**Emerson Cotta Bodevan**

# DETECÇÃO SIMULTÂNEA DE MÚLTIPLAS REGIÕES DE ALTO E BAIXO RISCO EM MAPAS DE DADOS PONTUAIS DE CASO-CONTROLE

Belo Horizonte/MG, Fevereiro 2012.

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação em Estatística

# DETECÇÃO SIMULTÂNEA DE MÚLTIPLAS REGIÕES DE ALTO E BAIXO RISCO EM MAPAS DE DADOS PONTUAIS DE CASO-CONTROLE

**Emerson Cotta Bodevan**

Tese de doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Doutor em Estatística.

Área de Concentração: Estatística e Probabilidade
Orientador: Luiz Henrique Duczmal

Belo Horizonte/MG, Fevereiro de 2012

# Dedicatória

*Dedico este trabalho aos meus filhos, Bernardo e Letícia, à minha esposa, Luciana, à minha mãe, Maria de Lourdes, e a Deus. Obrigado por me apoiarem, cada um a sua maneira.*

**Amo vocês!**

# Agradecimentos

*Em especial agradeço ao meu orientador Luiz Duczmal e aos amigos Alexandre Celestino, Anderson Duarte, Fernando Oliveira, Gladston Moreira, Marcelo Buosi e Nilson Brito. Agradeço também a todos os outros docentes do Departamento de Estatística que, de uma forma ou de outra, cotribuíram para minha formação. Agradeço a todos os outros familiares e amigos que me apoiaram e acreditaram em mim. Agradeço igualmente às meninas das secretarias, em especial a Rogéria e a Marcinha.*

*Agradeço a CAPES e a FAPEMIG.*

**Obrigado!**

# Resumo

A estatística *scan* espacial é a técnica mais comumente utilizada para detecção de *clusters*. Várias extensões desta técnica foram desenvolvidas, buscando flexibilizar o espaço de busca dos *clusters* assim como melhorar a precisão da sua detecção. Uma das extensões da estatística *scan* mais recentes, denominada *Voronoi Based Scan* (VBScan), se propõe a detectar *clusters* para o caso de dados pontuais do tipo caso-controle utilizando árvores geradoras mínimas. Outros esforços estão sendo empregados no sentido de utilizar a estatística *scan* no problema de detecção de múltiplos *clusters*.

Neste trabalho, propõe-se um método, baseado no VBScan, para detecção da partição de um mapa consistindo de dados pontuais do tipo caso-controle. O método visa identificar e delinear todas as múltiplas anomalias significativas, que podem ser de alto ou baixo risco. Neste novo método utiliza-se o VBScan recursivamente sobre um mapa com dados pontuais do tipo caso-controle, através de um procedimento bi-objetivo.

O método foi testado em diferentes mapas simulados, particionados em diferentes números de componentes. O poder de detecção e o *matching* (uma medida de *overlap* entre as partições verdadeiras e as detectadas) foram avaliados. O método também foi aplicado em dois conjuntos de dados reais.

O método mostrou-se rápido e com boa precisão na determinação das partições.

**Palavras-chave:** Árvore geradora mínima; Estatística *scan* espacial; Multiobjetivo; Voronoi.

# Abstract

The spatial scan statistic is the most commonly used technique for detecting clusters. Several extensions of this technique have been developed, seeking flexibility in the search space of the clusters, as well as improvement in the accuracy of the detection. One of the recents extensions of the scan statistic, called Voronoi Based Scan (VBScan), aims to detect clusters in a map consisting of point event data (case-control). Other efforts are being employed in order to use the scan statistic in the problem of detection of multiple clusters.

In this work, we developed a method based on the VBScan to detect the partition of a map consisting of point-event data (case-control). The method aims to identify all significant multiple anomalies, which can be of high or low risk. In this new method, we use the VBScan recursively on the map with case-control point-event data, através de um procedimento bi-objetivo. Our method was tested on different simulated maps and partitioned into different numbers of components. We evaluate the power of detection and matching (a measure of overlap between the true and detected partitions). We also apply the method on two case studies.

The method is fast, with good accuracy determination.

**Keywords:** Minimum spanning tree; Multi-objective; Spatial scan statistics; Voronoi.

# Lista de Figuras


2.1 Distribuição espacial e árvore geradora mínima de Voronoi para um conjunto hipotético de pontos caso-controle. . . . . 14

2.2 Visualização do procedimento guloso de retirada de arestas, em passos sucessivos enumerados de 1 a 10. . . . . 15

3.1 Conjunto de pontos ilustrando o conceito de dominância. . . . . 24

3.2 Conjunto Pareto-Ótimo (+) e pontos dominados (•). . . . . 25

3.3 (a) AGMV e (b) cluster mais verossímil detectado. . . . . 28

3.4 Partições candidatas compostas pelo cluster mais verossímil e respectivos clusters nas subárvores. . . . . 28

3.5 Demais partições candidatas. . . . . 29

3.6 Soluções factíveis encontradas no exemplo comentado. . . . . 29

3.7 Conjunto Pareto-ótimo para o exemplo comentado. . . . . 30

3.8 Fronteira entre $R_0$ e $R_1$. . . . . 31

3.9 Múltiplas Fronteiras e respectivas Superfícies de Aproveitamento. . . . . 32

3.10 Os 10.000 conjuntos Pareto-ótimo obtidos por simulações de Monte Carlo sob a hipótese nula são representados pelos pontos e são apresentadas as isolinhas correspondentes a 95%; 99% e 99,9%. . . . . 33

3.11 Localização e identificação dos casos de uma partição artificial gerada. . 35

3.12 Partição artificial simulada e detectada. . . . . 35

3.13 Configurações 123 e 132, e respectivo *matching*, entre as partições simulada e detectada. . . . . 36
3.14 Configurações 213 e 231, e respectivo *matching*, entre as partições simulada e detectada. . . . . 36
3.15 Configurações 312 e 321, e respectivo *matching*, entre as partições simulada e detectada. . . . . 37

4.1 Três cluster espaciais artificiais. . . . . 40
4.2 Populações homogênea e heterogênea . . . . . 42
4.3 Mapas artificias gerados. . . . . 43

5.1 Distribuição espacial dos casos observados de câncer de pulmão (pontos) e de câncer de laringe (círculos), na cidade de Lancashire-UK. . . . . 50
5.2 Distribuição espacial dos casos observados de dengue (círculos) e dos controles (pontos), na cidade de Lassance-BR. . . . . 51
5.3 Árvore geradora mínima de Voronoi, com cluster mais verossímil encontrado pelo VBScan e pelo Scan Elíptico, para os dados de Lancashire. . . 52
5.4 Árvore geradora mínima de Voronoi, com cluster mais verossímil encontrado pelo VBScan e pelo Scan Elíptico, para os dados de Lassance. Pontos em cinza correspondem ao cluster primário, enquanto que os pontos pretos, ao cluster secundário encontrado pelo VBScan. Os pontos de cor cinza marcados com uma cruz correspodem ao cluster primário encontrado pelo Scan Elíptico. . . . . 54
5.5 Soluções Pareto-ótimas para o conjunto de Lancashire. . . . . 56
5.6 Superfícies de aproveitamento de 95%, 99% e 99,9% , com respectiva solução Pareto-ótima (+), para os dados de Lancashire. Separadas por número de componentes: 2 (A), 3 (B), 4 (C) e 5 (D). . . . . 57
5.7 Soluções Pareto-ótimo para o conjunto de Lassance. . . . . 59

5.8 Superfícies de aproveitamento de 95%, 99% e 99, 9%, com respectiva solução Pareto-ótima (+), para os dados de Lassance. Separadas por número de componentes: 2 (A), 3 (B), 4 (C) e 5 (D). . . . . 60

# Lista de Tabelas


4.1 Comparação de poder, valor preditivo positivo e sensibilidade para os três formatos de clusters. . . . . 41

4.2 Riscos relativos utilizados em cada uma das componentes de todos os mapas artificias criados. . . . . 44

4.3 Resultados de poder e *matching* para cada um dos mapas artificias criados. . . . . 45

5.1 Comparação da detecção de *clusters* espaciais para os dados de Lancashire, resultados encontrados para os métodos scan elíptico e VBScan. . . . . 53

5.2 Resultados encontrados para detecção de *clusters* espaciais para os dados de Lassance utilizando o método VBScan. . . . 55

5.3 Resultados da detecção da melhor partição para os dados de Lancashire. . . . . 58

5.4 Resultados da detecção da melhor partição para os dados de Lassance. . . . . 61

# Sumário

**Resumo** ix

**Abstract** xi

**Lista de Figuras** xiii

**Lista de Tabelas** xvii

**1 Introdução** 1

1.1 Estatística Scan Espacial . . . . . 1

1.2 Tipos de dados . . . . . 4

1.2.1 Dados agregados (ou dados de área) . . . . . 4

1.2.2 Dados pontuais . . . . . 4

1.2.3 Dados pontuais (caso-controle) . . . . . 5

1.3 Referencial teórico . . . . . 5

1.4 Organizacão da Tese . . . . . 8

**2 Voronoi Based Scan (VBScan)** 9

2.1 Estimação da densidade populacional . . . . . 11

2.2 Conjunto dos possíveis clusters no espaço de coordenadas . . . 12

2.3 Inferência e medidas de eficiência para o VBScan . . . . . 16

2.3.1 Inferência . . . . . 16

2.3.2 Medidas de eficiência . . . . . 16

**3 Multi-Objective Multiple Clusters (MOMC-VBScan) 19**

3.1 Otimização multi-objetivo . . . . . 22

3.2 O processo de busca de partições candidatas . . . . . 25

3.3 Calculando a significância estatística das partições . . . . . 30

3.4 Medidas de eficiência do algoritmo . . . . . 34

3.4.1 Poder . . . . . 34

3.4.2 *Matching* . . . . . 35

**4 Resultados de Simulações 39**

4.1 VBScan . . . . . 39

4.1.1 Clusters artificiais . . . . . 39

4.1.2 Análises numéricas . . . . . 41

4.2 MOMC-VBScan . . . . . 42

4.2.1 Os mapas artificiais . . . . . 42

4.2.2 Análises numéricas . . . . . 45

**5 Aplicações em dados Reais 49**

5.1 VBScan . . . . . 52

5.1.1 Dados de Lancashire . . . . . 52

5.1.2 Dados de Lassance . . . . . 53

5.2 MOMC-VBScan . . . . . 56

5.2.1 Dados de Lancashire . . . . . 56

5.2.2 Dados de Lassance . . . . . 58

**6 Conclusões 63**

6.1 Considerações Finais . . . . . 63

6.2 Propostas de continuidade . . . . . 65

**Referências Bibliográficas** **67**

# Capítulo 1

# Introdução

Considerando a distribuição geográfica da incidência de algum fenômeno de interesse, como casos de doenças ou de homicídios, a verificação de padrões anômalos nesta distribuição é de extrema importância para que se possa planejar políticas de intervenção em saúde ou segurança pública.

Estudos referentes a conglomerados espaciais que apresentam discrepância na ocorrência do fenômeno de interesse são encontrados em diversos trabalhos. Neste texto, conglomerados espaciais serão tratados pela palavra em inglês, *clusters*, como já é bastante usual nesta área de pesquisa.

Para um bom entendimento do referencial teórico sobre o assunto de interesse dessa tese, a próxima seção se propõe a explicar o método da Estatística Scan Espacial, a principal técnica utilizada na detecção de clusters.

## 1.1 Estatística Scan Espacial

Nesta seção faremos uma breve revisão da estatística scan clássica introduzida em Kulldorff (1997).

Considere um mapa dividido em $m$ regiões $R_1, \ldots, R_m$, com população

total $N$ e número total de casos $C$, para o fenômeno de interesse. Assuma ainda que a população e o número de casos em cada uma das regiões sejam também conhecidos e denotados por $n_i$ e $c_i$ com $i \in \{1, \ldots, m\}$, respectivamente. Define-se como *zona* qualquer subconjunto conexo de regiões do mapa em estudo. Seja $Z$ o conjunto de todas as zonas do mapa. Com a suposição de que os casos se distribuem no mapa segundo o modelo de Poisson, temos que o número de casos $C_i$ em cada região $R_i$ é uma variável aleatória de Poisson, cujo parâmetro $\mu_i$ é dado por $C\frac{n_i}{N}$. Note que $\mu_i$ é o número esperado de casos na região $R_i$.

O procedimento proposto por Kulldorff constitui-se na construção de um teste de hipótese, cuja hipótese nula é a de não existência de *cluster* no mapa em estudo, enquanto que a hipótese alternativa pressupõe a existência de pelo menos um *cluster* no mapa em estudo. Trata-se do clássico teste da razão de verossimilhança.

Pode-se mostrar que a razão entre a função de verossimilhança sob a hipótese alternativa e a função de verossimilhança sob a hipótese nula para a distribuição de casos em alguma zona $z$ em estudo, é dada por (veja Kulldorff (1997)):

$$LR(z) = \begin{cases} \left(\frac{c_z}{\mu_z}\right)^{c_z} \left(\frac{C - c_z}{C - \mu_z}\right)^{C - c_z} & \text{se } c_z > \mu_z \\ 1 & \text{caso contrário} \end{cases} \tag{1.1}$$

A zona $z$ mais verossímil é aquela que maximiza a função $LR(z)$ com respeito ao conjunto $Z$. Desta forma, a estatística de teste fica definida por $\max_{z \in Z} LR(z)$.

Em geral, a função $LR(z)$ assume valores muito grandes. Para amenizar esse problema, utiliza-se o logaritmo da razão de verossimilhança, $LLR(z)$. Dado que a função logaritmo é monotonicamente crescente, a zona $z$ que

maximiza $LR(z)$ também maximiza $LLR(z)$. A expressão para $LLR(z)$ é dada por:

$$LLR(z) = \begin{cases} c_z \log\left(\frac{c_z}{\mu_z}\right) + (C - c_z)\log\left(\frac{C - c_z}{C - \mu_z}\right) & \text{se } c_z > \mu_z \\ 0 & \text{caso contrário} \end{cases} \tag{1.2}$$

É importante salientar que identificar a zona mais verossímil não constitui em identificar um cluster. Precisa-se ainda verificar a sua significância estatística para que a zona detectada seja considerada como cluster.

A significância estatística da zona mais verossímil identificada nos dados observados é calculada através de simulação Monte Carlo, de acordo com o procedimento descrito em Dwass (1957). Sob a hipótese nula, casos simulados são distribuídos sobre a região em estudo e a estatística de teste é calculada. Este procedimento é repetido uma grande quantidade de vezes, com o objetivo de produzir uma distribuição empírica para a estatística de teste, sob a hipótese nula. O valor da estatística de teste nos dados observados é então comparado com essa distribuição empírica afim de determinar seu nível de significância (o valor $p$).

A cardinalidade do conjunto $Z$, definido anteriormente, para um mapa dividido em $m$ regiões, pode ser muito grande. Note que existem $2^m - 1$ possíveis subconjuntos de regiões, dos quais deveríamos verificar quais são conexos, para construir o conjunto $Z$. Para $m$ da ordem de algumas centenas, esta tarefa já seria computacionalmente árdua.

Desta forma, alguma técnica para redução do espaço de busca deve ser utilizada. A técnica mais utilizada para este fim é denominada Scan Espacial Circular (Kulldorff & Nagarwalla (1995)). Nesta técnica, a redução do espaço de busca é feita através de janelas circulares centradas em cada um dos

centróides[1] do mapa em estudo. O raio do círculo varia de zero até um raio máximo segundo alguma métrica pré-estabelecida.

O método Scan Espacial como apresentado aqui, está definido para informações agregadas por regiões no mapa em estudo. No entanto, os dados disponíveis sobre o fenômeno de interesse nem sempre se apresentam desta forma, podendo aparecer em outros formatos, como descrito na próxima seção.

## 1.2 Tipos de dados

### 1.2.1 Dados agregados (ou dados de área)

Nesse tipo de dados, uma região de estudo é dividida em subregiões não sobrepostas (tais como bairros, municípios, setores censitários etc.). Cada subregião é geo-referenciada por um ponto arbitrário em seu interior (centróide). Comumente, associam-se a cada um desses centróides, o número de casos e a população referente à subregião que ele representa. Podem-se também agregar, às contagens de casos e populações, variáveis sócio-econômicas, ambientais, físicas, biológicas etc. Esse é o tipo mais comum de dados disponíveis, uma vez que, na maioria das vezes, a localização exata dos indivíduos não está disponível publicamente.

### 1.2.2 Dados pontuais

Nesse tipo de dados, a localização dos objetos em estudo (por exemplo, casos de doença, casos de homicídio etc.) são, individualmente, representados por

[1] Um centróide é um ponto arbitrário dentro de cada região que compõe o mapa em estudo.

coordenadas bi-dimensionais em um mapa no $\mathbb{R}^2$. Aqui, como no caso de dados agregados, pode-se associar a cada objeto outras variáveis de interesse (idade, gênero etc.).

### 1.2.3 Dados pontuais (caso-controle)

É similar à definição anterior, entretanto alguns indivíduos são classificados como casos, ou seja, possuem a característica de interesse do estudo, ao passo que os outros indivíduos são classificados como controles.

## 1.3 Referencial teórico

Os principais métodos para detecção de clusters espaciais foram revistos em Elliott *et al.* (1995), Waller & Jacquez (1995), Lawson & Kulldorff (1999), Moore & Carpenter (1999), Glaz *et al.* (2001), Lawson (2001), Balakrishnan & Koutras (2002), Buckeridge *et al.* (2005), e mais recentemente em Duczmal *et al.* (2009).

Algoritmos para a detecção de clusters são ferramentas úteis em estudos etiológicos, conforme Lawson *et al.* (1999). Para a advertência antecipada de surtos de doenças infecciosas, resultados podem ser encontrados em Duczmal & Buckeridge (2006); Kulldorff *et al.* (2005, 2006, 2007) e Neill (2009).

A estatística scan espacial de (Kulldorff (1997)), utilizada para a detecção de clusters, está implementada no *software* SaTScan (Kulldorff (1999)). A precisão das estimativas de valor $p$ fornecidas pelo Scan Circular é discutida em Abrams *et al.* (2010). Em Almeida *et al.* (2011) é proposta uma correção no cálculo do valor $p$ da estatística scan acrescentando a informação do tamanho dos clusters candidatos.

Várias tentativas têm sido desenvolvidas para relaxar a suposição da forma circular do cluster. A estatística scan ULS proposta em Patil & Taillie (2004) controla a excessiva liberdade na forma geométrica de possíveis candidatos através de uma razão que divide o número de casos pela população de risco na área em estudo. Em Modarres & Patil (2007) discute-se uma extenção da estatística scan ULS para dados bivariados. Em Duczmal & Assunção (2004) é proposto a utilização de um algoritmo *Simulated Annealing*, visando obter a solução que maximiza a estatística espacial scan. A estatística scan FS proposta em Tango & Takahashi (2005) faz uma busca exaustiva de todos os clusters conectados de primeira ordem contidos num conjunto abrangendo os $K$ vizinhos mais próximos de uma dada região. A técnica Scan Circular foi estendida para o caso de clusters com forma elíptica em Kulldorff *et al.* (2006), permitindo assim a detecção de clusters de forma alongada. Em Assunção *et al.* (2006) é proposto um método para detecção de clusters, baseado em Árvores Geradoras Mínimas. Uma medida para avaliar o grau de regularidade de possíveis soluções, denominada compacidade geométrica, é apresentada em Duczmal *et al.* (2006). Neste trabalho, ela é utilizada em associação com o algoritmo *Simulated Annealing*. Em Duczmal *et al.* (2007), esta medida é associada a um algoritmo genético visando maximizar o produto entre a medida de compacidade geométrica e a $LLR(z)$. Em Yiannakoulias *et al.* (2007) é proposta uma nova medida de avaliação de regularidade da forma do cluster, uma penalização topológica. Entretanto, utilizam uma técnica de otimização mais simples, um algoritmo guloso. Em Moura *et al.* (2007) desenvolve-se um método para analisar mais profundamente os vários níveis de agrupamentos que surgem naturalmente em mapas divididos em $m$ regiões. Ao invés de usar um algoritmo genético, este método incorpora a simplicidade do scan circular, sendo hábil também para detectar

cluster irregulares. Em Neill (2008) é proposta uma forma computacionalmente mais eficiente de busca de soluções. Uma abordagem mais recente utiliza medidas para avaliar a regularidade na forma de possíveis soluções através de técnicas multi-objetivo (Duczmal *et al.* (2008); Cançado *et al.* (2010); Duarte *et al.* (2010)).

Em Sahajpal *et al.* (2004) apresenta-se um algoritmo genético para encontrar clusters formados pela interseção de círculos de diferentes tamanhos e centros em dados pontuais. Em Conley *et al.* (2005) é proposto um algoritmo genético para explorar uma configuração espacial de múltiplos aglomerações de elipses em mapas de dados pontuais. Em Wieland *et al.* (2007) é proposto um método baseado em árvore geradora mínima e no diagrama de Voronoi, para dados pontuais caso-controle.

Em Dematteï *et al.* (2007) é proposto um método baseado na construção de uma trajetória para a detecção de múltiplos clusters utilizando a estatística espacial scan em dados pontuais. Nos estudos em que há a coexistência de múltiplos clusters, deparamo-nos com o problema do efeito de sombra que um cluster pode causar sobre o outro. Em Zhang *et al.* (2010) é proposto um procedimento sequencial para detecção de clusters no qual, na eventualidade da existência de múltiplos clusters, o grau de significância de cada um deles é obtido de forma mais precisa. O procedimento remove o efeito do cluster mais significativo, em um passo, sobre os clusters com menor significância, no passo posterior, pela exclusão sequencial dos clusters detectados anteriormente. Esta metodologia apresentou maior poder de detecção do cluster secundário do que o método scan circular sem o procedimento proposto. Li *et al.* (2011) apresentam uma técnica alternativa ao procedimento sequencial anterior, adotando uma estatística de teste única adaptada ao número de clusters presentes na área em estudo. A hipótese alternativa

é ajustada à presença de dois ou mais clusters. A técnica se mostrou com maior poder de detecção do que o método sequencial e com maior precisão no delineamenteo dos múltiplos clusters coexistentes.

Desta revisão, notamos que as técnicas de detecção de clusters podem evoluir para procedimentos mais eficientes e menos onerosos computacionalmente e que há uma carência em trabalhos que apresentam técnicas para procedimentos de detecção de múltiplos clusters em dados pontuais do tipo caso-controle.

Esta tese propõe uma técnica para particionar um mapa em estudo de dados pontuais do tipo caso-controle identificando todas as múltiplas anomalias que sejam significativas, ou seja, com risco de ocorrência acima ou abaixo daquele que previamente seria esperado.

## 1.4 Organizacão da Tese

Esta tese está organizada da seguinte forma. No capítulo 2 descreve-se completamente o método *Voronoi Based Scan* (VBScan). No final do capítulo apresenta-se a motivação para seu uso na detecção de múltiplas anomalias espaciais. No capítulo 3 apresenta-se e descreve-se um método para detecção e inferência de partições, baseado no VBScan, utilizando uma abordagem multi-objetivo. No capítulo 4 apresenta-se resultados de simulações para testar a eficiência do método. Essa eficiência é medida em termos de poder e em termos do *matching*, uma medida apresentada para verificar a similaridade entre a partição verdadeira e a detectada. No capítulo 5 aplica-se o método em dois conjuntos de dados reais e analisamos os resultados obtidos. Finalmente no capítulo 6 faz-se as considerações finais desta tese e apresenta-se as perspectivas de trabalhos futuros.

# Capítulo 2

# Voronoi Based Scan (VBScan)

Aqui descreve-se um método de detecção de clusters espaciais para dados do tipo caso-controle, denominado *Voronoi Based Scan* - VBScan (Duczmal *et al.* (2011)).

A idéia de empregar uma Árvore Geradora Mínima (AGM), a fim de caracterizar clusters, já foi utilizada em Assunção *et al.* (2006), no contexto de dados de áreas. Para lidar com conjuntos de dados pontuais, a aplicação da estatística scan requer uma definição própria da densidade populacional relacionada a cada caso. Como um único raio de círculo não é adequado para estimar a densidade populacional em todas as regiões, devido a heterogeneidade na distribuição geográfica da população, um procedimento de correção é necessário. O procedimento proposto por Wieland *et al.* (2007) faz uma transformação não-linear do mapa, levando a um novo mapa com a distribuição dos controles aproximadamente homogênea. Esse procedimento é computacionalmente intensivo. No método VBScan, um procedimento mais simples para a estimação de densidade populacional é proposto.

O método VBScan, baseia-se na construção do que se denomina *diagrama de Voronoi*. Para maior entendimento deste procedimento, o diagrama de

Voronoi será definido formalmente a seguir:

**Definição 2.1 (Diagrama de Voronoi):** *considere um conjunto $\mathcal{A}$ de $n$ pontos em $\mathbb{R}^2$, $\mathcal{A} = \{A_i\}$, $i = 1, \ldots, n$. Dado um elemento $A_j \in \mathcal{A}$, considere o lugar geométrico de todos os pontos em $\mathbb{R}^2$ que são mais próximos de $A_j$, segundo a métrica euclidiana, do que de qualquer $A_i$ com $i \in \{1, \ldots, n\}$ e $i \neq j$. Este lugar geométrico é denominado por célula de Voronoi associada ao ponto $A_j$. A coleção de todas as células de Voronoi, associadas a cada um dos pontos do conjunto $\mathcal{A}$, determina o diagrama de Voronoi associado a este conjunto de pontos.*

De posse do diagrama de Voronoi, uma nova métrica entre dois pontos, denominada *distância de Voronoi*, será definida, de forma a transformar o mapa com densidade populacional heterogênea em homogênea.

**Definição 2.2 (A Distância de Voronoi):** *dados dois pontos $c_i$ e $c_j$, com $i \neq j$, pertencentes ao conjunto de pontos em estudo, define-se a distância de Voronoi entre eles, $\delta(i, j)$, como o número de arestas do diagrama de Voronoi*[1] *interceptadas pelo segmento de reta que liga $c_i$ e $c_j$.*

O método VBScan inicia-se pela construção de um diagrama de Voronoi para um conjunto de dados do tipo caso-controle, considerando todos os pontos. Neste momento, o interesse é conhecer a distância de Voronoi entre todos os casos do mapa em estudo, com os objetivos de estabelecer uma estratégia para estimar a densidade populacional para cada caso e determinar os candidatos a clusters.

[1] Segmentos que limitam as células de Voronoi.

## 2.1 Estimação da densidade populacional

Seja $c_i = (x_i, y_i)$, com $i = 1, \ldots, n$, a localização de $n$ casos numa região de estudo. Considere as arestas ponderadas da AGM de Voronoi (AGMV)[2] obtida destes casos e tome a aresta mínima dentre as que incidem em $c_i$, um destes casos. Seja $\omega_i = 2r$, o peso desta aresta. A *estimativa populacional* associada ao caso $c_i$ será dada pela área do círculo de raio $r$, ou seja, $\pi r^2$. Este círculo, será denotada por $\mathcal{C}(c_i, r)$. Desta forma, a definição da distância de Voronoi contém a informação necessária para calcular, aproximadamente, a função da densidade local da população heterogênea, para uma escolha adequada de vizinhos de cada caso individualmente.

**Proposição 2.1** *Considere um conjunto de casos $\mathcal{D}$ e sua correspondente AGMV, denotada por $\mathcal{V}$. Seja $T_S$ um subgrafo conectado de $\mathcal{V}$. Denote por $f(c_i)$ a densidade de população local em $c_i \in S$, em que $S$ é o conjunto de nós de $T_S$. Para cada caso $c_i \in S$, seja $\omega_i$ o peso mínimo das arestas em $\mathcal{V}$ que incidem em $c_i$ e $\mathcal{B} = \bigcup \mathcal{C}(c_i, \omega_i/2)$. A população local de $S$ pode ser aproximada por* $\int_{\mathcal{B}} f(x)dx = \frac{1}{4} \sum_{c_i \in S} \pi \omega_i^2$.

Isso define uma "região de influência" do cluster candidato $S$ através da composição das regiões de influência de cada caso, que são definidos como regiões circulares, com raio $\omega_i/2$ escolhido tão grande quanto possível, tal que não exista interferência entre círculos vizinhos na AGMV.

Nota-se ainda que essa definição é robusta no seguinte sentido. Considere duas situações: na primeira, um conjunto de casos $D$ espalha-se uniformemente num mapa de controles, e na segunda, um conjunto de casos $D'$ com o mesmo número de pontos e forma global de $D$, mas geograficamente menor,

[2] AGM cujos pesos são definidos pelas distâncias de Voronoi.

está inserido no mesmo mapa de controles. É fácil ver que as regiões de influência do cluster associado a $D$ é maior do que as correspondentes regiões de influência associada com $D'$, como poderíamos esperar.

Vamos utilizar essa informação para estimar o número de controles individuais sob a "região de influência" de cada caso individual que, por sua vez, permitirá o uso da estatística scan e também para definir um algoritmo de busca de clusters empregando a AGMV.

Com essa proposta, o cálculo da distância de Voronoi e de todas as entidades associadas pode ser realizado com algoritmos polinomiais eficientes. Utilizando a AGMV, a detecção de clusters do fenômeno de interesse pode ser realizada muito rapidamente.

No procedimento do cálculo das distâncias de Voronoi, algum segmento pode interceptar tangencialmente alguma das células de Voronoi. Assim, um problema pode ocorrer no cálculo de $\delta(i, j)$. Porém, esse problema ocorre raramente, supondo que as coordenadas dos pontos seguem um padrão aleatório. Uma estratégia de correção foi implementada.

No próximo capítulo é proposta uma definição alternativa para estimar a população de controles associada a cada caso, que, como será explicado, apresenta vantagens em relação à definição deste capítulo.

## 2.2 Conjunto dos possíveis clusters no espaço de coordenadas

Seja $\mathcal{D}$ o conjunto dos casos pertencentes ao mapa em estudo. Na tentativa de identificar quais subconjuntos de $\mathcal{D}$ são verossímeis a ponto de constituir um cluster, o seguinte procedimento heurístico é aplicado aqui: um subconjunto $\mathcal{S}$ de $\mathcal{D}$ forma um candidato a cluster se a menor distância que separa os

conjuntos $\mathcal{S}$ e $\mathcal{D} - \mathcal{S}$ é maior que a distância máxima interna de $\mathcal{S}$, em que $\mathcal{D} - \mathcal{S}$ é o subconjunto de $\mathcal{D}$ retirado todos os pontos de $\mathcal{S}$. Portanto, um cluster potencial é um subgrafo conectado com estrutura de árvore ligando casos no espaço de domínio. O algoritmo proposto constrói um conjunto de sub-árvores da árvore geradora mínima do grafo completo dos casos, definindo um subconjunto do espaço de clusters em potencial.

Formalmente, seja $\mathcal{D} = \{c_i; i = 1, \ldots, n\}$, em que $n$ representa o número total de casos do fenômeno de interesse no mapa em estudo, cuja localização geográfica é dada pelo par ordenado $(x_i, y_i)$. Define-se o grafo completo ponderado $\mathcal{G}(\mathcal{D}) = (\mathcal{V}, \mathcal{E})$ com conjunto de vértices $\mathcal{V} = \{c_i; c_i \in \mathcal{D}\}$, e conjunto de arestas $\mathcal{E} = \{(c_i, c_j); c_i, c_j \in \mathcal{D}, i \neq j\}$. Cada aresta $(c_i, c_j) \in \mathcal{E}$ tem peso definido pela distância de Voronoi, $\delta(i, j)$. Uma árvore geradora mínima (AGM) de um grafo completo $\mathcal{G}(\mathcal{D})$ pode ser definida como um conjunto minimal de arestas de $\mathcal{G}(\mathcal{D})$ que conecta todos os vértices com distância mínima total. A árvore geradora mínima de Voronoi (AGMV), do grafo ponderado $\mathcal{G}(\mathcal{D})$ definido acima, é uma AGM com a menor distância total de Voronoi.

Um conjunto de valores discretos caracteriza as distâncias de Voronoi. Isso pode causar o surgimento de múltiplas soluções muito fequentemente. Este efeito é eliminado ordenando as arestas de peso idênticos de acordo com a distância euclidiana. Esse procedimento garante o seguinte lema, que é uma extensão do resultado proposto por Wieland *et al.* (2007):

**Lema 2.1** *Assuma que a distância Euclidiana entre dois pontos quaisquer, pertencentes a um conjunto $\mathcal{P}$, é diferente de qualquer outra distância entre dois pontos do mesmo conjunto. Então o conjunto dos clusters em potencial estão em correspondência biunívoca com os componentes conectados entre todos os grafos $T_{w'}$ com $T_w$ derivado da AGMV, retirando-se todas as arestas tendo peso maior que $w$.*

**Demonstração**: *Ordene os pesos $w$ das arestas de AGMV no sentido decrescente, desempatando pela distância euclidiana. O restante da prova segue um procedimento análogo ao que foi realizado considerando a distância euclidiana (veja Wieland et al. (2007)).*

O conjunto dos clusters em potencial pode ser rapidamente encontrado utilizando um procedimento guloso de retirada das arestas, melhorando e simplificando a estratégia empregada pelo método aplicado em Wieland *et al.* (2007). O procedimento é: após construir a AGMV do conjunto $\mathcal{D}$ das localizações dos casos, remove-se, iterativamente, a maior aresta remanescente, dando origem a dois cluster candidatos adicionais em cada iteração. Para um mapa com $n$ casos, obtém-se $2n - 1$ clusters candidatos, incluindo os $n$ clusters unitários.

A Figura 2.1 mostra a distribuição espacial de 70 coordenadas, com 10 casos observados (círculos) e 60 controles (pontos) em um conjunto artificial de dados e a árvore geradora mínima de Voronoi associada.

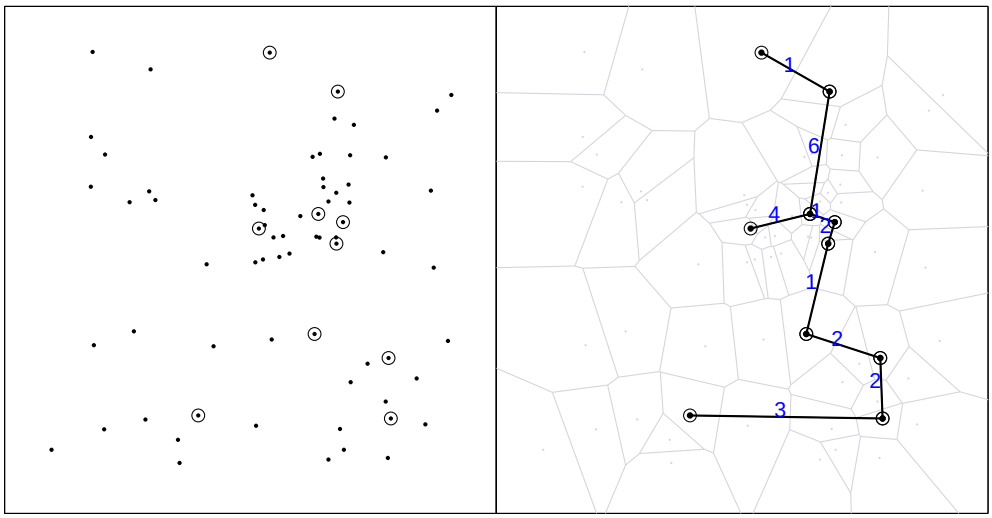


Figura 2.1: Distribuição espacial e árvore geradora mínima de Voronoi para um conjunto hipotético de pontos caso-controle.

A Figura 2.2 mostra uma visualização do procedimento guloso de retirada de arestas para o exemplo anterior. Os sucessivos passos da retirada de arestas estão representados com os novos clusters candidatos mostrados em cada iteração. Os sub-grafos ligando os círculos preenchidos representam os novos candidatos a clusters que aparecem em cada iteração, e os subgrafos ligando os círculos não preenchidos representam os candidatos a clusters que já apareceram em passos anteriores.

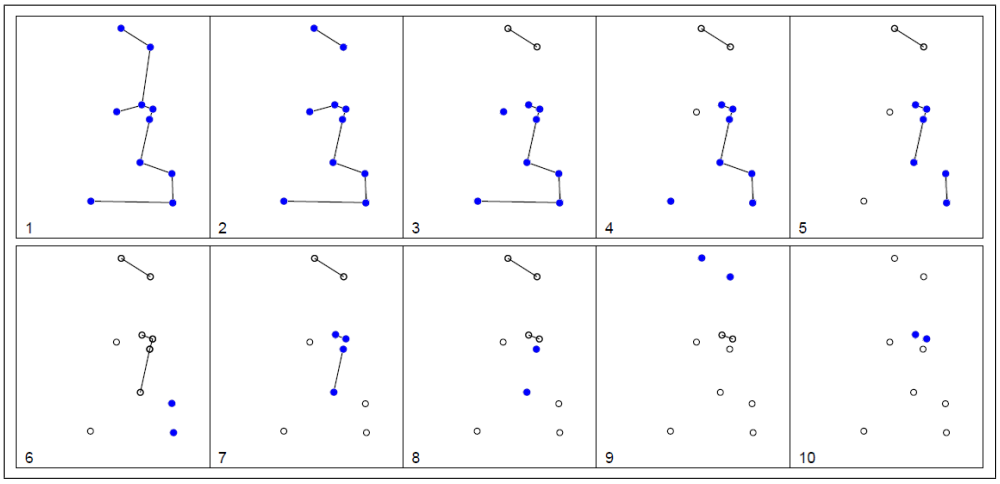


Figura 2.2: Visualização do procedimento guloso de retirada de arestas, em passos sucessivos enumerados de 1 a 10.

Proposto o método, é necessário avaliar a significância estatística das soluções encontradas e também avaliar quão bom o método é, através de métricas adequadas. Na próxima seção descreve-se o processo de inferência assim como define-se as medidas de eficiência, comumente utilizadas, para avaliação de métodos de detecção e inferência de clusters.

## 2.3 Inferência e medidas de eficiência para o VBScan

### 2.3.1 Inferência

De forma análoga ao procedimento descrito no Capítulo 1, que aborda a Estatística Scan Espacial, o processo do cálculo da significância do cluster mais verossímil do VBScan é feito da seguinte forma. Aplica-se o algoritmo do VBScan para os dados observados e obtém-se a sua estatística de teste, denominada $LLR_{obs}$. Sob a hipótese de não-existência de cluster (hipótese nula), realiza-se $n$ distribuições aleatórias para os casos (mantendo-se fixas as coordenadas dos pontos), ao longo do mapa em estudo, através do procedimento de Monte Carlo e obtém-se uma distribuição empírica para a estatística de teste. Conta-se então a quantidade $k$ de vezes que o valor da estatística de teste, obtido sob a hipótese nula, ultrapassa o valor $LLR_{obs}$. O valor $p$ para a solução encontrada nos dados observados é estimado por $\frac{k+1}{n+1}$.

### 2.3.2 Medidas de eficiência

Espera-se que um bom método de detecção de cluster seja sensível o suficiente para detectar um cluster quando este realmente existe. A eficiência do algoritmo será avaliada calculando-se seu poder de detecção, sua sensibilidade e seu valor de predição positiva (VPP).

**Definição 2.3 (Poder do teste):** *O poder de um teste de hipóteses é definido como a probabilidade de que a hipótese nula seja rejeitada quando esta é, de fato, falsa.*

O poder do método é, então, a probabilidade de que o método detecte um cluster quando este realmente existe. O poder será estimado através

de simulações Monte Carlo, executando o algoritmo $n$ vezes em cenários artificiais, construídos de forma que sabe-se que neles há a presença de um cluster. Assim, deve-se fazer a contagem da quantidade $m$ de vezes em que um cluster foi detectado no mapa em estudo, visando estimar a probabilidade desejada. Desta forma, o poder será dado pela proporção, $m/n$, de detecções em relação ao número total de execuções.

Além do poder, outras medidas bastante utilizadas para avaliação da eficiência de algoritmo de detecção de cluster são a *sensibilidade* e o *valor de predição positivo* (VPP). Considere que os $n$ casos, no mapa em estudo, são denotados por $\mathcal{D} = \{c_i\}$, $i = 1, \ldots, n$. A sensibilidade e o VPP são definidos, aqui, como:

$$Sensibilidade = \frac{\sum_{i=1}^{n} \mathbb{1}(c_i \in \text{Cluster Detectado} \cap \text{Cluster Verdadeiro})}{\sum_{i=1}^{n} \mathbb{1}(c_i \in \text{Cluster Verdadeiro})} \tag{2.1}$$

$$VPP = \frac{\sum_{i=1}^{n} \mathbb{1}(c_i \in \text{Cluster Detectado} \cap \text{Cluster Verdadeiro})}{\sum_{i=1}^{n} \mathbb{1}(c_i \in \text{Cluster Detectado})} \tag{2.2}$$

, em que $\mathbb{1}(.)$ é a função indicadora.

No próximo capítulo, será apresentada uma proposta de detecção de múltiplas anomalias, baseado no método VBScan. Não serão apresentados aqui resultados numéricos que atestem a eficiência do algoritmo VBScan. Todos as avaliações numéricas serão apresentadas em conjunto, considerando esses dois métodos, nos capítulos 4 e 5.

# Capítulo 3

# Multi-Objective Multiple Clusters (MOMC-VBScan)

Este novo método proposto baseia-se na aplicação da técnica apresentada no capítulo anterior (VBScan) de forma recursiva, visando possibilitar a busca de múltiplas anomalias, sendo elas de alto ou baixo risco, num mapa composto por dados pontuais do tipo caso-controle. O método propõe-se, também, a avaliar as múltiplas anomalias através de uma técnica multi-objetivo.

O método pressupõe, primeiramente, uma mudança na estimação da densidade populacional, associada a cada caso, em relação à utilizada no VBScan. Esta mudança visa, além de estimar a densidade populacional de um modo mais preciso, identificar quais os pontos que estão sob a influência de cada um dos casos no mapa em estudo. Nesta formulação, os pontos correspondentes a controles que não estão sob influência de qualquer um dos casos, constituirão o que, doravante, passaremos a denominar por *região branca*.

**Definição 3.1 (A estimativa populacional):** *para cada caso no mapa em estudo é calculada da seguinte forma: dado um caso $c_i$, considere as arestas ponderadas na AGMV, e tome a aresta mínima dentre todas as que incidem*

*sobre $c_i$. A população estimada associada a este caso é dada pelo próprio $c_i$ além dos controles que estão mais próximos de $c_i$ do que de qualquer outro caso, segundo a métrica de Voronoi, e cuja distância em relação ao caso $c_i$ seja menor ou igual ao comprimento da aresta mínima incidente sobre ele, também segundo a métrica de Voronoi.*

Uma segunda alteração no VBScan é o relaxamento da restrição de busca por clusters de alto risco ($c_z > \mu_z$) na função $LR(z)$ (veja Eq. 1.1). A intenção é que, agora, ao se utilizar o VBScan, ele seja hábil para identificar também os clusters de baixa intensidade.

A função $LR(z)$ fica, então, assim:

$$LR(z) = \left(\frac{c_z}{\mu_z}\right)^{c_z} \left(\frac{C - c_z}{C - \mu_z}\right)^{C - c_z} \tag{3.1}$$

Dado que o interesse central deste método está na identificação de múltiplas anomalias, introduziremos o conceito de partição do mapa em estudo.

**Definição 3.2 (Partição):** *seja $\mathcal{P}$ o conjunto de pontos (caso-controle) no mapa em estudo. Considere os subconjuntos $\mathcal{P}_1, \ldots, \mathcal{P}_k$ de $\mathcal{P}$, tais que $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset, i \neq j$ e $\bigcup_{i=1}^{k} \mathcal{P}_i = \mathcal{P}$. Dizemos que $\mathcal{P}_1, \ldots, \mathcal{P}_k$ é uma partição de $k$ componentes do mapa em estudo.*

O interesse do estudo é a detecção da partição mais significativa do mapa. Para tanto, a expressão da $LR(z)$ (como em 3.1) será adaptada para a situação de partição. Considere que o mapa esteja particionado em $m$ componentes ditas $z_1, \ldots, z_m$. A $LR(z_1, \ldots, z_m)$ desta partição, é definida como:

$$\begin{aligned} LR(z_1, \ldots, z_m) &= \left(\frac{C - (c_1 + \ldots + c_{m-1})}{C - (\mu_1 + \ldots + \mu_{m-1})}\right)^{C - (c_1 + \ldots + c_{m-1})} \times \prod_{i=1}^{m-1} \left(\frac{c_i}{\mu_i}\right)^{c_i} \\ &= \prod_{i=1}^{m} \left(\frac{c_i}{\mu_i}\right)^{c_i}, \end{aligned} \tag{3.2}$$

em que $c_i$ e $\mu_i, i = 1, \ldots, m$, representam o número de casos observados e esperados em cada componente, respectivamente.

Tomando-se o logaritmo, tem-se:

$$LLR(z_1, \ldots, z_m) = \sum_{i=1}^{m} c_i \log\left(\frac{c_i}{\mu_i}\right) \tag{3.3}$$

O relaxamento da restrição original na função $LR(z)$ torna a função $LLR(z_1, \ldots, z_m)$ não decrescente com respeito ao número de componentes $m$. Essa propriedade é uma consequência do seguinte lema:

**Lema 3.1** *Sejam* $\{u, v, x, y\} \subset (0, +\infty)$*, então*

$$\left(\frac{u+v}{x+y}\right)^{u+v} \leq \left(\frac{u}{x}\right)^{u} \left(\frac{v}{y}\right)^{v}.$$

**Demonstração**:

$$\begin{array}{rcl} \left(\frac{u+v}{x+y}\right)^{u+v} & \leq & \left(\frac{u}{x}\right)^{u} \left(\frac{v}{y}\right)^{v} \\ \log\left(\left(\frac{u+v}{x+y}\right)^{u+v}\right) & \leq & \log\left(\left(\frac{u}{x}\right)^{u} \left(\frac{v}{y}\right)^{v}\right) \\ (u+v)\big(\log(u+v) - \log(x+y)\big) & \leq & u\big(\log(u) - \log(x)\big) + v\big(\log(v) - \log(y)\big) \\ u\big(\log(u) - \log(x)\big) + v\big(\log(v) - \log(y)\big) & + & (u+v)\big(\log(x+y) - \log(u+v)\big) \geq 0. \end{array} \tag{3.4}$$

*Agora considere a função* $f : \mathbb{R}^4 \to \mathbb{R}$ *definida por:*

$$\begin{aligned} f(u, v, x, y) = u\big(\log(u) - \log(x)\big) &+ v\big(\log(v) - \log(y)\big) + \\ &+ (u+v)\big(\log(x+y) - \log(u+v)\big). \end{aligned} \tag{3.5}$$

*Se fixarmos* $u, v$ *e* $y$*, teremos uma função na variável* $x$ *que, por simplicidade, denotaremos também por* $f$*.*

*Nessas condições temos*

$$f'(x) = \frac{u+v}{x+y} - \frac{u}{x} = \frac{vx - uy}{x(x+y)}.$$

*Resulta daí que, o único ponto crítico de $f$ é $x = \frac{uy}{v}$. Se $x < \frac{uy}{v}$ então $f'(x) < 0$, e se $x > \frac{uy}{v}$ então $f'(x) > 0$; conclui-se que $f$ possui um mínimo global no ponto $x = \frac{uy}{v}$. Agora vamos calcular o valor desse mínimo global.*

$$\begin{aligned} f\left(\tfrac{uy}{v}\right) &= u\left(\log(u) - \log\left(\tfrac{uy}{v}\right)\right) + v\left(\log(v) - \log(y)\right) + \\ &\quad + (u+v)\left(\log\left(\tfrac{uy}{v} + y\right) - \log(u+v)\right) \\ &= u\log\left(\tfrac{v}{y}\right) + v\log\left(\tfrac{v}{y}\right) + \\ &\quad + (u+v)\left(\log\left(\tfrac{y}{v}\right) + \log(u+v) - \log(u+v)\right) = 0. \end{aligned} \tag{3.6}$$

*Como os pontos $u, v$ e $y$ são arbitrários, conclui-se que a função $f$ dada pela expressão (3.5) possui mínimo global que é 0. Resulta daí e de (3.4) que a desiguladade expressa no Lema 3.1 é verdadeira.*

Desta forma, avaliar somente a função $LLR(z_1, \ldots, z_m)$ se torna insuficiente, pois necessariamente, uma das soluções do problema seria considerar todos os casos separadamente, ou seja, todos os pontos individuais seriam componentes da partição do mapa. Assim sendo, a motivação da utilização de uma técnica multi-objetivo surge naturalmente.

Dito isto, faremos na próxima seção uma descrição da abordagem multi-objetivo.

## 3.1 Otimização multi-objetivo

A abordagem de otimização multi-objetivo é utilizada para modelar diversos problemas reais. Consiste em encontrar soluções ótimas considerando uma função objetivo com imagem em um espaço de dimensão superior a 1, podendo ser formalmente definido como:

$$\begin{aligned} \max_{x} \quad & f(x) = (f_1(x), f_2(x), \ldots, f_n(x)) \\ \text{sujeito a:} \quad & \textit{possíveis restrições de factibilidade} \end{aligned}$$

Ao analisarmos um único objetivo, o conjunto imagem desta função possui elementos pertencentes à reta. Portanto, podendo ser classificados pela ordem existente na reta. Quando partimos para uma abordagem multi-objetivo, o conjunto imagem da função objetivo possui elementos pertencentes ao $\mathbb{R}^n$, não possuindo, então, uma relação de ordem total. Para estabelecer uma relação de ordem neste tipo de conjunto, definiremos o conceito de *dominância*.

**Definição 3.3 (Dominância):** *Seja* $f(x) = (f_1(x), f_2(x), \ldots, f_n(x))$ *uma função definida em um espaço* $\mathcal{X}$. *Um ponto* $x_1 \in \mathcal{X}$ *domina outro ponto* $x_2 \in \mathcal{X}$ *(denota-se* $x_1 \succ x_2$*) se* $f_i(x_1) \geq f_i(x_2), i = 1, \ldots, n$ *e se existe, pelo menos, um índice* $k \in \{1, ..., n\}$ *tal que* $f_k(x_1) > f_k(x_2)$.

De outra forma, um ponto $x_1$ domina o ponto $x_2$ se a avaliação de $x_1$ for superior que a avaliação de $x_2$ em, pelo menos, um objetivo e não for inferior nos demais. Se o problema for de minimização, basta trocar os sinais, na definição de dominância por $\prec$, $\leq$, e $<$, respectivamente.

A Figura 3.1 ilustra o conceito de dominância, considerando a situação na qual se deseja maximizar os objetivos $f_1$ e $f_2$. Na ilustração temos que os pontos B, D e E representam soluções não dominadas, e os pontos A e C soluções dominadas. Note que o ponto C domina o ponto A, entretanto é dominado pelo ponto D. Observe ainda, que entre os pontos B, D e E não existe relação de dominância.

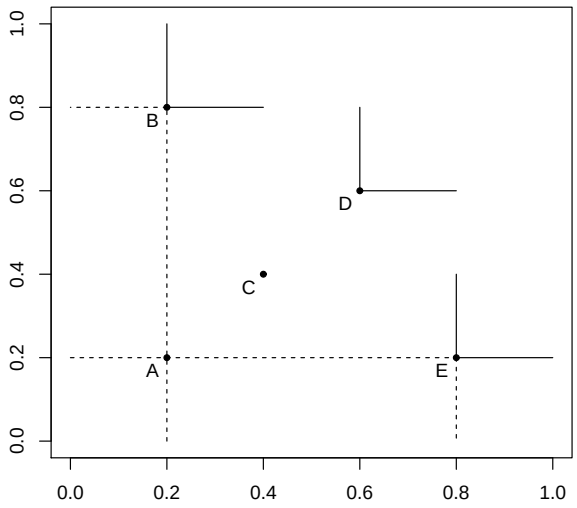


Figura 3.1: Conjunto de pontos ilustrando o conceito de dominância.

Uma vez definido o conceito de dominância, podemos definir a *solução Pareto-ótima*, um objeto de extrema relevância na resolução de problemas de otimização multi-objetivo.

**Definição 3.4 (Solução Pareto-ótima):** *uma solução $x^* \in \mathcal{X}$ é Pareto-ótima se não existe outra solução $x \in \mathcal{X}$, tal que $x$ domina $x^*$.*

Em outras palavras, uma solução é Pareto-ótima (solução não-dominada) se não é inferior à nenhuma outra. O *conjunto Pareto-ótimo* é formado por todas as soluções Pareto-ótimas. Assim, ao contrário do que ocorre em problemas de otimização mono-objetivo, temos um conjunto de soluções que são, em um certo sentido, ótimas. A Figura 3.2 ilustra o conceito de pontos dominados e o de pontos que formam o conjunto de Pareto-Ótimo.

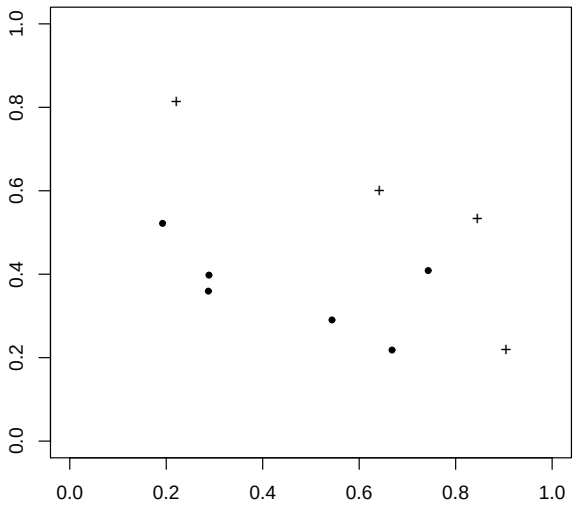


Figura 3.2: Conjunto Pareto-Ótimo (+) e pontos dominados (•).

Voltando à questão de detecção de partições, o problema de encontrar soluções candidatas ao melhor particionamento do mapa, será tratado como um problema de maximização bi-objetivo, em que as funções objetivo são definidas como: $LLR$, redefinida para partições (equação 3.3) , e a $\Delta LLR$ que representa o ganho da $LLR$ ao se acrescentar uma componente na partição do mapa em estudo.

Na próxima seção, descreve-se o processo para se encontrar o conjunto Pareto-ótimo (partições candidatas) para este problema.

## 3.2 O processo de busca de partições candidatas

Considere um mapa com população igual a $N$, sendo que, destes, $n$ são casos. O processo de busca das soluções candidatas às partições do mapa é descrito

através dos seguintes passos:

(1) construa o diagrama de Voronoi para todos os casos e controles, como descrito anteriormente no VBScan;

(2) construa a AGM dos casos utilizando a distância de Voronoi, assim como no método VBScan;

(3) encontre a região de influência dos casos, segundo a definição 3.1. Neste momento, o mapa é particionado em duas regiões: aquela composta por todos os casos e controles sob sua influência, e a região desprovida de casos (região branca), que pode não ser necessariamente conexa;

(4) calcule o valor da função $LLR$ para a partição definida no terceiro passo. Para essa partição define-se o valor da segunda função objetivo como sendo $\Delta LLR = LLR$. Esta será a primeira partição do mapa, pertencente ao conjunto solução Pareto-ótimo do nosso problema de otimização bi-objetivo. O valor de $LLR$ e $\Delta LLR$, desta primeira solução, é o ponto de partida para o cálculo do valor das funções objetivo de todas as outras partições candidatas. Este passo é considerado o nível de recursividade zero;

(5) aplica-se o VBScan, na região de domínio dos casos, e encontra-se o cluster mais verossímil. Uma nova partição é encontrada; composta por três componentes: a região branca, a região composta pelos casos e controles do cluster mais verossímil e a região composta por todos os outros casos e controles que não pertencem à região branca e nem à região do cluster mais verossímil. Calcula-se o valor das funções objetivos para essa nova partição. Essa partição é uma solução factível, candidata à pertencer ao conjunto Pareto-ótimo. Este passo é considerado

o nível de recursividade um;

(6) o cluster mais verossímil, encontrado no passo anterior, é retirado da árvore geradora mínima, produzindo assim, $k$ subárvores, em que $1 \leq k \leq n-1$, cada uma delas com sua respectiva área de influência. Em cada uma das $k$ subárvores, aplica-se o VBScan, encontrando-se o cluster mais verossímil de cada uma delas. Nesse momento, encontram-se novas $2^k - 1$ partições candidatas, sendo que cada uma delas é composta por: a região branca, a região do cluster do nível de recursividade 1, a(s) região(ões) do(s) cluster(s) mais verossímil(meis) da(s) subárvore(s) considerada(s) e a região composta por todos os outros casos e controles que não pertencem nem à região branca e nem à região dos clusters encontrados em níveis de recursivade anteriores. Para cada uma das partições, calcula-se o valor de suas funções objetivo, $LLR$ e $\Delta LLR$. Este é o segundo nível de recursividade;

O passo 6 é realizado, repetidas vezes, até que se atinja um nível máximo, pré-estabelecido, de recursividade. Terminado este processo, o conjunto de todas as soluções factíveis estarão disponíveis.

(7) determina-se o conjunto Pareto-ótimo das soluções factíveis encontradas, até o nível máximo de recursividade utilizado.

Para o bom entendimento quanto aos passos 4 a 7, apresenta-se um exemplo ilustrado, nas Figuras 3.3, 3.4 e 3.5. Neste exemplo, o nível de recursividade máximo considerado é 2.

Na Figura 3.3 (a) observa-se a AGMV obtida para o exemplo, correspondendo ao nível zero de recursividade. Na Figura 3.3 (b) é destacado o cluster mais verossímil, assim como as subárvores encontradas, correspondendo ao nível um de recursividade.

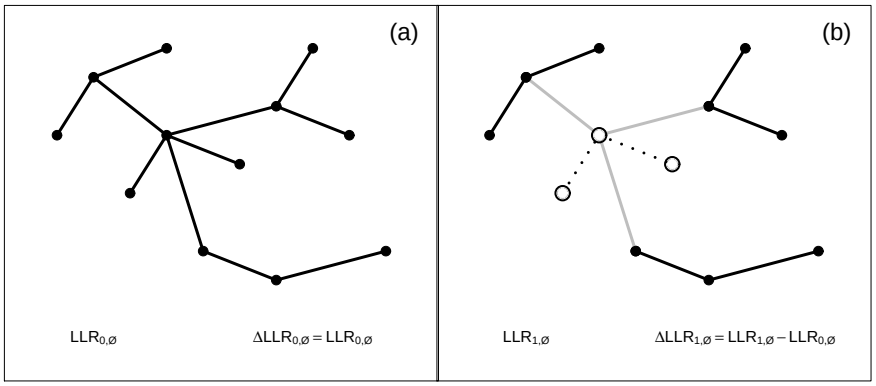


Figura 3.3: (a) AGMV e (b) cluster mais verossímil detectado.

Na Figura 3.3 (b), nota-se que, ao se retirar o cluster mais verossímil, foram obtidas três subárvores. Desta forma, o número de partições candidatas a serem avaliadas é, em princípio, $2^3 - 1 = 7$. Na Figura 3.4 são mostradas as partições candidatas composta pelo cluster mais verossímil e por cada um dos clusters mais verossímeis das subárvores.

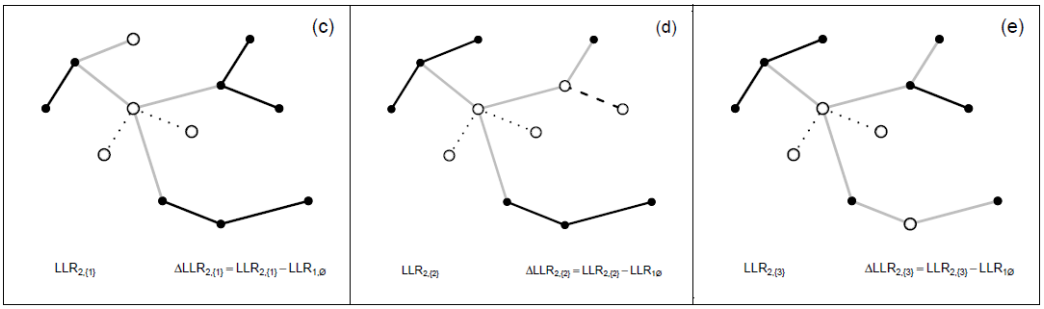


Figura 3.4: Partições candidatas compostas pelo cluster mais verossímil e respectivos clusters nas subárvores.

Na Figura 3.5 são apresentadas as demais partições candidatas, compostas por todas as combinações possíveis considerando os clusters já encontrados.

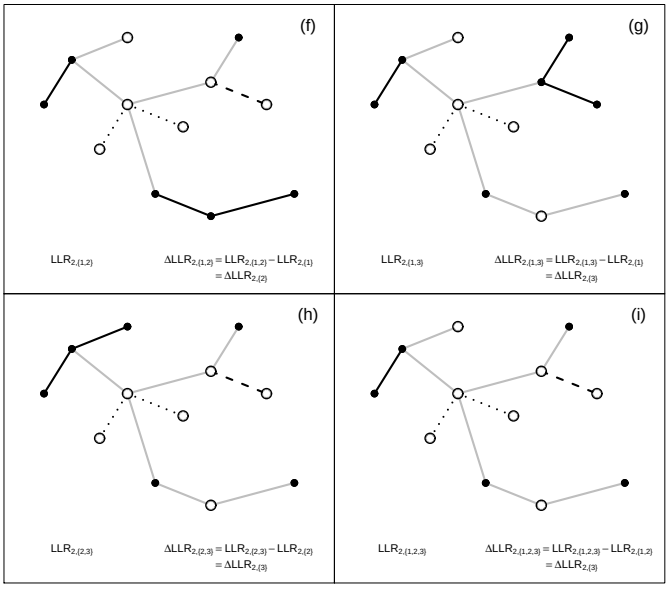


Figura 3.5: Demais partições candidatas.

Os pontos representando todas as soluções factíveis são apresentados na Figura 3.6.

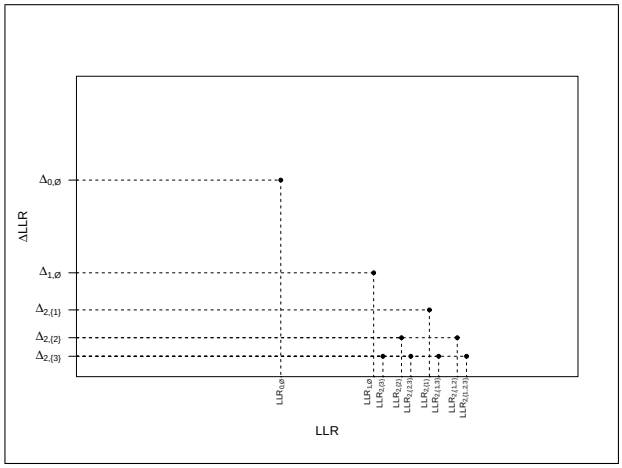


Figura 3.6: Soluções factíveis encontradas no exemplo comentado.

Na Figura 3.7 mostra-se o conjunto Pareto-Ótimo correspondente às soluções factíveis encontradas em uma realização do algoritmo. Percebe-se que o conjunto Pareto-Ótimo de soluções no nível de recursividade 2 é formado por apenas três das soluções factíveis. Em geral, se $k$ subárvores são formadas

em um nível de recursividade, precisa-se avaliar apenas $k$ das $2^k - 1$ partições candidatas possíveis, pois as demais $(2^k - 1) - k$ são dominadas. Isto representa uma economia significativa no tempo de execução do algoritmo.

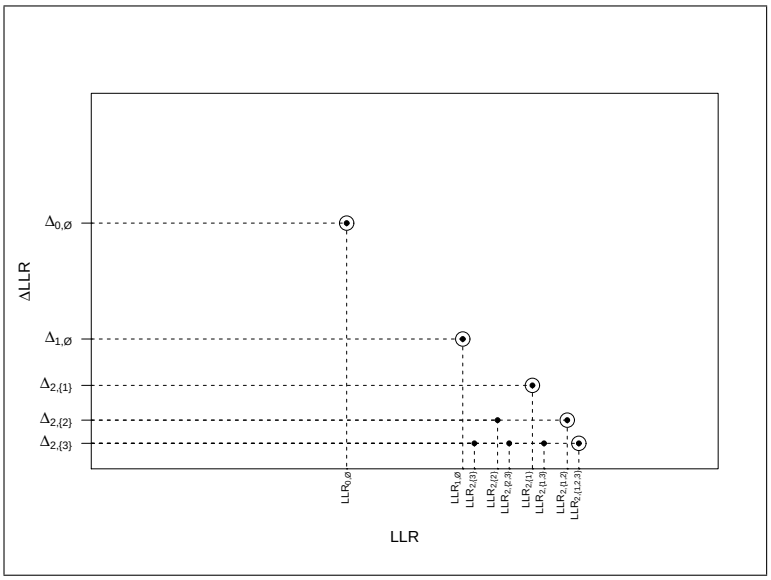


Figura 3.7: Conjunto Pareto-ótimo para o exemplo comentado.

Encontrado o conjunto Pareto-ótimo, deseja-se inferir sobre o conjunto de soluções encontradas. Como este é um problema bi-objetivo, precisa-se de uma nova forma para calcular a significância estatística para estas soluções. Adotaremos o método proposto por Cançado *et al.* (2010), em que utiliza-se o conceito de *Superfície de Aproveitamento* (SA). Na próxima seção descrevemos esse método adequando-o ao nosso problema.

## 3.3 Calculando a significância estatística das partições

Nesta seção descrevemos o processo para se encontrar a Superfície de Aproveitamento (SA) e a adequação ao estudo da tese.

As definições deste parágrafo são discutidas em Cançado *et al.* (2010), a partir dos trabalhos de da Fonseca *et al.* (2001) e Fonseca *et al.* (2005). Considere um problema de maximização bi-objetivo, com objetivos $f_1$ e $f_2$. Seja $\mathcal{E} = \{x_j, j = 1, \ldots, Q\}$ o conjunto de todas as soluções obtidas em uma realização da estratégia de otimização, e defina sua imagem como $\mathcal{I} = \{Y_j = (f_1(x_j), f_2(x_j)), j = 1, \ldots, Q\}$, contida no espaço de objetivos contido no $\mathbb{R}^2$. Uma solução $x_j$ é chamada de *não-dominada* se $x_j$ não é dominada por qualquer outra solução em $\left\{x_j^*, j = 1, \ldots, q\right\} \subset \mathcal{E}$. Seja $\left\{x_j^*, j = 1, \ldots, q\right\} \subset \mathcal{E}$ o conjunto de soluções não-dominadas de $\mathcal{E}$. O subconjunto $\mathcal{Y} = \left\{Y_j^* = (f_1(x_j^*), f_2(x_j^*)), j = 1, \ldots, q\right\} \subset \mathcal{I}$ é definido como o *resultado* de uma única execução de um algoritmo bi-objetivo.

Pode-se associar a $\mathcal{Y}$ uma fronteira que divide o espaço de objetivos em duas regiões $R_1$ e $R_0$: $R_1$ é a região consistindo de pontos dominados por, ou igual a, pelo menos um ponto em $\mathcal{Y}$; e $R_0$ consistindo dos pontos que não são dominados por nenhum dos pontos em $\mathcal{Y}$ (veja Figura 3.8).

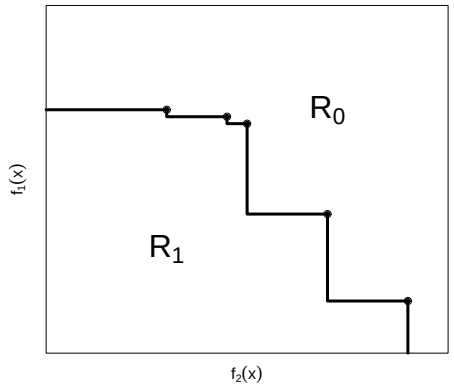


Figura 3.8: Fronteira entre $R_0$ e $R_1$.

Quando a solução $x$ é dominada por, pelo menos, uma solução de $\mathcal{Y}$, tem-se que $x$ é atingida por $\mathcal{Y}$. Na Figura 3.8, qualquer solução localizada na região $R_1$ é atingida por $\mathcal{Y}$. Agora, considere $n$ realizações do algoritmo.

Considerando entradas de dados distintas, a cada realização do algoritmo serão produzidas diferentes saídas, obtendo-se assim múltiplas fronteiras, como pode ser visto na Figura 3.9 (a).

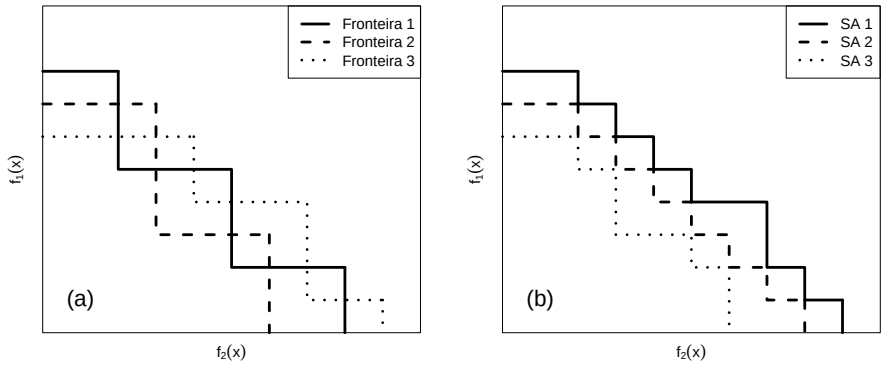


Figura 3.9: Múltiplas Fronteiras e respectivas Superfícies de Aproveitamento.

Pontos situados no canto superior direito da Figura 3.9 (a) não são atingidos por nenhuma das fronteiras. Pontos localizados no canto inferior esquerdo são atingidos por todas as fronteiras. E pontos situados entre as diferentes fronteiras foram atingidos em algumas realizações, mas em outras não. Assim podemos dividir o espaço em $n + 1$ regiões de acordo com a frequência em que estas regiões são atingidas. As fronteiras dessas regiões são chamadas de *superfícies de aproveitamento* (veja Figura 3.9 (b)). Essas frequências são utilizadas para estimar a probabilidade de atingirmos um ponto no espaço de objetivos, quando um grande número de realizações é executado.

A função de aproveitamento avaliada em um ponto $Y$ no espaço de objetivos pode ser estimada pelos conjuntos das saídas $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ obtidas através de $n$ realizações independentes do algoritmo, como:

$$A_n(Y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\mathcal{Y}_i \trianglerighteq Y)$$

, em que o símbolo "$\trianglerighteq$" significa que $\mathcal{Y}_i$ atingiu $Y$ e $\mathbb{1}(.)$ é a função indicadora assumindo valor 1 se $\mathcal{Y}_i \trianglerighteq Y$ e zero caso contrário.

No problema específico, em estudo nesta tese, o interesse está em estimar o valor $p$ das partições candidatas não-dominadas representadas por pontos no espaço de objetivos $(LLR, \Delta LLR)$, em que $\Delta LLR$ é a diferença de $LLR$ da partição candidata em relação a uma partição de referência. Formalmente, define-se $A(Y)$ como o $\lim_{n\to\infty} A_n(Y)$ quando ele existe. Agora, dado que $0 < p \leq 1$, a *isolinha do valor p* é definida como a imagem inversa $A^{-1}(p)$. Sob certas condições de suavidade, $A^{-1}(p)$ é uma superfície uni-dimensional dividindo o espaço de objetivos em duas regiões $R_0$ e $R_1$, tais que se $Y \in R_1$ então $A(Y) > p$, e se $Y \in R_0$ então $A(Y) \leq p$. Na prática, dadas $n$ saídas $Y_1, \ldots, Y_n$ pode-se construir aproximações das isolinhas de valores $p$ para cada $p = i/(n+1), i = 1, \ldots, n+1$ através das funções de aproveitamento estimadas $A_n(Y)$. A Figura 3.10 ilustra estas aproximações de algumas isolinhas de valores $p$ resultantes de $n = 10.000$ conjuntos Pareto-ótimos obtidos por simulações de Monte Carlo sob a hipótese nula.

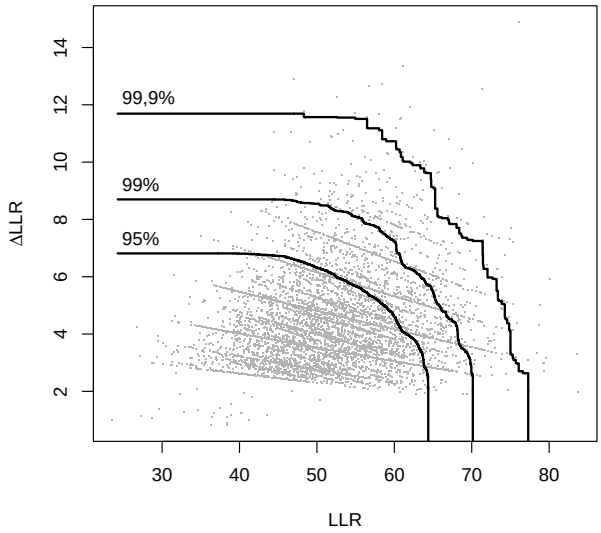


Figura 3.10: Os 10.000 conjuntos Pareto-ótimo obtidos por simulações de Monte Carlo sob a hipótese nula são representados pelos pontos e são apresentadas as isolinhas correspondentes a 95%; 99% e 99,9%.

O algorítmo desenvolvido vasculha somente parte do conjunto de soluções potenciais, e portanto, não existe garantia de que a solução ótima não-dominada foi encontrada. Isto, claro, pode nos levar a uma estimação viciada da significância, produzindo p-valores subestimados. Assim os p-valores calculados, são de fato bordas inferiores dos p-valores teóricos. Visando reduzir esta subestimação, as superfícies de aproveitamento obtidas neste trabalho foram separadas por número de componentes das partições encontradas (a idéia dessa separação originou-se do trabalho de Almeida *et al.* (2011)), eliminando assim o efeito de número de componentes no cálculo do valor $p$.

Precisa-se, de alguma maneira, avaliar a eficiência do algoritmo proposto. Uma das medidas usadas é o poder do teste. Como medida de sensibilidade apresenta-se o *matching*, que será definido e detalhado na próxima seção.

## 3.4 Medidas de eficiência do algoritmo

### 3.4.1 Poder

O poder do teste de hipótese para o método de detecção de partição, será calculado da seguinte forma. Considere a superfície de aproveitamento de $(1-\alpha)\%$, gerada sob a hipótese nula. Dado um cenário sob a hipótese alternativa, realiza-se $n$ simulações Monte Carlo. Conta-se em quantas destas $n$ simulações pelo menos uma das soluções pertencente ao conjunto Pareto-ótimo encontrado supera a SA de $(1-\alpha)\%$. Supondo que seja $m$ essa quantidade, então, o poder será dado pela proporção $m/n$.

### 3.4.2 *Matching*

Aqui, descreve-se uma medida para avaliar a sensibilidade do método MOMC-VBScan em identificar a verdadeira partição gerada, denominada *matching*. A Figura 3.11 considera um exemplo para o melhor entendimento do conceito da medida proposta. Nesta Figura 3.11 (à esquerda) está a localização dos casos de uma mapa artificial gerado. Na mesma Figura (à direita) está a identificação de cada um destes casos.

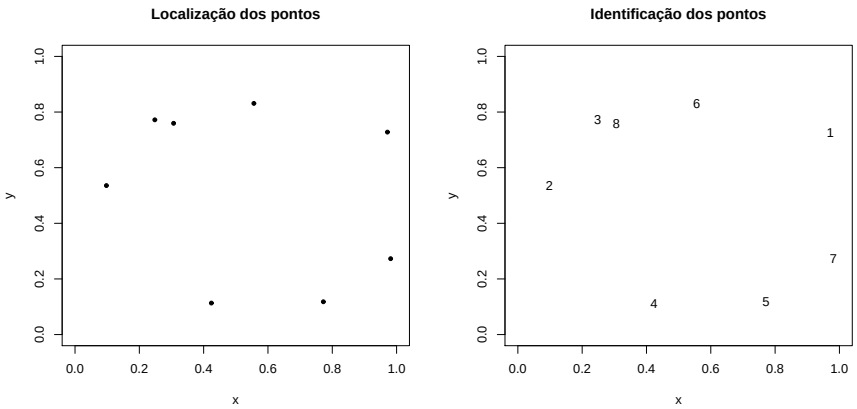


Figura 3.11: Localização e identificação dos casos de uma partição artificial gerada.

Suponha que as componentes da partição artificial simulada, e da detectada pelo algoritmo, sejam aquelas mostradas na Figura 3.12.

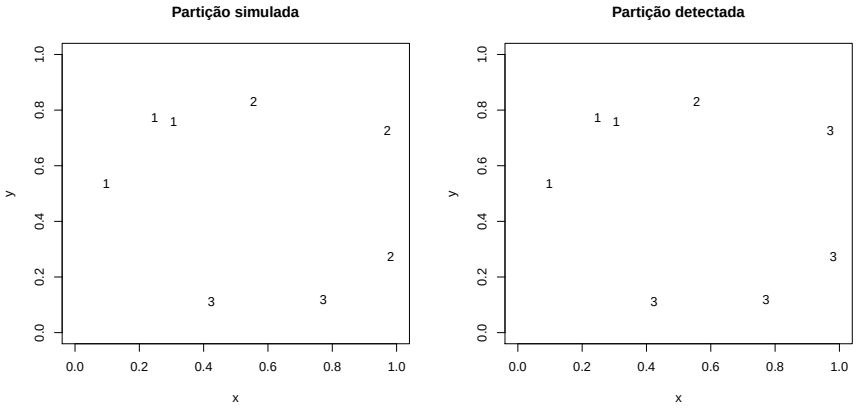


Figura 3.12: Partição artificial simulada e detectada.

Para cada componente da partição detectada, verifica-se o número de interseções com cada uma das componentes da partição simulada. O *matching* é a razão entre o maior número de interseções obtida pelo número de casos. As Figuras 3.13, 3.14 e 3.15 ilustram o cálculo do *matching* para todas as configurações possíveis para o exemplo citado.

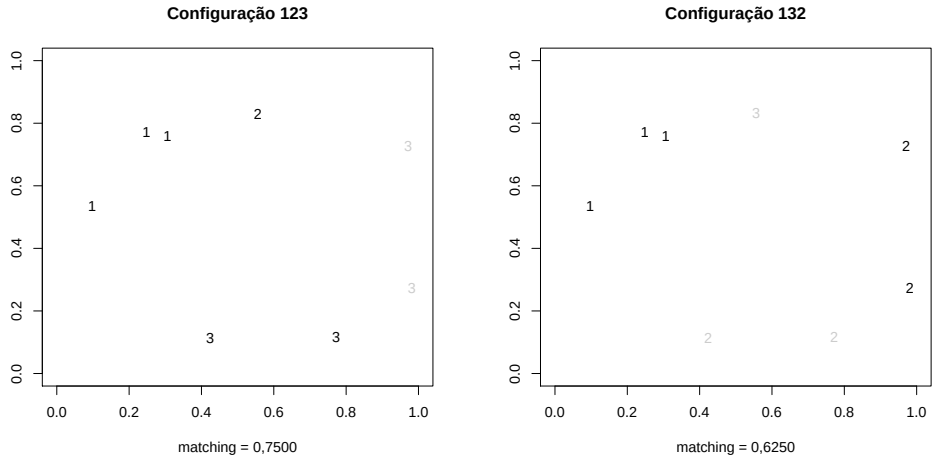


Figura 3.13: Configurações 123 e 132, e respectivo *matching*, entre as partições simulada e detectada.

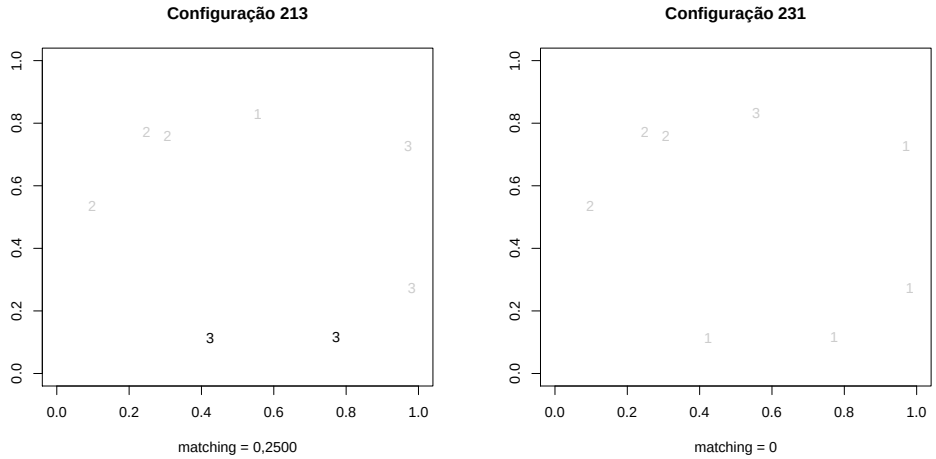


Figura 3.14: Configurações 213 e 231, e respectivo *matching*, entre as partições simulada e detectada.

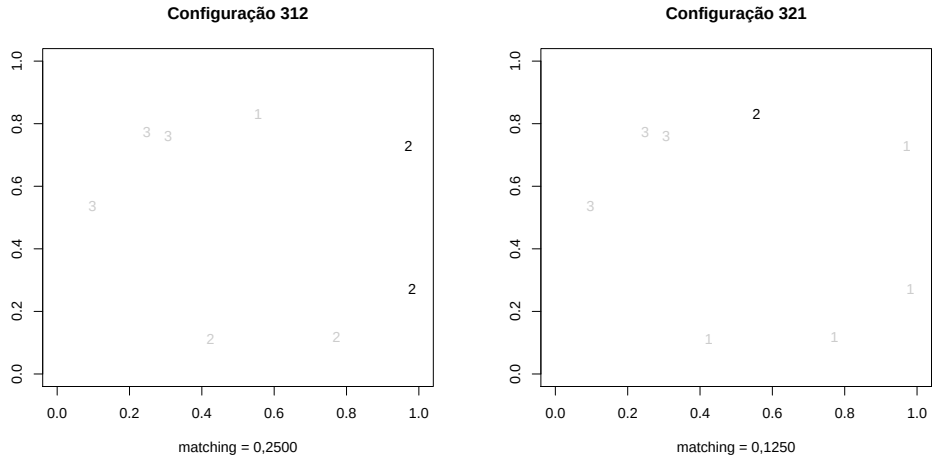

Figura 3.15: Configurações 312 e 321, e respectivo *matching*, entre as partições simulada e detectada.

O *matching* entre a partição artificial simulada e a detectada, deste exemplo comentado, é o maior deles $(0,7500)$.

Pelo exposto, percebe-se que se o número de componentes da partição simulada, $m$, é menor ou igual ao número de componentes da partição detectada, $n$, então o número de configurações a ser calculado o *matching* é equivalente ao número de arranjos de $n$ elementos tomados $m$ a $m$. Caso contrário, será equivalente ao número de arranjos de $m$ elementos tomados $n$ a $n$.

Supondo, agora, que são realizadas $n$ simulações sob a hipótese alternativa, o *matching* é dado pelo valor médio do máximo dos *matching* de cada realização. Assim, o valor do *matching* é definido por:

$$Matching = \frac{1}{n}\sum_{i=1}^{n} \max_{i}\{match_j\}, \tag{3.7}$$

em que $match_j$ é o *matching* de cada realização de Monte Carlo sob a hipótese alternativa.

No próximo capítulo apresentaremos resultados de várias simulações numericas realizadas, com o intuito de verificar a eficiência do método proposto.

# Capítulo 4

# Resultados de Simulações

Neste capítulo apresentam-se e discutem-se os resultados obtidos em simulações numéricas, com o objetivo de verificar a eficiência dos métodos VBScan e MOMC-VBScan.

## 4.1 VBScan

Nesta seção, a estatística scan baseada em Voronoi (VBScan) é comparada numericamente com a versão elíptica[1] da estatística scan, de acordo com o poder de detecção, sensibilidade e valor de predição positiva.

### 4.1.1 Clusters artificiais

Para se medir a eficiência do algoritmo VBScan para detecção de clusters espaciais foram criados: (1) um conjunto de dados artificiais, caso-controle, com população total em risco de 1.000 indivíduos, sendo 100 casos, e (2) três clusters espaciais com formas geométricas diferentes. Os cenários simulados foram criados dentro de um quadrado unitário $[0, 1] \times [0, 1]$ com a população

[1] Procedimento análogo ao scan circular considerando janela de varredura elíptica.

em risco seguindo um processo pontual uniforme. A Figura 4.1 mostra a forma destes três clusters:

(1) Um cluster circular com raio igual a $0,195$.

(2) Um cluster com formato de um "T" sendo constituído por: $T = T_1 \cup T_2$, onde $T_1 = [0,2; 0,4] \times [0,5; 0,8]$, $T_2 = [0,0; 0,6] \times [0,8; 0,9]$.

(3) Um cluster com formato de um "L" sendo constituído por: $L = L_1 \cup L_2$, onde $L_1 = [0,2; 0,4] \times [0,5; 0,8]$, $L_2 = [0,2; 0,8] \times [0,8; 0,9]$.

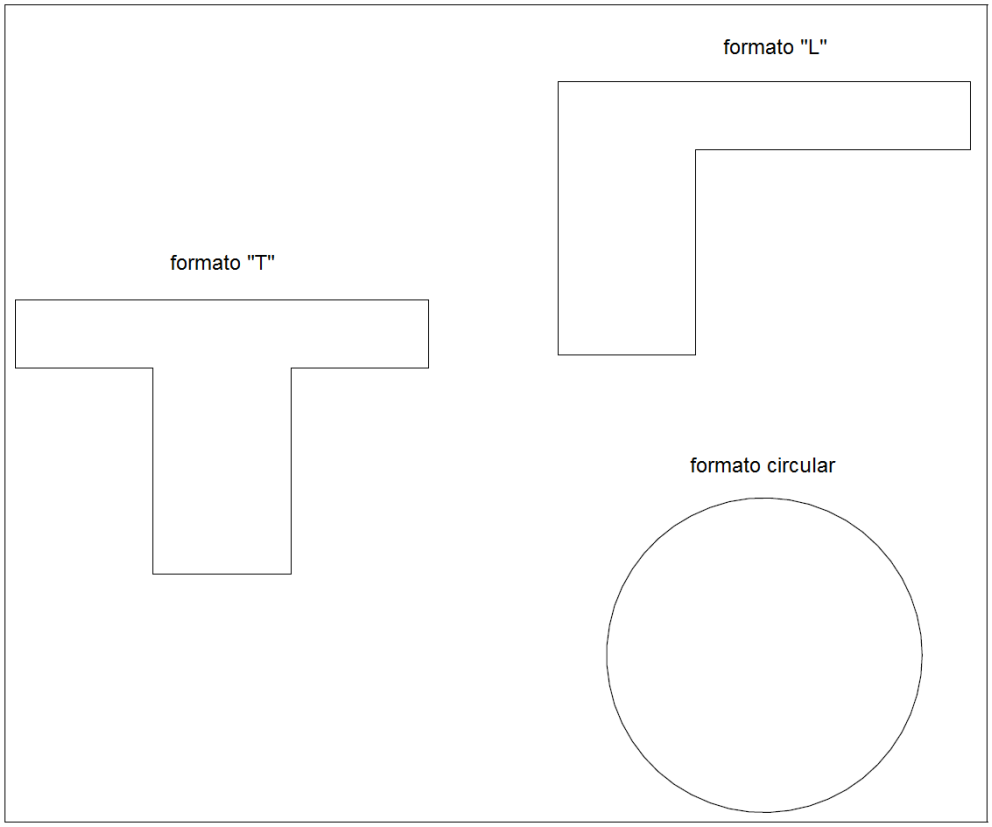


Figura 4.1: Três cluster espaciais artificiais.

Um risco relativo igual a $1,0$ foi definido para todo controle fora do cluster real, e maior que $1,0$ e idêntico em cada controle dentro do cluster. Os riscos

relativos para cada cluster foram definidos de tal forma que, se a localização exata do cluster fosse conhecida de antemão, o poder de detecção seria de $0,999$ (veja Kulldorff *et al.* (2003)).

### 4.1.2 Análises numéricas

Para todos os três cenários criados, as mesmas coordenadas geográficas do conjunto de dados foram utilizadas por todos os algoritmos. Foram executadas 10.000 realizações do algoritmo Monte Carlo sob a hipótese nula, o mesmo para cada um dos três modelos de hipótese alternativa. As medidas de eficiência: poder, sensibilidade e PPV, foram calculadas para o cluster mais verossímil em cada replicação.

A Tabela 4.1 apresenta os resultados. Os valores de poder e VPP do VBScan mostram-se sensivelmente menores, se comparados com o Scan Elíptico, apesar disso o VBScan mostra-se significativamente superior, demonstrando uma maior capacidade de identificação do cluster, quando ele realmente existe.

Tabela 4.1: Comparação de poder, valor preditivo positivo e sensibilidade para os três formatos de clusters.

| | **Poder** | | **Sensibilidade** | | **VPP** | |
|---|---|---|---|---|---|---|
| cluster | Elíptico | VBScan | Elíptico | VBScan | Elíptico | VBScan |
| Circular | 0,8400 | 0,7963 | 0,7257 | 0,8199 | 0,8347 | 0,7871 |
| T | 0,7320 | 0,7067 | 0,5508 | 0,7270 | 0,7837 | 0,7398 |
| L | 0,7206 | 0,6696 | 0,5501 | 0,7144 | 0,7740 | 0,6932 |

Os resultados da aplicação do VBScan em dois conjuntos de dados reais são apresentados e discutidos no próximo capítulo.

## 4.2 MOMC-VBScan

Nesta seção o método MOMC-VBScan é testado em diferentes mapas artificiais. Os mapas artificiais foram alterados no que diz respeito a: (1) homogeneidade da população de risco, (2) número de clusters no mapa artificial em estudo e (3) risco relativo de cada cluster, perfazendo-se, assim, 16 cenários diferentes. Para cada um dos cenários criados, o poder de detecção, e o *matching* foram avaliados.

### 4.2.1 Os mapas artificiais

Um conjunto de dados, do tipo caso-controle, foi criado com uma população total de risco de 2.000 indivíduos, sendo 50 casos. Análogo ao VBScan, os cenários foram criados dentro de um quadrado unitário $[0, 1] \times [0, 1]$. A Figura 4.2 mostra a distribuição das duas populações artificiais criadas. Uma população de risco homogênea, distribuída uniformemente pelo mapa, e a outra heterogênea, onde 750 indivíduos foram distribuídos dentro do quadrado $[0,50; 0,75] \times [0,50; 0,75]$, e os outros 1250 fora deste quadrado, ambos de maneira uniforme.

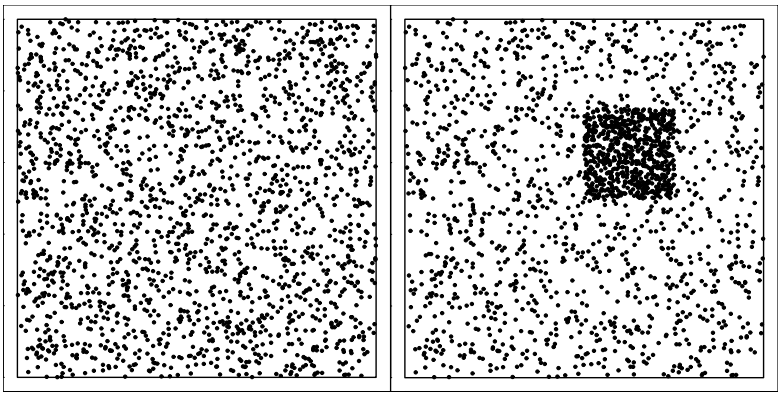

Figura 4.2: Populações homogênea e heterogênea

Para cada tipo de população criada, mapas artificiais contendo diferentes números de clusters, com diferentes riscos relativos, também foram criados. A Figura 4.3 mostra a forma destes cenários, assim como a forma geométrica dos clusters e demarca a região mais populosa no mapa de população heterogênea em linha tracejada:

(A) : um único cluster $[0,4;0,7] \times [0,4;0,7]$;

(B) : um único cluster $[0,0;0,3] \times [0,0;0,3]$;

(C) : dois clusters $[0,0;0,3] \times [0,0;0,3]$ e $[0,6;0,9] \times [0,6;0,9]$;

(D) : três clusters $[0,0;0,3] \times [0,0;0,3]$, $[0,6;0,9] \times [0,6;0,9]$ e $[0,0;0,3] \times [0,7;1,0]$.

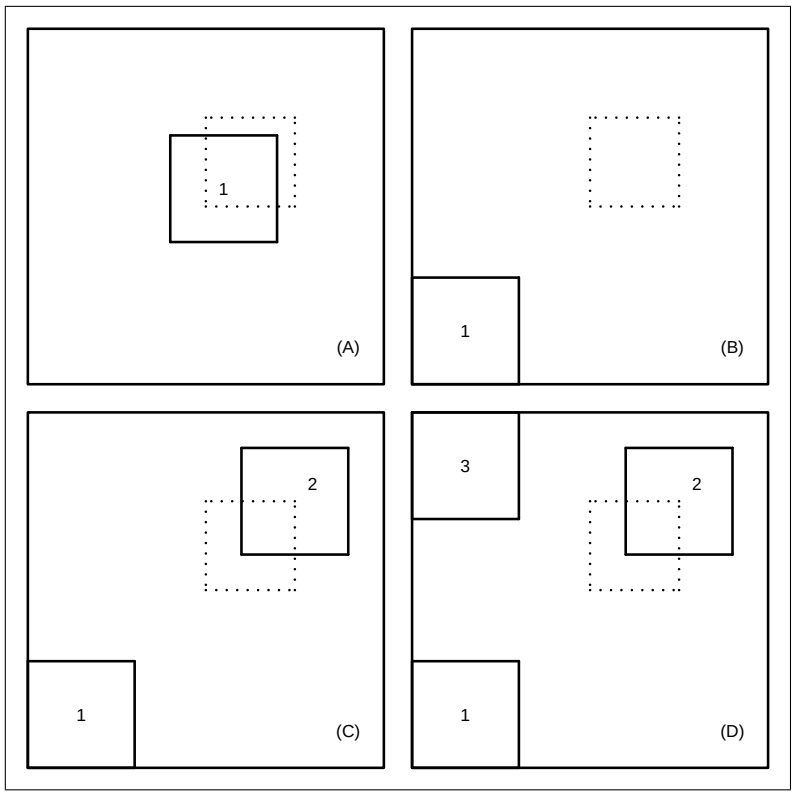


Figura 4.3: Mapas artificias gerados.

O procedimento para cálculo do risco relativo dos clusters dentro dos mapas em estudo foi o mesmo utilizado no VBScan (veja Kulldorff *et al.* (2003)). No caso do cluster ser de baixo risco, adotou-se o valor do inverso do risco relativo encontrado como se o cluster fosse de alto risco. Para todos os controles fora dos clusters, adotou-se um risco relativo igual a $1,0$. A Tabela 4.2 mostra os riscos relativos de cada cluster em todos os mapas artificiais utilizados.

Tabela 4.2: Riscos relativos utilizados em cada uma das componentes de todos os mapas artificias criados.

| Mapa | População | Componente | | | Risco relativo | | |
|---|---|---|---|---|---|---|---|
| A1 | Homogênea | 1 | | | 6,08 | | |
| A2 | Homogênea | 1 | | | 0,17 | | |
| A3 | Heterogênea | 1 | | | 3,90 | | |
| A4 | Heterogênea | 1 | | | 0,26 | | |
| B1 | Heterogênea | 1 | | | 6,94 | | |
| B2 | Heterogênea | 1 | | | 0,14 | | |
| C1 | Homogênea | 1 | 2 | | 5,76 | 6,06 | |
| C2 | Homogênea | 1 | 2 | | 5,76 | 0,17 | |
| C3 | Heterogênea | 1 | 2 | | 6,94 | 4,32 | |
| C4 | Heterogênea | 1 | 2 | | 6,94 | 0,23 | |
| D1 | Homogênea | 1 | 2 | 3 | 5,76 | 6,06 | 5,94 |
| D2 | Homogênea | 1 | 2 | 3 | 5,76 | 0,17 | 5,94 |
| D3 | Homogênea | 1 | 2 | 3 | 5,76 | 6,06 | 0,17 |
| D4 | Heterogênea | 1 | 2 | 3 | 6,94 | 4,32 | 7,41 |
| D5 | Heterogênea | 1 | 2 | 3 | 6,94 | 0,23 | 7,41 |
| D6 | Heterogênea | 1 | 2 | 3 | 6,94 | 4,32 | 0,13 |

### 4.2.2 Análises numéricas

Dada as populações, homogênea ou heterogênea, todos os cenários utilizaram as mesmas coordenadas geográficas do conjunto de dados. De forma análoga ao VBScan, 10.000 realizações do algoritmo Monte Carlo, sob a hipótese nula, foram realizadas, o mesmo para cada um dos mapas de hipótese alternativa. As medidas de poder e *matching* para cada cenário mais verossímil foram calculadas para cada cenário. A Tabela 4.3 apresenta os resultados.

Tabela 4.3: Resultados de poder e *matching* para cada um dos mapas artificias criados.

| Mapa | Poder | *Matching* | Mapa | Poder | *Matching* |
|---|---|---|---|---|---|
| A1 | 0,8298 | 0,8424 | C3 | 0,8814 | 0,5972 |
| A2 | 0,2023 | 0,9090 | C4 | 0,9704 | 0,8506 |
| A3 | 0,6933 | 0,6290 | D1 | 0,8848 | 0,4981 |
| A4 | 0,3236 | 0,8388 | D2 | 0,9489 | 0,6391 |
| B1 | 0,8873 | 0,9021 | D3 | 0,9418 | 0,6374 |
| B2 | 0,2112 | 0,9125 | D4 | 0,8796 | 0,5019 |
| C1 | 0,9105 | 0,6508 | D5 | 0,9859 | 0,7016 |
| C2 | 0,9249 | 0,8370 | D6 | 0,9123 | 0,5883 |

Nos mapas em que não há presença de cluster de alto risco, mapas $A2$, $A4$ e $B2$, o poder apresenta-se com valores inferiores. Como nestes mapas só há um cluster, três fenômenos podem estar ocorrendo: (1) ao gerar o cenário sob a hipótese alternativa, com risco relativo inferior, pode ocorrer que na região delimitada para o cluster não haja casos. Desta forma, todos os casos estão sendo distribuídos na região em que o risco relativo é $1,0$. Assim, essa região do cluster será agregada à região branca. Isto provoca um efeito tal

que o algoritmo dificilmente encontra uma partição significativa no mapa; (2) pode ocorrer também que, mesmo que na região delimitada para o cluster haja algum caso, estes podem estar localizados na fronteira com a região de risco relativo $1,0$. Isto também provoca uma dificuldade de detecção para o algoritmo, pois estes poucos casos estariam se juntando aos casos da região de risco $1,0$, novamente fazendo com que o poder de detecção seja inferior; e (3) a atribuição de risco relativo no procedimento de simulação foi feita de forma experimental, pois ainda não existem resultados definitivos para esta atribuição. Dado estes argumentos, os valores de *matching* encontrados para os mapas $A2$, $A4$ e $B2$ não retratam uma real sensibilidade do algoritmo para detecção de clusters de baixo risco.

Considere agora os mapas com dois clusters, em particular, os mapas $C1$ e $C2$, ambos com população homogênea. No mapa $C1$, geramos dois clusters de alto risco, resultando num poder de $0,9105$ e um *matching* de $0,6508$. Nota-se que o poder e o *matching* encontrados para o mapa $C2$, composto por um cluster de alto risco e um de baixo risco, são superiores ao encontrado no mapa $C1$. Novamente, surge a dificuldade do algoritmo em lidar com partições que possuem, pelo menos, uma componente de baixo risco: (1) se na componente de baixo risco não há caso, o algoritmo tende a agregar essa região à região branca, fora do domínio dos casos, que faz com que a detecção de partição tenda a encontrar partições com apenas uma componente de alto risco; e (2) se houver algum caso na componente de baixo risco, e se ele estiver perto da sua borda, o algoritmo pode agregá-lo à componente de alto risco. A redução no valor do poder mostra como o algoritmo perde a capacidade de detectação de partições com maior número de componentes. Comportamenteo análogo ocorre com o valor de *matching*. Se considerarmos os mapas $C3$ e $C4$, com população heterogênea, percebemos comportamento

semelhante ao que foi descrito para os mapas com população homogênea. Isso mostra que o algoritmo tende a ser estável no que diz respeito às populações de fundo.

No que tange aos mapas com três clusters, se um destes clusters é de baixo risco, observa-se a mesma dificuladade do algoritmo em identificar essa região de baixo risco. Com relação ao *matching*, observam-se valores inferiores quando a partição possui maior número de clusters. Os valores de matching vão reduzindo de ordem de grandeza à medida em que se aumenta o número de componentes. Isso mostra uma redução na "sensibilidade" do algoritmo em presença de um maior número de anomalias.

Os resultados da aplicação do MOMC-VBScan em dois conjuntos de dados reais são apresentados e discutidos no próximo capítulo.

# Capítulo 5

# Aplicações em dados Reais

Os dois métodos propostos, o VBScan e o MOMC-VBScan, são aplicados em dois conjuntos de dados reais.

Um dos conjuntos de dados é o já bem conhecido conjunto de dados de câncer de laringe e câncer de pulmão de Chorley-Ribble em Lancashire-UK, de 1973 a 1984 (Diggle (1990)). Neste conjunto de dados são apresentados 917 casos de câncer de pulmão e 57 casos de câncer de laringe e a residência de cada doente é conhecida (veja http://cran.r-project.org/web/ packages/splancs/splancs.pdf - pag. 55). Neste trabalho foi investigada a suspeita de que uma incineradora localizada na região poderia ser a causa dos casos de câncer de laringe. Os casos de câncer de pulmão, que parecem ser menos dependentes de fatores ambientais, são denotados CONTROLES, enquanto os casos de câncer de laringe são denotados CASOS.

A Figura 5.1 mostra a distribuição da população de risco (pontos) e dos casos (círculos) para os dados de câncer de conjunto de dados de Lancashire.

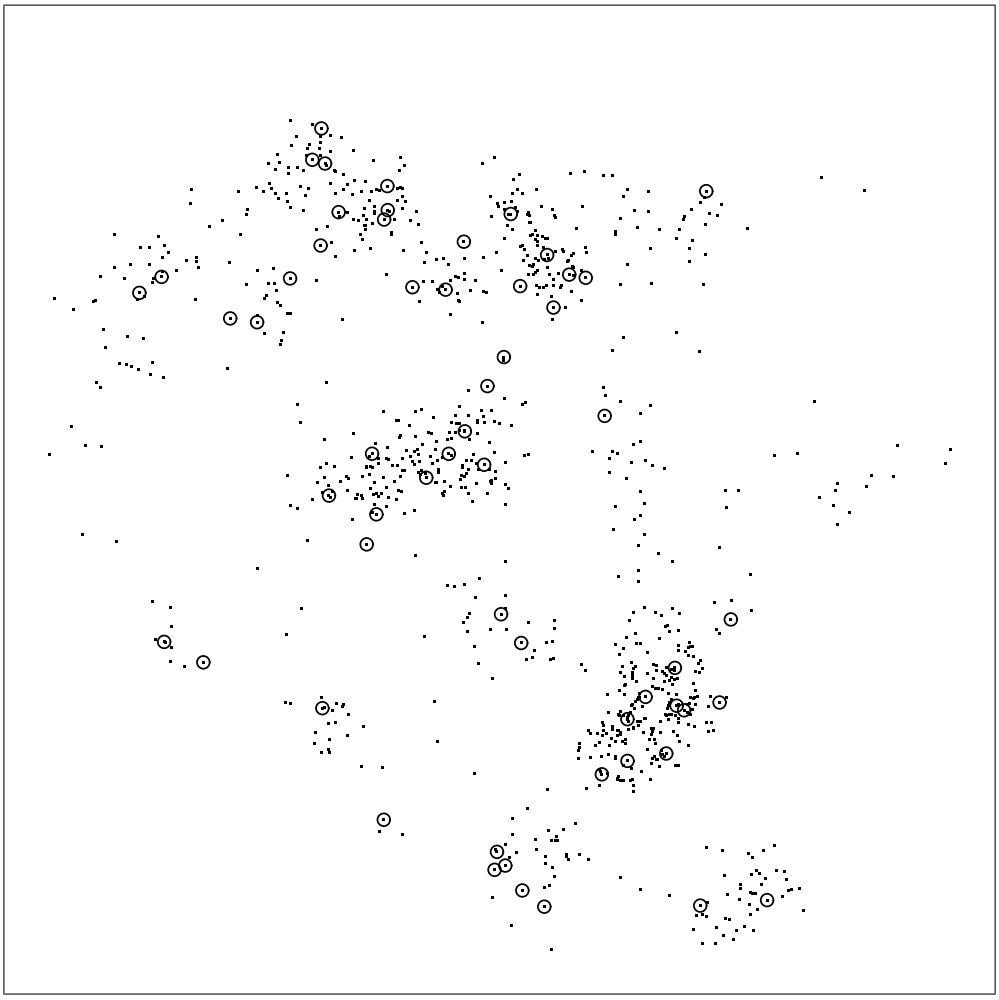

Figura 5.1: Distribuição espacial dos casos observados de câncer de pulmão (pontos) e de câncer de laringe (círculos), na cidade de Lancashire-UK.

O segundo conjunto de dados é o relativo a casos de dengue numa pequena cidade no sudeste brasileiro, Lassance, MG. Num período de 6 meses em 2010, entre janeiro e junho, um total de 57 casos foram registrados, de uma população total de 3986 indivíduos na população de risco. A distribuição espacial da população de risco (pontos) e dos casos (círculos) para os dados de dengue de Lassance estão na Figura 5.2. Este conjunto de dados está completamente descrito em Duczmal *et al.* (2011).

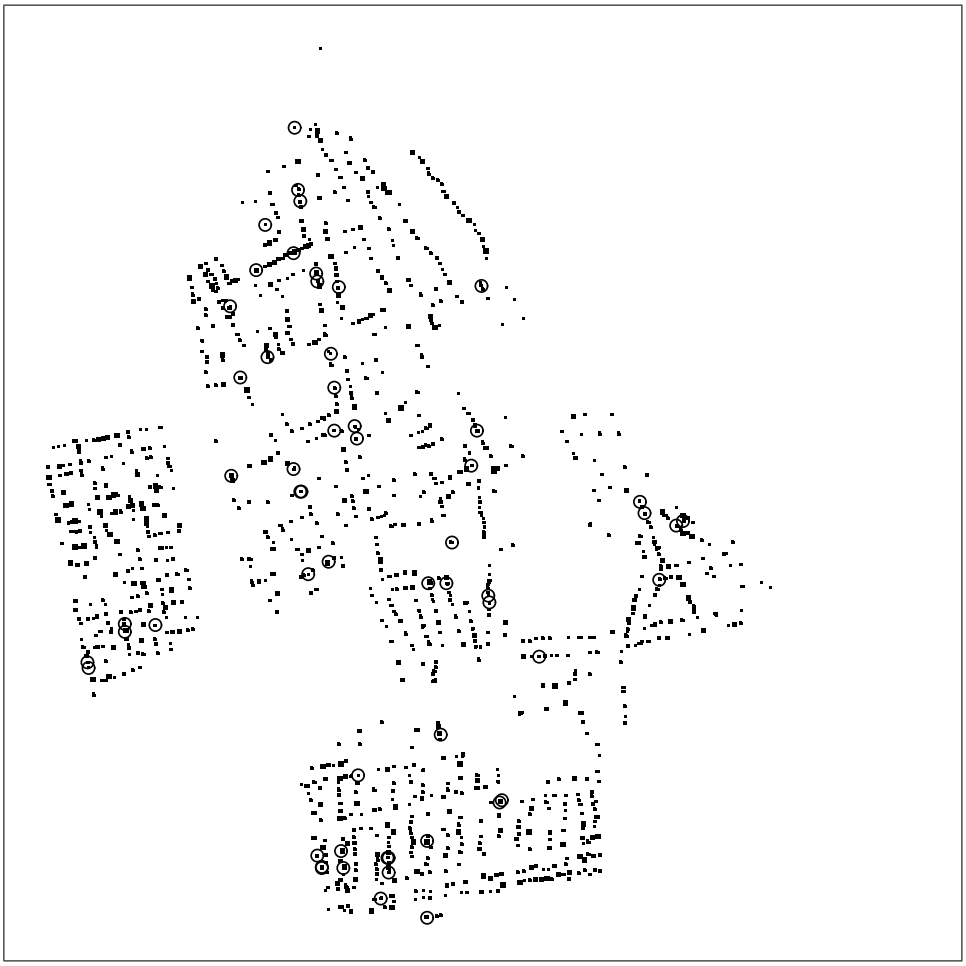

Figura 5.2: Distribuição espacial dos casos observados de dengue (círculos) e dos controles (pontos), na cidade de Lassance-BR.

Na próxima seção, os resultados da aplicação do método VBScan nos dois conjuntos de dados reais são apresentados e analisados. Os procedimentos foram executados utilizando um processador Dual Core 2.1GHz, 4Gb de memória RAM e sistema operacional windows 7. Os códigos foram implementados em linguagem C++ e se encontram disponíveis para fins de pesquisa para quaisquer interessados.

## 5.1 VBScan

Os resultados encontrados para a estatística scan baseada na distância de Voronoi foi comparada com a estatística scan elíptica.

### 5.1.1 Dados de Lancashire

Na Figura 5.3, é mostrada a Árvore Geradora Mínima de Voronoi, para os dados de Lancashire, com respectivos pesos das arestas, com o diagrama de Voronoi como plano de fundo. A estatística espacial scan elíptica é também executada para comparação.

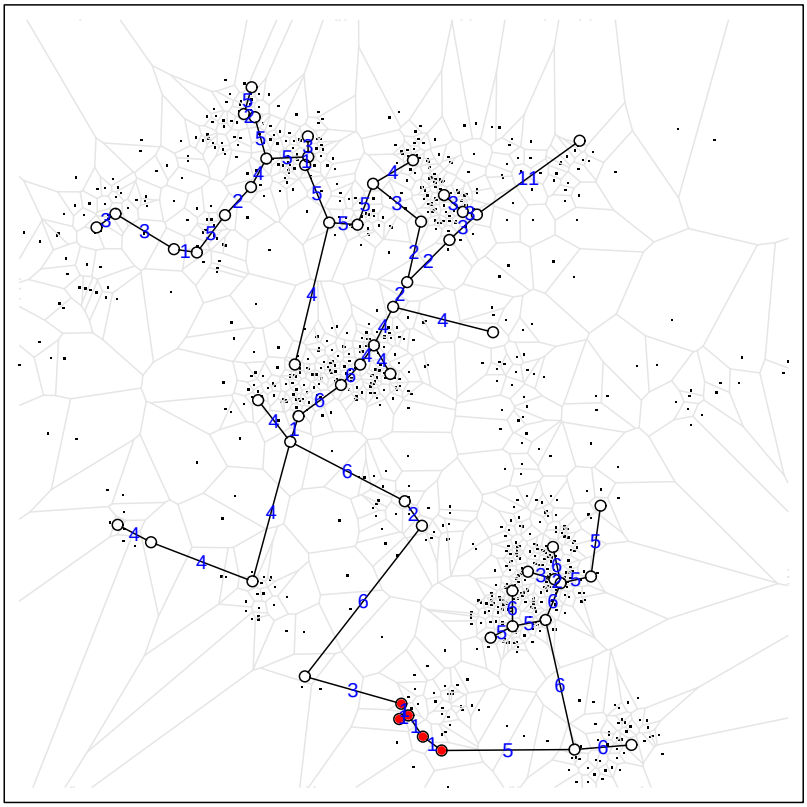

Figura 5.3: Árvore geradora mínima de Voronoi, com cluster mais verossímil encontrado pelo VBScan e pelo Scan Elíptico, para os dados de Lancashire.

Os valores $p$ associados às duas estatísticas scan são calculadas baseadas em 9.999 simulações de Monte Carlo sob a hipótese nula. O cluster mais verossímil encontrado em ambas as técnicas são idênticos, consistindo de cinco casos (círculos preenchidos) da Figura 5.3, que se encontram próximos da indústria incineradora localizada na região.

A Tabela 5.1 mostra os valores de verossimilhança ($LLR$), número de casos, valores $p$ e tempo de execução para ambas estatísticas. O conjunto dos possíveis clusters elípticos forma um espaço mais restritivo de configurações do que o conjunto de clusters de forma irregular. Portanto não é surpresa que o valor $p$ seja menor em relação ao p-valor do VBScan, já que o cluster mais verossímil se ajusta muito bem em uma elipse alongada. Nota-se que o método baseado na distância de Voronoi, requer menos tempo computacional para dados pontuais, se comparado com Scan Elíptico, como apresentado. Percebe-se ainda que os valores de $LLR$ obtidos por cada método são distintos, apesar de corresponderem à mesma solução. Isto se deve às diferentes estratégias de estimativa populacional empregadas.

Tabela 5.1: Comparação da detecção de *clusters* espaciais para os dados de Lancashire, resultados encontrados para os métodos scan elíptico e VBScan.

| Método | LLR | Nº Casos | valor $p$ | CPU(seg.) |
|---|---|---|---|---|
| Scan Elíptico | $14,4049$ | 5 | $0,0089$ | $896,0$ |
| VBScan | $10,8357$ | 5 | $0,0470$ | $449,5$ |

### 5.1.2 Dados de Lassance

De maneira análoga aos dados de Lancashire, na Figura 5.4 mostra-se a AGMV para os dados de Lassance, com respectivos pesos das arestas, com

o diagrama de Voronoi como plano de fundo. Para detectar os possíveis clusters, o método VBScan foi aplicado. Os dois clusters mais verossímeis apresentaram 10 e 9 casos, respectivamente para os clusters primário e secundário, como mostra a Figura 5.4.

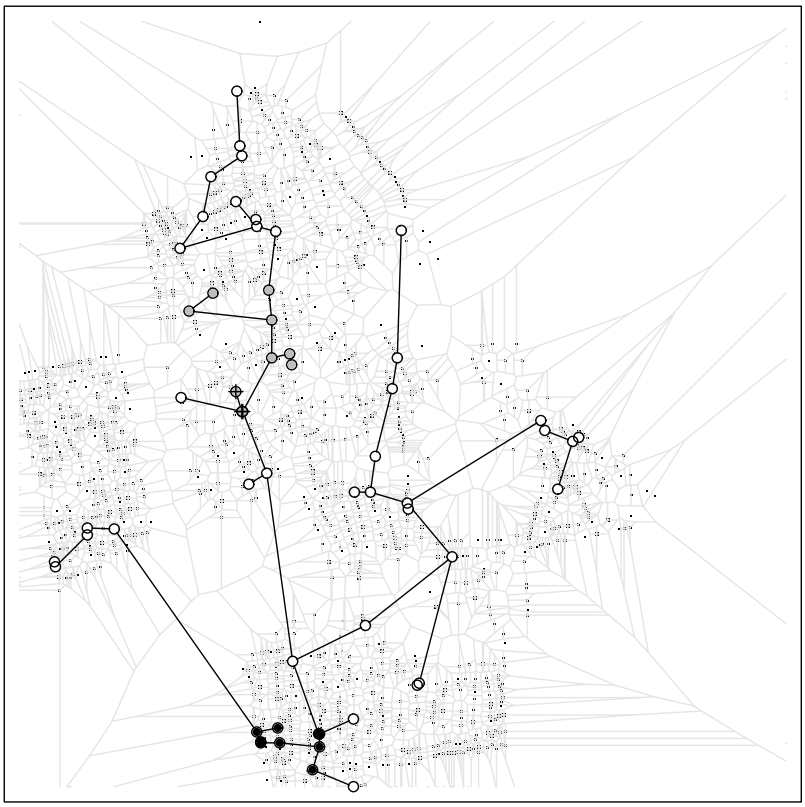

Figura 5.4: Árvore geradora mínima de Voronoi, com cluster mais verossímil encontrado pelo VBScan e pelo Scan Elíptico, para os dados de Lassance. Pontos em cinza correspondem ao cluster primário, enquanto que os pontos pretos, ao cluster secundário encontrado pelo VBScan. Os pontos de cor cinza marcados com uma cruz correspodem ao cluster primário encontrado pelo Scan Elíptico.

Para o cluster primário um valor $p$ igual a $0,004$ foi encontrado (veja Tabela 5.2). A Tabela 5.2 mostra que o cluster secundário também é estatisticamente significante. Os valores $p$ são calculados através de 999 simulações de Monte Carlo sob a hipótese nula. Assim, concluimos que existe evidência significativa de regiões de alto risco de dengue na área urbana da cidade de Lassance.

Tabela 5.2: Resultados encontrados para detecção de *clusters* espaciais para os dados de Lassance utilizando o método VBScan.

| *Clusters* | LLR | Nº Casos | valor $p$ |
|---|---|---|---|
| Primário | $17,5686$ | 10 | $0,0040$ |
| Secundário | $15,2390$ | 09 | $0,0160$ |

Empregando a versão elíptica da estatística scan, também com 999 simulações Monte Carlo, o cluster mais verossímel encontrado tem somente 3 casos, contido dentro do cluster primário encontrado pelo VBScan, veja na Figura 5.4 (valor $p = 0,054$). Este interessante resultado acontece devido a características peculiar desse problema:(1) a população não segue uma distribuição aleatória espacial. Ao invés disto os indivíduos são mais ou menos alinhados de acordo com a geometria das ruas e (2) a estrutura de vizinhança induzida pela métrica da distância Euclidiana, que é utilizada pelo scan elíptico, é muito diferente da estrutura de vizinhança imposta pela distância de Voronoi.

O tempo de execução das 999 replicações de Monte Carlo para o conjunto de dados da Dengue foi cerca de 187 segundos para o VBScan e 764 segundos para a estatística scan elíptica, confirmando o ganho computacional já observado anteriormente.

## 5.2 MOMC-VBScan

Os resultados encontrados para o método MOMC-VBScan são apresentados e discutidos nesta seção.

### 5.2.1 Dados de Lancashire

A Figura 5.5 mostra as partições relativas ao conjunto Pareto-ótimo encontrado para o conjunto de dados de Lancashire-UK.

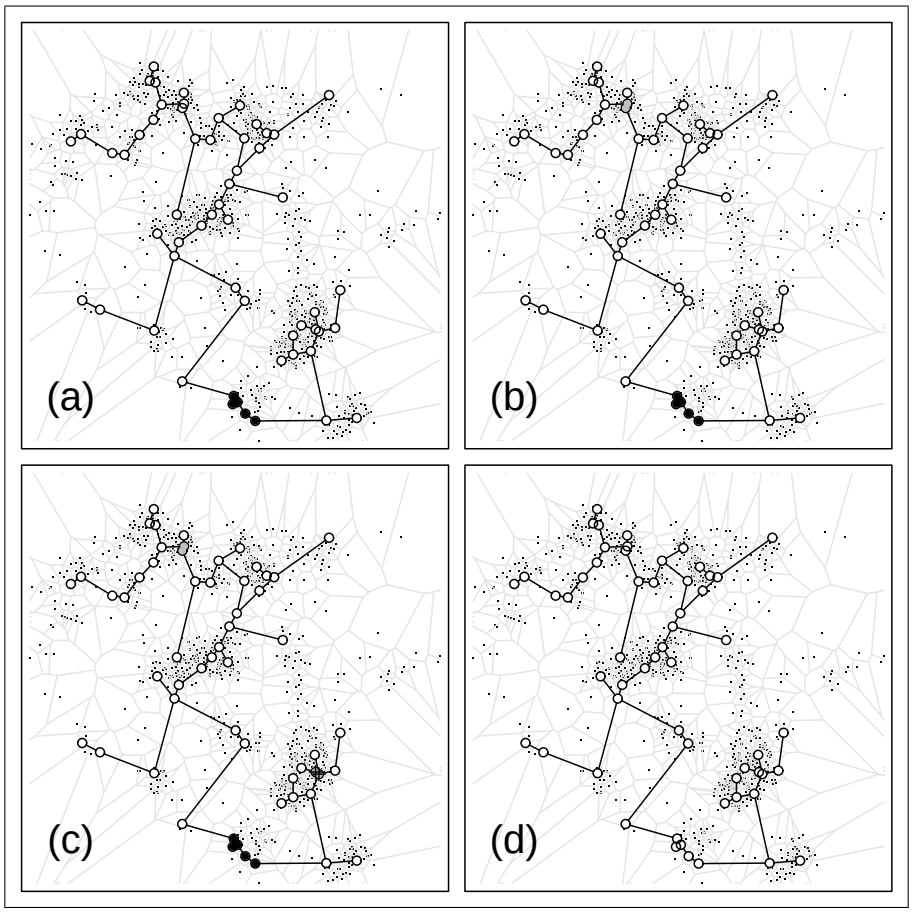


Figura 5.5: Soluções Pareto-ótimas para o conjunto de Lancashire.

A partição mais verossímil é composta por três componentes: (1) uma componente corresponde à região desprovida de casos (região branca), (2) a segunda componente corresponde ao cluster mais verossímil, e (3) a terceira componente corresponde a todos os outros casos e controles do mapa em estudo. Esta solução está em consonância com o resultado encontrado pelo VBScan. O conjunto Pareto-ótimo é composto por quatro soluções factíveis.

Na Figura 5.6 são apresentadas as Superfícies de Aproveitamenteo (SA) para cada solução do conjunto Pareto-ótimo.

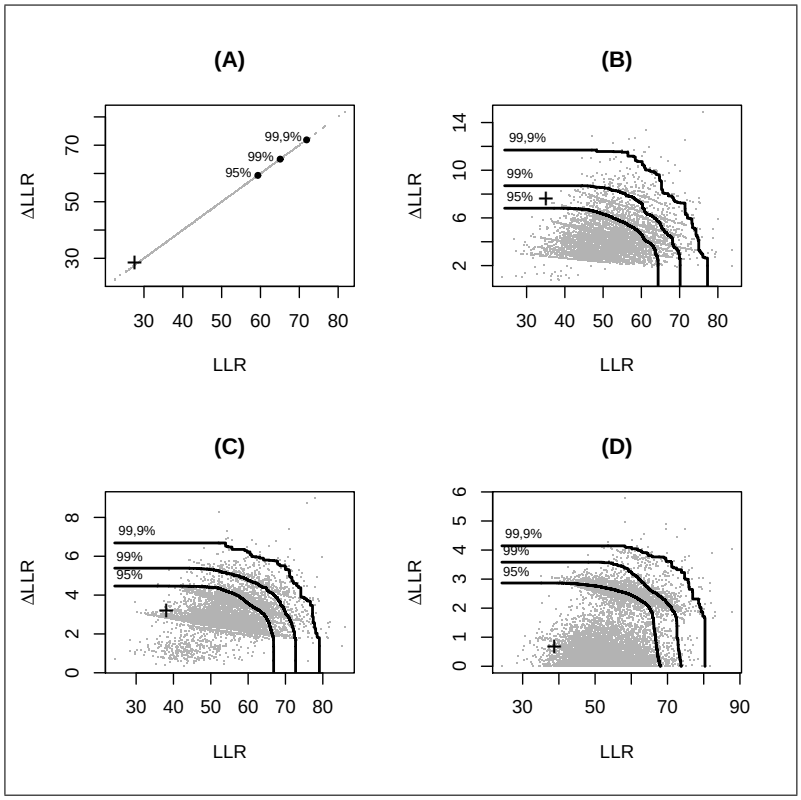


Figura 5.6: Superfícies de aproveitamento de 95%, 99% e 99, 9% , com respectiva solução Pareto-ótima (+), para os dados de Lancashire. Separadas por número de componentes: 2 (A), 3 (B), 4 (C) e 5 (D).

Na Tabela 5.3 estão os valores das funções-objetivo, $LLR$ e $\Delta LLR$, bem como o número de componentes de cada solução do conjunto Pareto-ótimo, assim como os respectivos valores $p$ e tempo de CPU(seg.). A solução Pareto-ótima significativa encontrada corresponde à Figura 5.5 (a).

Tabela 5.3: Resultados da detecção da melhor partição para os dados de Lancashire.

| LLR | $\Delta$ LLR | Nº Componentes | valor $p$ | CPU(seg.) |
|---|---|---|---|---|
| 35, 046002 | 7, 430548 | 3 | 0, 0350 | 29 |
| 38, 127952 | 3, 081950 | 4 | 0, 3210 | |
| 38, 712310 | 0, 584357 | 5 | 0, 6260 | |
| 27, 615454 | 27, 615454 | 2 | 0, 9980 | |

### 5.2.2 Dados de Lassance

A Figura 5.7 mostra as partições relativas ao conjunto Pareto-ótimo encontrado para o conjunto de dados de Lassance. Neste conjunto de dados, a partição mais verossímil é composta por quatro componentes: (1) uma componente corresponde à região desprovida de casos (região branca), (2) a segunda componente composta por 10 casos (pontos pretos), (3) a terceira componente composta por 9 casos (pontos cinza), e (4) todos os outros casos e controles do mapa em estudo. Esta solução também se encontra em consonância com o resultado encontrado pelo VBScan. As duas componentes com 10 e 9 casos, correspondem aos clusters primário e secundário encontrados pelo VBScan. O conjunto Pareto-ótimo é composto por quatro soluções factíveis, sendo três significativas.

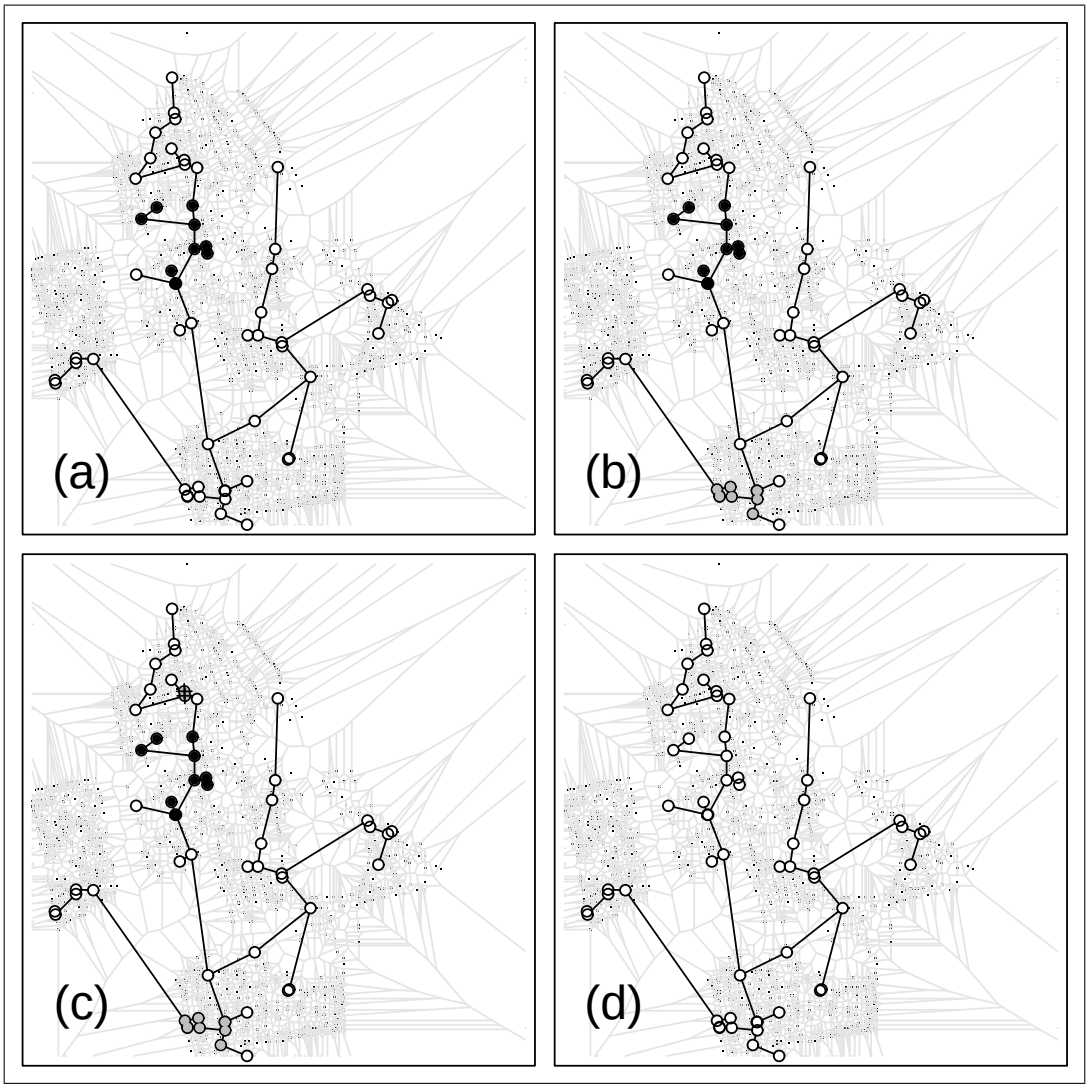


Figura 5.7: Soluções Pareto-ótimo para o conjunto de Lassance.

Na Figura 5.8 são apresentadas as Superfícies de Aproveitamenteo (SA) para cada solução do conjunto Pareto-ótimo.

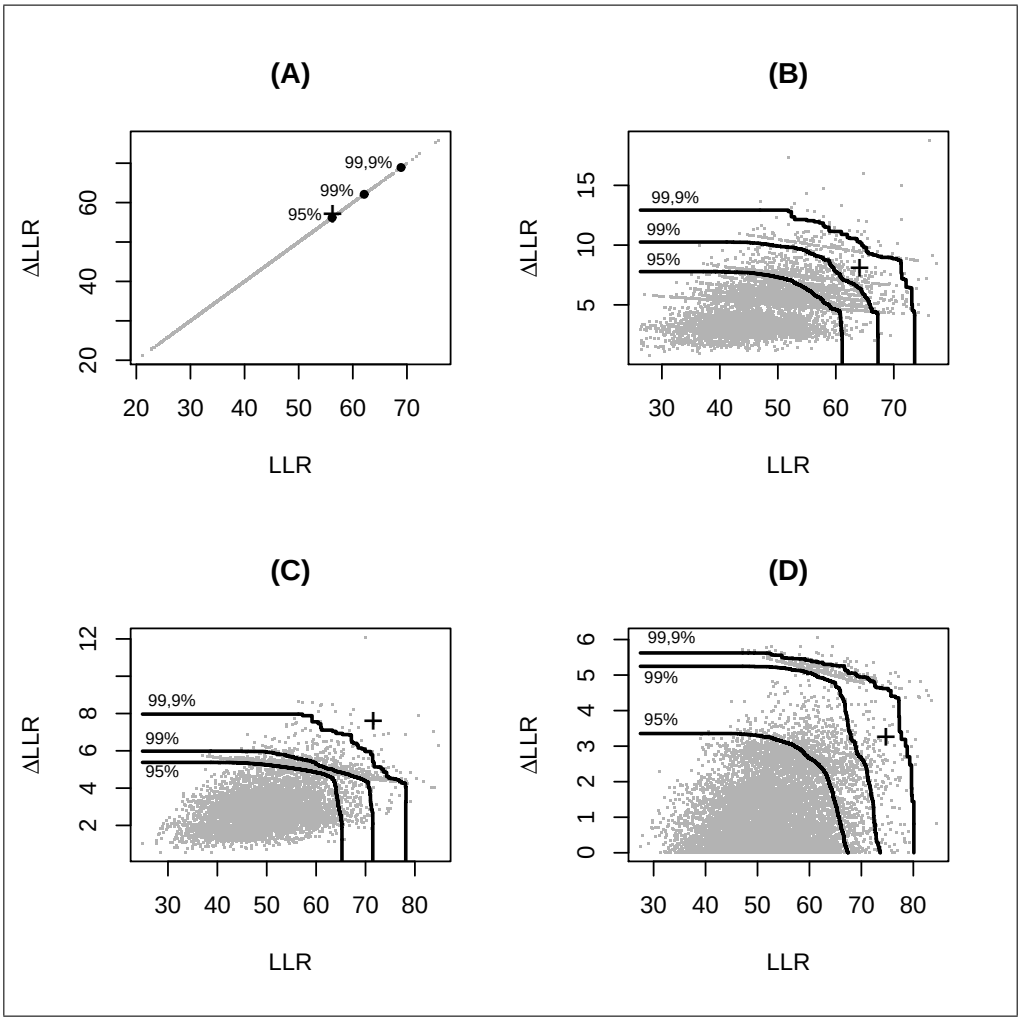


Figura 5.8: Superfícies de aproveitamento de 95%, 99% e 99, 9%, com respectiva solução Pareto-ótima (+), para os dados de Lassance. Separadas por número de componentes: 2 (A), 3 (B), 4 (C) e 5 (D).

Na Tabela 5.4 estão os valores das funções-objetivo, $LLR$ e $\Delta LLR$, o número de componentes de cada solução do conjunto Pareto-ótimo, assim como os respectivos valores $p$ e tempo de CPU. As soluções Pareto-ótimas significativas encontradas correspondem, respectivamente às Figuras 5.7 (b), (c) e (a). O maior número de soluções significativas se deve, provavelmente,

ao fato de que a população de risco não está distribuída de maneira aleatória pelo espaço.

Tabela 5.4: Resultados da detecção da melhor partição para os dados de Lassance.

| LLR | $\Delta$ LLR | N$^{\text{O}}$ Componentes | valor $p$ | CPU(seg.) |
|---|---|---|---|---|
| $71,558025$ | $7,449102$ | 4 | $0,0010$ | 1259 |
| $74,741006$ | $3,182981$ | 5 | $0,0030$ | |
| $64,108923$ | $7,828637$ | 3 | $0,0060$ | |
| $56,280287$ | $56,280287$ | 2 | $0,0590$ | |

# Capítulo 6

# Conclusões

## 6.1 Considerações Finais

Um novo algoritmo para detecção e inferência de clusters espaciais foi desenvolvido e testado, o *Voronoi Based Scan (VBScan)*. O conceito de Árvore Geradora Mínima foi adaptado para a nova distância de Voronoi, que é utilizada para detectar os clusters em potencial. Este conjunto é avaliado utilizando a estatística espacial scan.

A classe de problemas considerada aqui assume que os conjuntos de dados sejam do tipo caso-controle, com um domínio no espaço $\mathbb{R}^2$. Os clusters são modelados no espaço de coordenadas como um grafo conectado com estrutura de árvore, ligando os casos.

A distância de Voronoi entre dois pontos quaisquer pode ser interpretada como uma aproximação da integral de linha para função de densidade populacional sobre o seguimento que une os dois pontos. Por esta razão, a AGM de Voronoi é uma extensão natural da AGM Euclidiana, levando em conta a heterogeneidade da densidade populacional. Esta distância também é utilizada para estimar o número de controles individuais sob a região de

influência de cada um dos casos. Isso permite definir uma população associada a cada cluster em potencial, podendo assim, o cluster ser avaliado pela estatística scan espacial.

Os resultados das simulações numéricas mostram que o algoritmo proposto, VBScan, tem maior sensibilidade e maior velocidade computacional do que o scan elíptico. Sendo competitivo no que diz respeito ao poder e VPP.

O VBScan também inclui informação topológica da estrutura de vizinhança, além da informação geométrica. Por esta razão, ele se mostra mais robusto que um método puramente geométrico, como o scan elíptico. Estas vantagens foram ilustradas com a aplicação do método no conjunto de dados de dengue, em que a população em risco se distribui ao longo de um *grid* de linhas retas, de acordo com a configuração das ruas de Lassance. Isto faz com que o VBScan seja recomendável em estudos de ambientes urbanos.

No que tange ao MOMC-VBScan, uma mudança na proposta de estimação da população associada a cada caso, usada no VBScan, fez com que a região de influência de cada caso fosse visualizada. O método se mostrou rápido e com boa precisão na determinação das partições.

Os resultados encontrados na aplicação do método nos conjuntos de dados de casos de câncer e de dengue mostram que o método está em boa concordância com a análise prévia do VBScan. No estudo do conjunto de dados de Lancashire, a partição mais significativa detectada é composta por exatamente o mesmo cluster mais verossímil encontrado pelo VBScan, uma região branca e outra região composta por todos os outros casos e controles que não pertencem nem ao cluster mais verossímil, nem à região branca. No conjunto de dados de dengue (Lassance) a partição mais significativa tem como componentes os clusters primário e secundário encontrados no VBS-

can, uma região branca e outra região composta por todos os outros casos e controles que não pertencem nem aos clusters primário e secundário, nem à região branca.

## 6.2 Propostas de continuidade

Tópicos interesssantes para trabalhos futuros nesta área incluem:

1. Incluir co-variáveis na estimação das densidades populacionais para a detecção do cluster, seja pela estatística frequentista, seja pela bayesiana;

2. Mudar a métrica na obtenção do diagrama de Voronoi (incluindo nesse momento as co-variáveis);

3. Incluir a idéia da plausibilidade de um ponto pertencer a um cluster ou a uma componente, através de técnicas como a função intensidade (Oliveira *et al.* (2011));

4. Propor alternativas de medidas de eficiência para o algoritmo MOMC-VBScan.

5. Propor uma alternativa que seja mais eficiente na geração de clusters de baixo risco, para que se possa, efetivamente, testar a eficiência do algoritmo na presença de clusters de baixo risco;

6. Criar um *software* com interface amigável para os usuários, que implemente os algoritmos desenvolvidos.

# Referências Bibliográficas


Abrams, A. M., Kleinman, K., & Kulldorff, M. 2010. Gumbel based p-Value approximations for spatial scan statistics. *International Journal of Health Geographics*, **9**, 61.

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

Assunção, R. M., Costa, M. A., Tavares, A., & Ferreira, S. J. 2006. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, **25**, 723–742.

Balakrishnan, N., & Koutras, M. V. 2002. *Runs and Scans with Applications*. London: John Wiley & Sons.

Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R., & Moore, A. W. 2005. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, **38**, 99–113.

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

Conley, J., Gahegan, M., & Macgill, J. 2005. A Genetic Approach to Detecting Clusters in Point Data Sets. *Geographical Analysis*, **37**, 286–314.

da Fonseca, V. G., Fonseca, C. M., & Hall, A. O. 2001. Inferential Performance Assessment of Stochastic Optimisers and the Attainment Function. *Pages 213–225 of: Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization*, vol. 1993. Lecture Notes In Computer Science, Berlin: Springer-Verlag.

Dematteï, C., Molinari, N., & Daurès, J. P. 2007. Arbitrarily shaped multiple spatial cluster detection for case event data. *Computational Statistics & Data Analysis*, **51**, 3931–3945.

Diggle, P. J. 1990. A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point. *Journal of the Royal Statistical Society*, **153(3)**, 349–362.

Duarte, A. R., Duczmal, L., Ferreira, S. J., & Cançado, A. L. F. 2010. Internal cohesion and geometric shape of spatial clusters. *Environmental and Ecological Statistics*, **17**, 203–229.

Duczmal, L., & Assunção, R. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, **45**, 269–286.

Duczmal, L., & Buckeridge, D. L. 2006. A Workflow Spatial Scan Statistic. *Statistics in Medicine*, **25**, 743–754.

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

Duczmal, L., Cançado, A. L. F., Takahashi, R. H. C., & Bessegato, L. F. 2007. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, **52**, 43–52.

Duczmal, L., Cançado, A. L. F., & Takahashi, R. H. C. 2008. Geographic Delineation of Disease Clusters through multi-objective Optimization. *Journal of Computational & Graphical Statistics*, **17**, 243–262.

Duczmal, L., Duarte, A. R., & Tavares, R. 2009. Extensions of the scan statistic for the detection and inference of spatial clusters. *Pages 157–182 of:* Balakrishnan, N, & Glaz, J (eds), *Scan Statistics*. Boston, Basel and Berlin: Birkhäuser.

Duczmal, L. H., Moreira, G. J. P., Burgarelli, D., Takahashi, R. H. C., Magalhães, F. C. O., & Bodevan, E. C. 2011. Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. *International Journal of Health Geographics*, **10**, 29.

Dwass, M. 1957. Modified Randomization Tests for Nonparametric Hypotheses. *Annals of Mathematical Statistics*, **28**, 181–187.

Elliott, P., Martuzzi, M., & Shaddick, G. 1995. Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research*, **4:2**, 137–159.

Fonseca, C. M., da Fonseca, V. G., & Paquete, L. 2005. Exploring the Performance of Stochastic Multiobjective Optimisers with the Second-Order Attainment Function. *Pages 250–264 of: Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization*, vol. 3410. Lecture Notes In Computer Science, Berlin: Springer-Verlag.

Glaz, J., Naus, J., & Wallestein, S. 2001. Disease Mapping and Risk Assessment for Public Health. *In: Springer Series in Statistics*. Berlin Heidelberg New York: Springer.

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

Kulldorff, M. 1999. Spatial Scan Statistics: Models, Calculations, and Applications. *Pages 303–322 of:* Balakrishnan, N, & Glaz, J (eds), *Scan Statistics and Applications*. Birkhäuser.

Kulldorff, M., & Nagarwalla, N. 1995. Spatial disease clusters: detection and inference. *Statistics in Medicine*, **14**, 799–810.

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

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F. 2005. A Space Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS Med*, **2**(3).

Kulldorff, M., Huang, L., Pickle, L., & Duczmal, L. 2006. An elliptic spatial scan statistic. *Statistics in Medicine*, **25**, 3929–3943.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, K., Kleinman, K., & Platt, R. 2007. Multivariate Scan Statistics for Disease Surveillance. *Statistics in Medicine*, **26**, 1824–1833.

Lawson, A. 2001. Statistical methods in spatial epidemiology. *Pages 197–206 of:* Lawson, A (ed), *Large scale: surveillance*. London: Wiley.

Lawson, A., & Kulldorff, M. 1999. A review of cluster detection methods. *Pages 99–110 of:* Lawson, A (ed), *Disease Mapping and Risk Assessment for Public Health*. London: John Wiley and Sons.

Lawson, A., Biggeri, A., BVohning, D., Lesare, E., Viel, J. F., & Bertollini, R. 1999. *Disease Mapping and Risk Assessment for Public Health*. London: Wiley.

Li, X-Z., Wang, J-F., Yang, W-Z., Li, Z-J., & Lai, S-J. 2011. A Spatial Scan Statistic for Multiple Clusters. *Mathematical Biosciences*, **233**, 135–142.

Modarres, R., & Patil, G. P. 2007. Hotspot detection with bivariate data. *Journal of Statistical planning and inference*, **137**, 3643–3654.

Moore, D. A., & Carpenter, T. E. 1999. Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiologic Reviews*, **21**, 143–161.

Moura, F. R., Duczmal, L., Tavares, R., & Takahashi, R. H. C. 2007. Exploring Multi-cluster structures with the multi-objective Circular Scan. *Advances in Disease Surveillance*, **2**, 48.

Neill, D. B. 2008. Fast and flexible outbreak detection by linear-time subset scanning [abstract]. *Advances in Disease Surveillance*, **5**, 48.

Neill, D. B. 2009. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, **8**, 20.

Oliveira, F. L. P., Duczmal, L. H., F., Cançado A. L., & R., Tavares. 2011. Nonparametric intensity bounds for the delineation of spatial clusters. *International Journal of Health Geographics*, **10:1**.

Patil, G. P., & Taillie, C. 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, **11**, 183–197.

Sahajpal, R., Ramaraju, G. V., & Bhatt, V. 2004. Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. *In: International Conference on Intelligent Sensing and Information Processing.*

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

Waller, L. A., & Jacquez, G. M. 1995. Disease models implicit in statistical tests of disease clustering. *Epidemiology*, **6:6**, 584–590.

Wieland, S. C., Brownstein, J. S., Berger, B., & Mandl, K. D. 2007. Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes. *PNAS*, **104(22)**, 904–909.

Yiannakoulias, N., Rosychuk, R. J., & Hodgson, J. 2007. Adaptations for finding irregularly shaped disease clusters. *International Journal of Health Geographics*, **6**, 28.

Zhang, Z., Assuncao, R., & Kulldorf, M. 2010. Spatial Scan Statistics Adjusted for Multiple Clusters. *Journal of Probability and Statistics*, 11.