões do mapa e distribuindo os casos de acordo com uma distribuição multinomial.

A Figura 4.5 mostra as regiões que constituem o cluster primário (a) e secundário (b), enquanto a Figura 4.6 apresenta o cluster terciário, ambos identificados pelo SaTScan, sendo esses clusters formados por até 15% da população total do mapa.

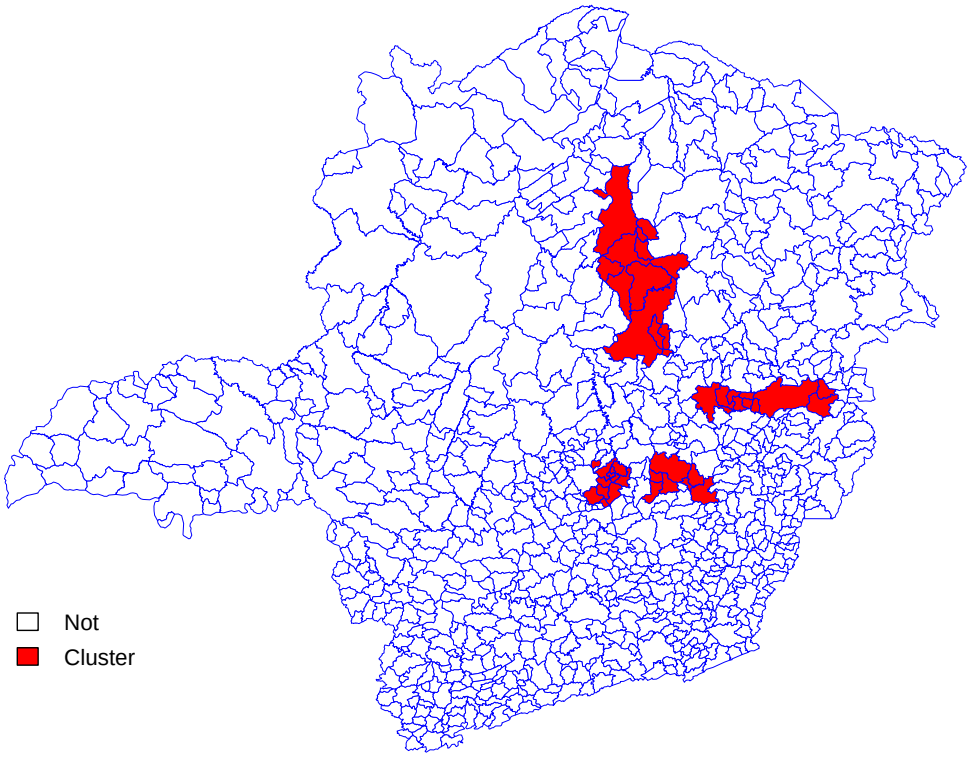
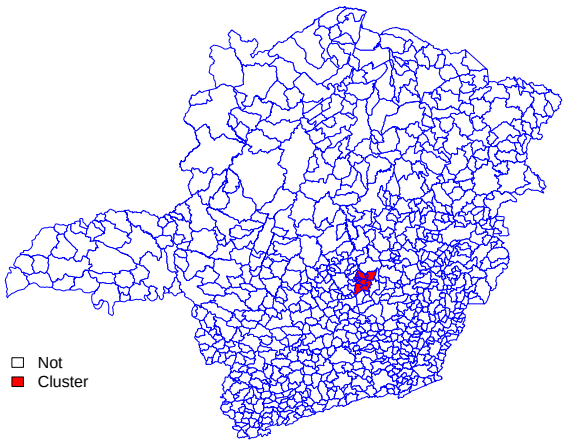
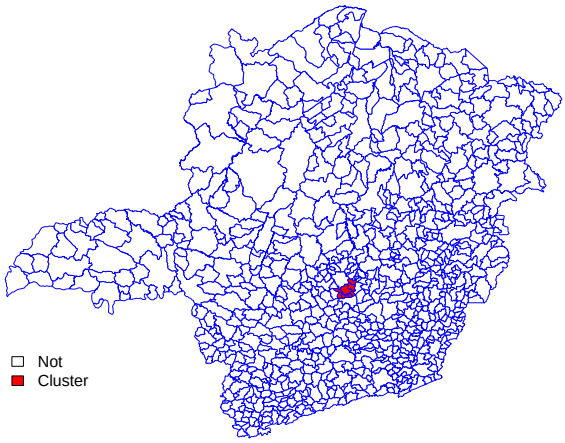



Figura 4.4: Regiões com maior risco.

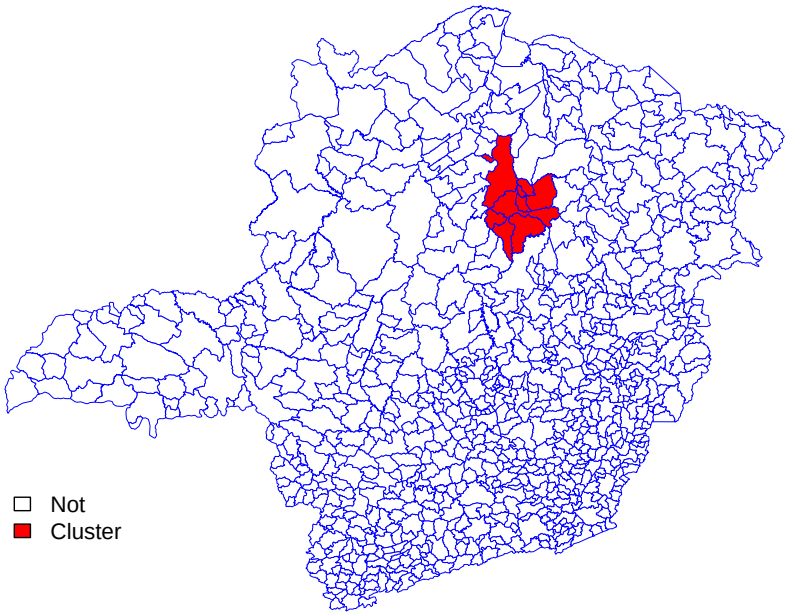



(a)




(b)

Figura 4.5: Regiões que formam os clusters primário (a) e secundário (b) identificado pelo SaTScan.




(a)

Figura 4.6: Regiões que formam o cluster terciário identificado pelo SaTScan.

A Figura 4.7 mostra o cluster identificados pelo Scan Seletivo para o parâmetro de 5% das regiões de maiores LLR e formado por janelas circulares cujos os raios englobem um máximo de 15% da população total do mapa. Podemos notar que as regiões que pertencem ao cluster detectado pelo Scan Seletivo aparecem nos clusters primário e secundário detectado pelo SaTScan.

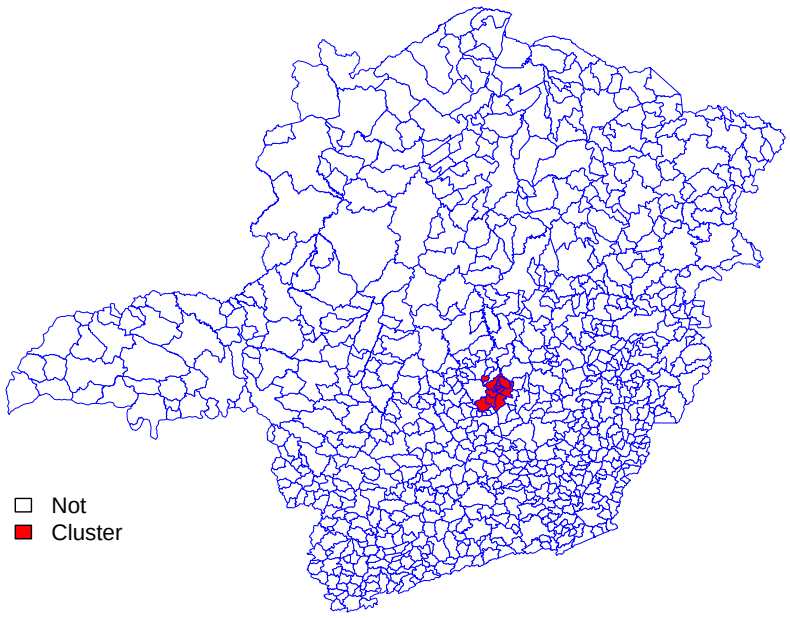



Figura 4.7: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 5% das regiões.

As Figuras 4.8 e 4.9 mostram o cluster identificado pelo Scan Seletivo para o parâmetro de 30% e 100%, respectivamente, das regiões de maiores LLR e formado por 15% da população total do mapa. Novamente notamos que regiões pertencentes as partes conexas do cluster detectado pelo Scan Seletivo correspondem às regiões dos clusters primário e secundário detectados pelo SaTScan. Percebemos que as regiões identificadas pelo Scan Seletivo são desconexas e formado por regiões do cluster primário e secundário do SaTScan.

Os clusters identificados pelo Scan Seletivo para os parâmetros de 45%, 55%, 80%, 90% e 95% encontram-se no Apêndice A, todos os clusters são formados por no máximo 15% da população total do mapa e com o risco relativo de 2,0.

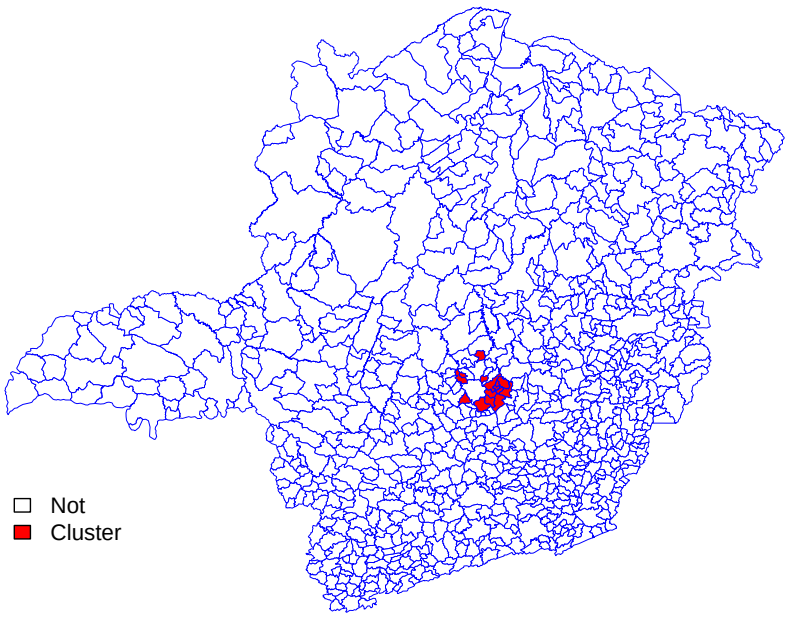



Figura 4.8: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 30% das regiões.

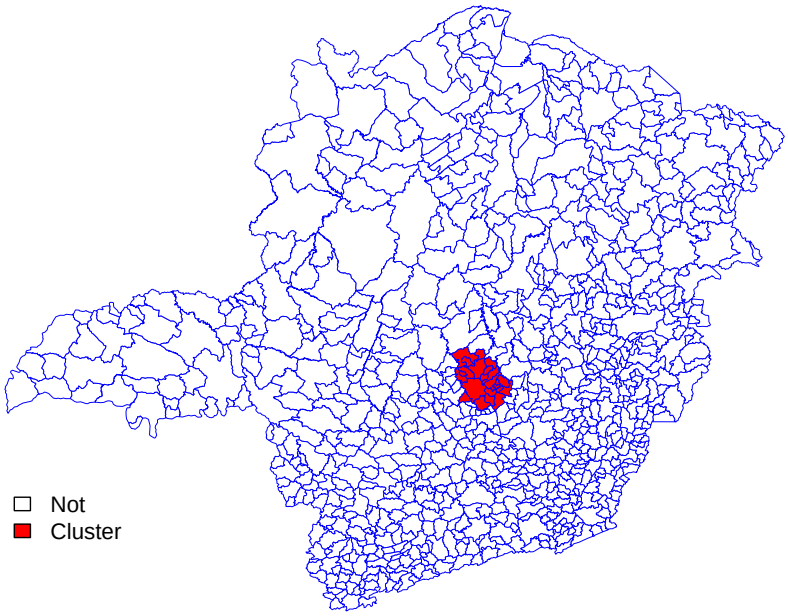



Figura 4.9: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 100% das regiões.

### 4.2.2 Avaliação da significância estatística

A Tabela 4.4 mostra os clusters detectados para cada conjunto seletivo com até 30% da população total do mapa.

A identificação das regiões numeradas, em ordem alfabética, com os municípios de Minas Gerais pode ser vista em http://www.ibge.gov.br/cidadesat/link.php?uf=mg. Por exemplo, as regiões numeradas: 66, 72 e 205 correspondem aos municípios de Belo Horizonte, Betim e Contagem, respectivamente.

Tabela 4.4: *Resultado do algoritmo Scan Seletivo para cada Conjunto Seletivo (CS)*

| CS | LLR | OC | $LLR \times OC$ | centro | raio | regiões |
|---|---|---|---|---|---|---|
| 0,05 | 2848,621 | 0,9692 | 2760,7822 | 637 | 7 | 637-205-843-66-72-675-338 |
| 0,10 | 2848,621 | 0,9692 | 2760,7822 | 637 | 7 | 637-205-843-66-72-675-338 |
| 0,15 | 2814,149 | 0,9832 | 2766,8088 | 637 | 8 | 637-205-843-576-66-72-675-338 |
| 0,20 | 2814,149 | 0,9832 | 2766,8088 | 637 | 8 | 637-205-843-576-66-72-675-338 |
| 0,25 | 2814,149 | 0,9832 | 2766,8088 | 637 | 8 | 637-205-843-576-66-72-675-338 |
| 0,30 | 2814,149 | 0,9832 | 2766,8088 | 637 | 8 | 637-205-843-576-66-72-675-338 |
| 0,35 | 2814,149 | 0,9832 | 2766,8088 | 637 | 8 | 637-205-843-576-66-72-675-338 |
| 0,40 | 2814,149 | 0,9832 | 2766,8088 | 637 | 8 | 637-205-843-576-66-72-675-338 |
| 0,45 | 2814,149 | 0,9832 | 2766,8088 | 637 | 8 | 637-205-843-576-66-72-675-338 |
| 0,50 | 2802,781 | 0,9871 | 2766,5076 | 637 | 9 | 637-740-205-843-576-66-72-675-338 |
| 0,55 | 2802,781 | 0,9871 | 2766,5076 | 637 | 9 | 637-740-205-843-576-66-72-675-338 |
| 0,60 | 2791,250 | 0,9904 | 2764,4158 | 637 | 11 | 637-740-205-843-576-197-66-72-136-675-338 |
| 0,65 | 2791,250 | 0,9904 | 2764,4158 | 637 | 11 | 637-740-205-843-576-197-66-72-136-675-338 |
| 0,70 | 2791,250 | 0,9904 | 2764,4158 | 637 | 11 | 637-740-205-843-576-197-66-72-136-675-338 |
| 0,75 | 2791,250 | 0,9904 | 2764,4158 | 637 | 11 | 637-740-205-843-576-197-66-72-136-675-338 |
| 0,80 | 2791,250 | 0,9904 | 2764,4158 | 637 | 11 | 637-740-205-843-576-197-66-72-136-675-338 |
| 0,85 | 2762,932 | 1,0000 | 2762,9319 | 637 | 12 | 637-740-205-843-576-197-66-72-136-675-431-338 |
| 0,90 | 2762,932 | 1,0000 | 2762,9319 | 637 | 12 | 637-740-205-843-576-197-66-72-136-675-431-338 |
| 0,95 | 2762,932 | 1,0000 | 2762,9319 | 637 | 12 | 637-740-205-843-576-197-66-72-136-675-431-338 |
| 1,00 | 2762,932 | 1,0000 | 2762,9319 | 637 | 12 | 637-740-205-843-576-197-66-72-136-675-431-338 |

A adequação do ajuste foi realizada pelo teste de Kolmogorov-Smirnov (Conover, 1971) que avalia se duas amostras vêm de uma mesma distribuição ($H_0$) ou não ($H_1$). Nesse trabalho, o teste de Kolmogorov-Smirnov (KS) verificará se uma GEV com determinados parâmetros se ajusta bem às estatísticas de teste calculadas, sob a hipótese nula de não existir cluster no mapa, em cada um dos conjuntos seletivos adotados.

A Tabela 4.5 mostra os resultados do p-valor para cada conjunto seletivo utilizando o teste de Kolmogorov-Smirnov.

Tabela 4.5: *Ajuste semi-paramétrico para a distribuição GEV.*

| conjunto seletivo | *a* | *b* | *s* | valor p (teste KS) |
|---|---|---|---|---|
| 0.05 | 2.13 | 0.89 | 0.01 | 0.1773 |
| 0.10 | 2.67 | 0.89 | -0.01 | 0.1484 |
| 0.15 | 2.96 | 0.87 | 0.01 | 0.1359 |
| 0.20 | 3.23 | 0.90 | -0.01 | 0.4794 |
| 0.25 | 3.45 | 0.90 | -0.02 | 0.6596 |
| 0.30 | 3.66 | 0.90 | -0.03 | 0.9901 |
| 0.35 | 3.80 | 0.90 | -0.02 | 0.8435 |
| 0.40 | 3.93 | 0.91 | -0.02 | 0.9854 |
| 0.45 | 4.08 | 0.95 | -0.03 | 0.9596 |
| 0.50 | 4.24 | 0.94 | -0.02 | 0.9846 |
| 0.55 | 4.39 | 0.95 | -0.02 | 0.9691 |
| 0.60 | 4.54 | 0.96 | -0.03 | 0.8763 |
| 0.65 | 4.68 | 0.95 | -0.03 | 0.8332 |
| 0.70 | 4.85 | 0.97 | -0.02 | 0.7541 |
| 0.75 | 5.00 | 0.99 | -0.03 | 0.8932 |
| 0.80 | 5.17 | 1.02 | -0.03 | 0.9173 |
| 0.85 | 5.37 | 1.04 | -0.03 | 0.8609 |
| 0.90 | 5.53 | 1.08 | -0.04 | 0.8754 |
| 0.95 | 5.70 | 1.09 | -0.04 | 0.9743 |
| 1.00 | 5.95 | 1.15 | -0.04 | 0.9470 |

A Tabela 4.5 mostra que para todos os conjuntos seletivos a distribuição GEV se ajuste bem aos dados de máximos referentes a estatística $LLR(z) \times OC(z)$.

A Figura 4.10 ilustra esse ajuste para o conjunto seletivo contendo as 5% e 30% regiões mais verossímeis.

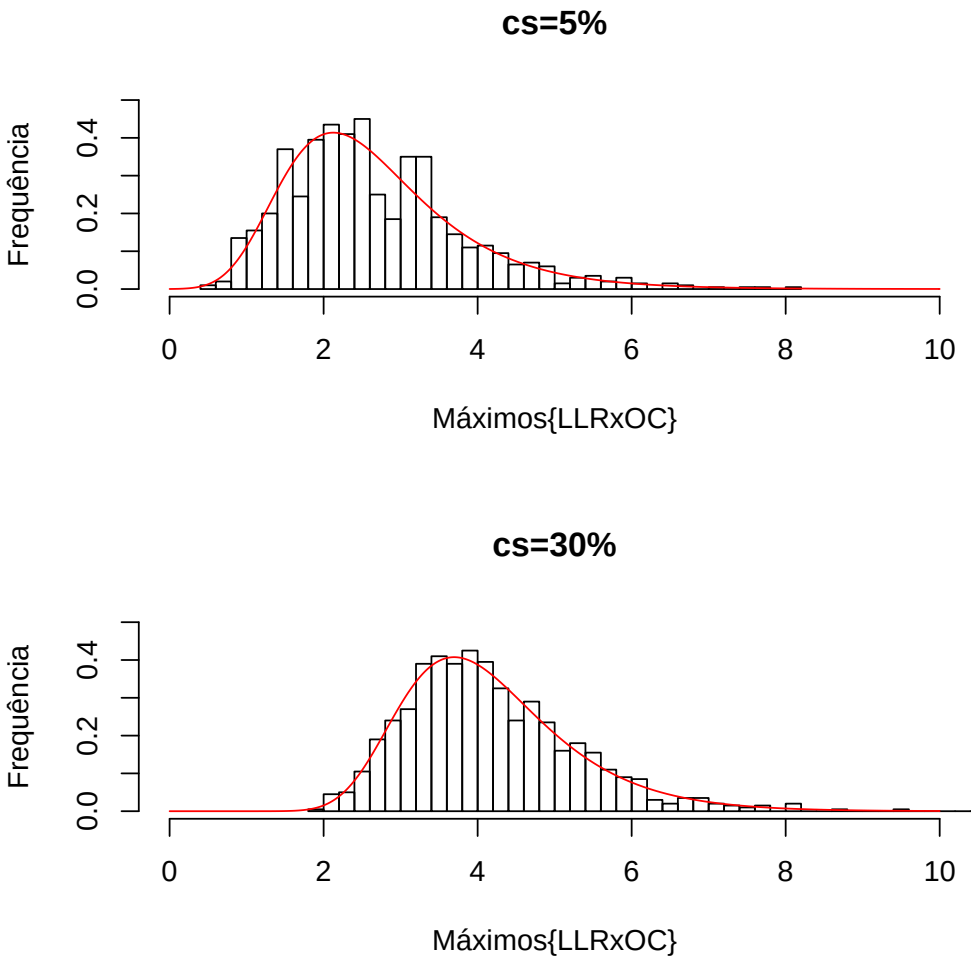



Figura 4.10: Ajuste da GEV para os dados de máximos do conjunto seletivo composto pelas 5% e 30% regiões mais verossímeis.

A Tabela 4.6 mostra cada conjunto seletivo e seus respectivos $LLR(z) \times OC(z)$ e p-valor.

Tabela 4.6: *Estatísticas de teste observadas e p-valores das zonas candidatas a cluster em cada conjunto seletivo, calculados a partir do ajuste da distribuição GEV.*

| conjunto seletivo | $LLR(z) \times OC(z)$ | valor p |
|---|---|---|
| 0,05 | 2760.78 | $< 2.06 \times 10^{-21}$ |
| 0,10 | 2760.78 | $< 2.06 \times 10^{-21}$ |
| 0,15 | 2766.81 | $< 2.06 \times 10^{-21}$ |
| 0,20 | 2766.81 | $< 2.06 \times 10^{-21}$ |
| 0,25 | 2766.81 | $< 2.06 \times 10^{-21}$ |
| 0,30 | 2766.81 | $< 2.06 \times 10^{-21}$ |
| 0,35 | 2766.81 | $< 2.06 \times 10^{-21}$ |
| 0,40 | 2766.81 | $< 2.06 \times 10^{-21}$ |
| 0,45 | 2766.81 | $< 2.06 \times 10^{-21}$ |
| 0,50 | 2766.51 | $< 2.06 \times 10^{-21}$ |
| 0,55 | 2766.51 | $< 2.06 \times 10^{-21}$ |
| 0,60 | 2764.42 | $< 2.06 \times 10^{-21}$ |
| 0,65 | 2764.42 | $< 2.06 \times 10^{-21}$ |
| 0,70 | 2764.42 | $< 2.06 \times 10^{-21}$ |
| 0,75 | 2764.42 | $< 2.06 \times 10^{-21}$ |
| 0,80 | 2764.42 | $< 2.06 \times 10^{-21}$ |
| 0,85 | 2762.93 | $< 2.06 \times 10^{-21}$ |
| 0,90 | 2762.93 | $< 2.06 \times 10^{-21}$ |
| 0,95 | 2762.93 | $< 2.06 \times 10^{-21}$ |
| 1,00 | 2762.93 | $< 2.06 \times 10^{-21}$ |

**Fonte**: Pacote evd.

De acordo com a Tabela 4.6 para todos os conjuntos seletivos os p-valores são muito próximos de zero. Isso evidencia que todos os clusters são significativos, mas não temos estratégia ainda para escolher a melhor solução.

# Capítulo 5

# Conclusões

Nossas comparações de performance com outros algoritmos foram feitas com versões mono-objetivo destes. Sabemos (Cançado et al., 2010) que a perfomance dos algoritmos multi-objetivo são melhores do que suas versões mono-objetivo. Porém o objetivo deste trabalho foi mostrar que as propriedades originais do scan seletivo, ou seja, sua capacidade de detectar clusters formados por um conjunto não necessariamente conexo de regiões e a nova função de penalização, denominada ocupação circular, geram um algoritmo eficiente de deteção de clusters. Neste sentido fizemos a comparação do Scan Seletivo com algoritmos de comprovada eficiência de detecção.

Considerando o quadro geral de resultados do poder de detecção, sensibilidade e PPV obtidos pelo Scan Seletivo em comparação com os demais algoritmos podemos dizer que eles se equivalem. Em algumas situações como o poder para os clusters F, Boston e Washington (DC), o PPV para os clusters A, C e F e a sensibilidade para os clusters NY e DC o Scan Seletivo apresentou o melhor resultado.

Outra qualidade do Scan Seletivo se mostrou quando simulamos clusters com casos de homicídios usando dados populacionais reais do estado de Minas Gerais. Os clusters não conexos detectados pelo Scan Seletivo não somente se aproximam bastante do cluster real bem como suas partes conexas correspondem em boa parte com as regiões dos clusters

primário e secundário detectados pelo SaTScan.

Além disso, nesse trabalho propomos o uso da distribuição generalizada de valores extremos para a estimativa do p-valor da estatística de teste. Diante dos bons resultados obtidos acreditamos que seja uma contribuição relevante para novos trabalhos.

Como continuidade deste trabalho planejamos implementar a versão multi-objetivo incluindo medidas da sensibilidade e do PPV. Além de fazer estudos exaustivos com diversos tipos de clusters simulados que tenham partes desconexas.

# Referências Bibliográficas


Abrams, A., Kulldorff, M., and Kleinman, K. (2006). Empirical/assymptotic p-values for monte carlo-based hypothesis testing: an application to cluster detection using the scan statistic. *Advances in Disease Surveillance*, 1:1–1.

Cançado, A. L. F., Duarte, A. R., Duczmal, L. H., Ferreira, S. J., Fonseca, C. M., and Gontijo, E. C. D. M. (2010). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographics*, 9:55.

Conover, W. J. (1971). *Practical Nonparametric Statistics*. John Wiley & Sons, New York.

Duczmal, L. H., Cancado, A. L. F., and Takahashi, R. H. C. (2008). Delineation of irregularly shaped disease clusters through multi-objective optimization. *Journal of Computational & Graphical Statistics*, 17:243–262.

Duczmal, L. H., Kulldorff, M., and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters. *Journal of Computational & Graphical Statistics*, 15:428–442.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28:181–187.

Huang, L., Kulldorff, M., and Gregorico, D. (2007). A spatial scan statistic for survival data. *Biometrics*, 63:109–118.

Jenkinson, A. F. (1985). The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Quart. J. R. Met. Soc.*, 81:158–171.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496.

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

Lima, M. (2004). Avaliação do poder do teste da estatística scan para múltiplos clusters. Mestrado em estatística, Universidade Federal de Minas Gerais, Belo Horizonte / MG.

Moura, F. d. R. (2006). Detecção de clusters espaciais via algoritmo scan circular seletivo. Mestrado em estatística, Universidade Federal de Minas Gerais, Belo Horizonte / MG.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72:67–90.

Stephenson, A. G. (2002). evd: Extreme value distributions. *R News*, 2(2):0.

# Apêndice A

# Clusters identificados pelo Scan Seletivo

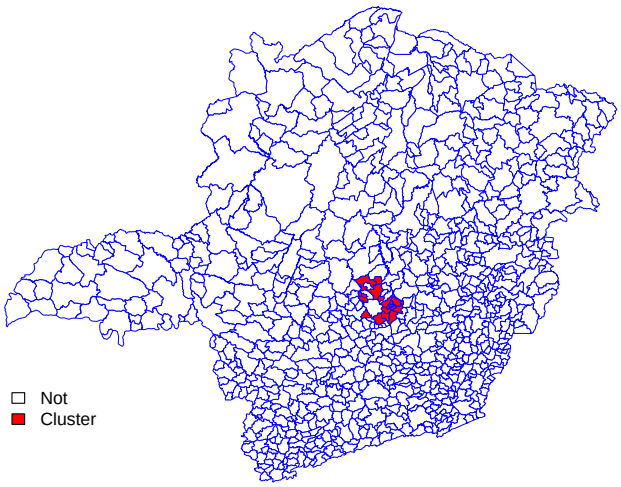



Figura A.1: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 45% das regiões.

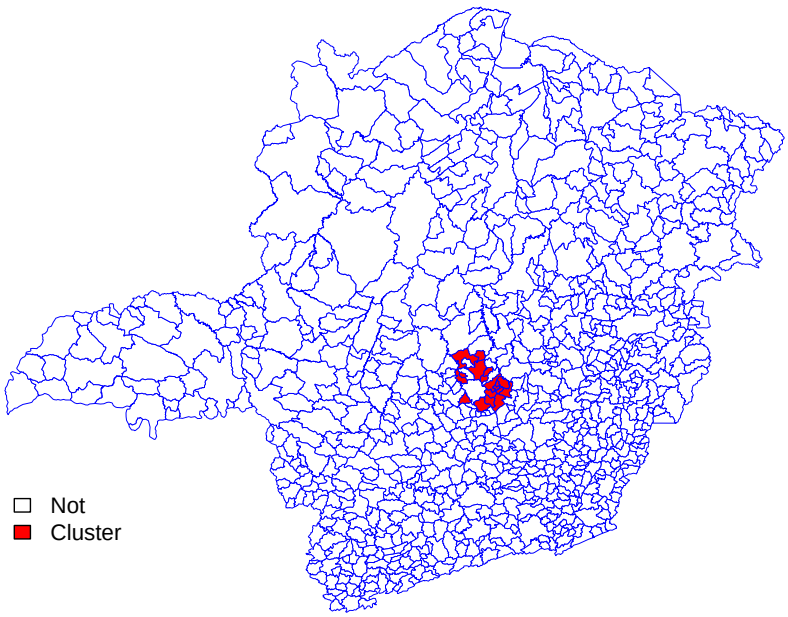



Figura A.2: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 55% das regiões.

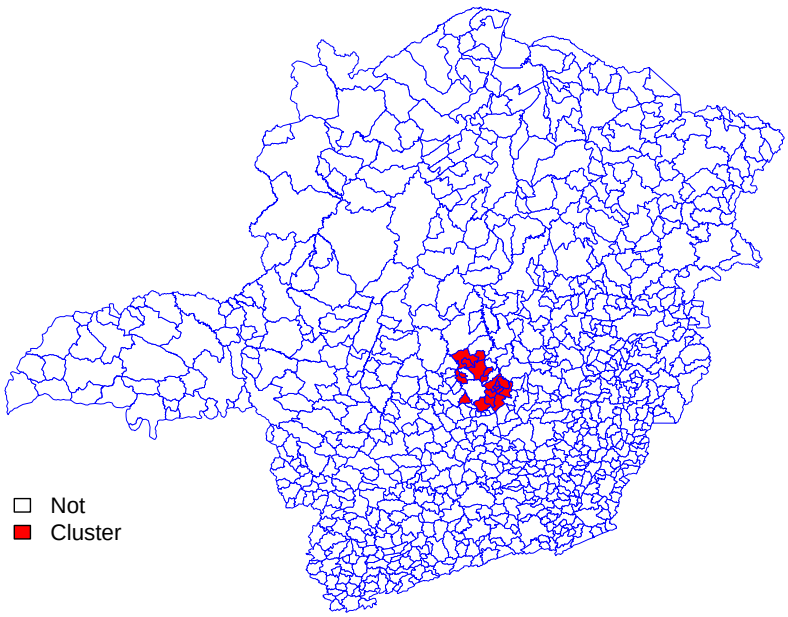



Figura A.3: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 80% das regiões.

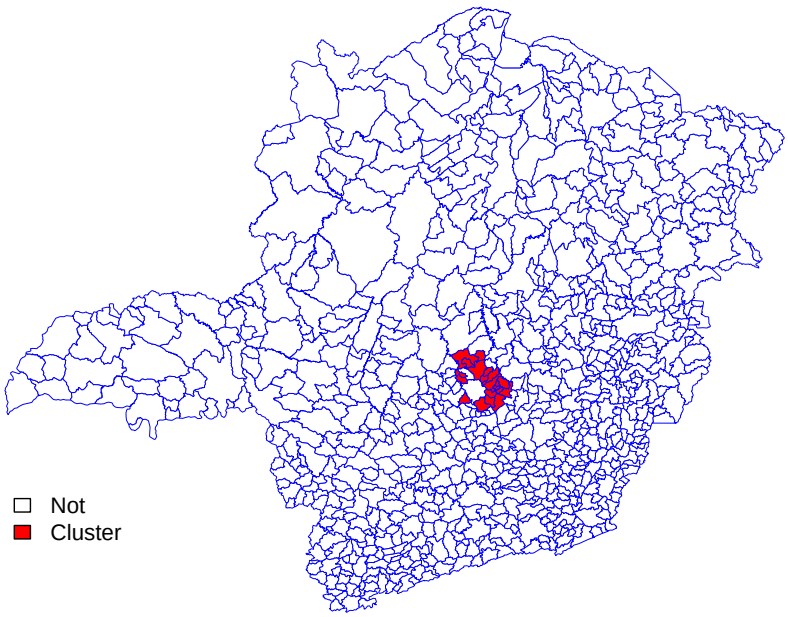



Figura A.4: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 90% das regiões.

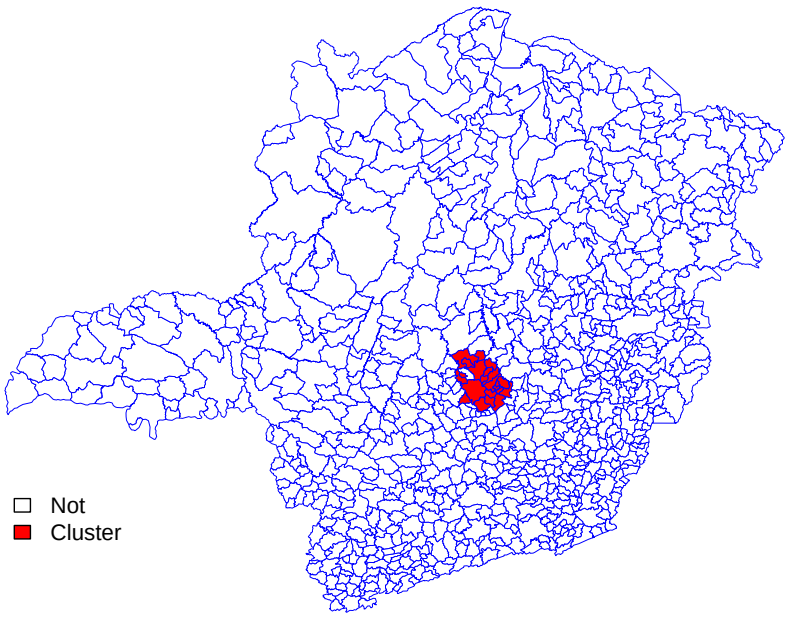



Figura A.5: Cluster identificado pelo Scan Seletivo com o conjunto seletivo de 95% das regiões.