Angélica Ferreira Carvalho

Detecção de clusters irregulares para dados pontuais através da Não-conectividade Ponderada de Grafos

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

Orientador: Luiz Henrique Duczmal
Co-orientador: Anderson Ribeiro Duarte

Universidade Federal de Minas Gerais
Belo Horizonte, Fevereiro de 2011

Dedicatória

Dedico essa vitória ao meu querido esposo Antônio e à nossa querida filha Laís. Principais incentivadores de minha constante busca pelo meu aperfeiçoamento pessoal e profissional.

Agradecimentos

Agradeço primeiramente a Deus pela presença constante em minha vida. Por me dar força e sabedoria para vencer esse grande desafio.

Agradeço toda minha família, meus pais Carlos e Diva pela educação que me deram, pelo amor incondicional. Minhas irmãs queridas Daniella e Vivian, pelo carinho.

Agradeço meu querido esposo Antônio pelo amor, compreensão, incentivo e companheirismo nos momentos tristes e alegres durante a jornada do curso.

Agradeço também meus sogros, Aparecida e Antônio, meu cunhado André, que sempre torceram por mim.

Agradeço a Vera por fazer parte da nossa família, compartilhando conosco a educação de minha filha Laís e me apoiando nessa vitória.

Agradeço os professores Anderson Duarte e Luiz Duczmal pela confiança e dedicação dispensados. Exemplos de profissionalismo e caráter que levarei sempre em minha caminhada profissional.

Agradeço as amizades que conquistei durante o curso. Cleide, Luciano, Ronaldo, Spencer, Emerson, Fernando, Nilson, Flávio, Ricardo, todos os professores, amigos sempre prontos a compartilhar os desafios e as alegrias.

Agradeço também a Antônia, minha gerente e amiga na Fiscalização da PBH, pela oportunidade concedida de realizar meu aperfeiçoamento profissional.

Resumo

Estratégias para detecção de conglomerados (*clusters*) espaciais tanto para dados por regiões quanto para dados pontuais são já bastante difundidas. Entenda-se por dados pontuais situações em que cada elemento da população é tratado individualmente, sabendo-se sua localização no mapa em estudo. Entretanto os problemas com clusters de formato irregular não estão completamente fechados. O cluster mais verossímil geralmente se espalha em grandes parcelas do mapa, impactando seu significado geográfico. Métodos que empregam a estatística Scan Espacial de Kulldorff, associados a medidas de penalização, foram usados para controlar a liberdade excessiva de forma dos clusters. Estes métodos em geral não foram aplicados para dados pontuais. Neste contexto, será apresentado um novo algoritmo multi-objetivo utilizando a estatística Scan Espacial e a *penalização por Não-conectividade Ponderada* para dados pontuais. A solução é um *conjunto de Pareto*, consistindo de todos os clusters não simultaneamente piores em ambos os objetivos. A melhor solução é determinada pela avaliação da significância através de simulações de Monte Carlo. Utilizamos uma teoria estatística para avaliar o significado estatístico de soluções obtidas através do algoritmo multi-objetivo empregando o conceito de *funções de aproveitamento*.

Palavras-chave: vigilância sindrômica; estatística espacial Scan de Kulldorff; dados pontuais; clusters espaciais de formato irregular; algoritmos multi-objetivos; função de Não-conectividade; Conjunto de Pareto.

Abstract

Strategies for detecting clusters for both spatial regions data and for point data are already quite widespread, it is understood by data point, situations in which each element in the population is treated individually, knowing its location on the map under study. The problems with irregularly shaped clusters are not closed. The most likely cluster generally spreads in large portions of the map, impacting its geographic significance. Statistical methods that use the Kulldorff's Spatial Scan, combined with penalty functions were used to control the excessive freedom of clusters' shapes. These methods have been not applied to point data. In this context, we will present a novel multi-objective algorithm using the Spatial Scan Statistic and *penalty function for Non-connectivity Weighted* to points data. The solution is a *Pareto set*, consisting of all clusters not less in both objectives than the others. The best solution is determined by evaluating the significance through Monte-Carlo simulations. We use a statistical theory to evaluate the statistical significance of the solutions obtained by multi-objective algorithm that employs the concept of *attainment functions*.

Keywords: disease surveillance; Kulldorff's Spatial Scan; point data; Irregular spatial clusters; multi-objective algorithms, Non-connectivity function; Pareto-set.

Índice

Capítulo 1

Introdução

1.1 Motivação

Pesquisadores da área de saúde pública constantemente se vêem obrigados a identificar áreas em que o risco de incidência de algum fenômeno de interesse seja significativamente discrepante, ou seja, muito elevado ou muito baixo se comparado aos valores esperados.

Existem diversos métodos para detecção de clusters espaciais, alguns deles tratando de dados por regiões, bem como alguns tratando de dados pontuais (caso-controle). Neste enfoque dados pontuais são situações em que cada elemento da população é tratado individualmente, sabendo-se sua localização no mapa em estudo e alguns deles serão considerados CASOS, ou seja, o fenômeno em estudo ocorreu para o indivíduo, enquanto outros serão considerados CONTROLES, ou seja, o fenômeno de interesse não ocorreu para o indivíduo.

Dentre os métodos existentes, alguns são discutidos para os dois enfoques. Entretanto existem métodos bastante eficientes para dados por regiões que não foram avaliados para dados pontuais.

Quanto ao processo de detecção em si, este pode ser realizado em intervalos de tempo *(cluster temporal)* ou então, para localizações no espaço *(cluster espacial)*, ou em ambos *(cluster espaço-temporal)*. Existem diversas razões para estudar o processo de avaliação e detecção de clusters. Elas podem ser de natureza reativa (investigação de alarme de alta incidência da doença), proativa (monitoramento contínuo de áreas com alta incidência) ou etiológica (busca por características de incidência de uma doença, previamente desconhecida). Podemos relacionar o problema de detecção de clusters em diversas situações tais como problemas associados a saúde pública (epidemiologia, vigilância sindrômica), poluição, criminologia, pesquisas de mercado, entre outros.

Existe uma diferença importante entre as abordagens para dados pontuais e para dados por regiões. Quando analisamos dados por áreas, cada área além de possuir um número de casos observados possui ainda uma população de risco. Já a análise de dados pontuais trata cada ponto como um *caso* ou um *não caso*. Portanto não temos posse de um valor populacional associado a cada ponto. Talvez este seja o principal ponto a ser resolvido para estender métodos para dados por regiões até estudos para dados pontuais.

Dentre os muitos trabalhos já existentes, alguns utilizam abordagens de otimização multi-objetivo. A eficiência dos resultados observados através de métodos de otimização multi-objetivo serviram como motivação fundamental para a discussão de uma metodologia de otimização multi-objetivo para o problema de detecção de clusters para dados caso-controle.

Um dos objetivos deste trabalho é determinar e desenvolver estratégias para a detecção de clusters espaciais para dados pontuais. De fato notamos que não existe um melhor método, mas sim uma extensa gama de métodos que se adequam bem em diferentes cenários.

Buscamos um algoritmo que seja capaz de delinear, o mais corretamente possível, o subconjunto de pontos dentre os pontos do conjunto completo de casos e controles, que serve como delineamento do possível cluster, apresentando justificativas estatísticas para o funcionamento adequado do método. Apesar da dissertação restringir-se ao estudo para detecção de clusters espaciais para dados pontuais, as propostas aqui discutidas podem ser estendidas para a busca de clusters espaço-temporais.

De uma forma geral podemos definir como principal contribuição a formulação de algoritmos para detecção e inferência de clusters espaciais. As principais contribuições são as seguintes:

- apresentação de uma revisão bibliográfica atualizada da área;
- proposição de um novo modelo de penalização para a estrutura topológica de um cluster;
- proposição de um modelo matemático para a topologia de clusters baseado em grafos;
- utilização de algoritmos genéticos multi-objetivo para resolução do problema;
- obtenção de resultados, para uma ampla gama de experimentos computacionais, que atestam a qualidade das soluções que podem ser obtidas pela metodologia proposta.

Dentre os métodos mais difundidos para detecção de clusters em dados pontuais, podemos mencionar Bessag and Newell [1990] que utiliza janelas circulares centradas nos casos para formação de possíveis clusters. Já em Kulldorff [1997] é abordada a estratégia Scan Circular considerando todos os pontos (casos e controles) como centros de janelas circulares.

1.2 Revisão Bibliográfica

Apresentaremos uma revisão bibliográfica dos métodos já existentes, discutindo métodos para detecção de clusters com dados divididos por regiões e com dados pontuais.

Os métodos de detecção e inferência para clusters espaciais são muito importantes em diversas áreas como vigilância sindrômica, epidemiologia entre outras. Isto pode ser observado em Lawson et al. [1999], Moore and Carpenter [1999], Lawson [2001], Glaz et al. [2001], Balakrishnan and Koutras [2002], Buckeridge et al. [2005]. A estatística Espacial Scan definida em Kulldorff [1997] como uma razão de verossimilhança busca detectar o cluster mais verossímil dentre algumas possíveis configurações de clusters no mapa em estudo. A utilização deste método requer a escolha de um formato específico de janela para o procedimento de busca. Uma escolha já muito utilizada, é o formato circular das janelas para o referido método, denominado Scan Circular, apresentado em Kulldorff and Nagarwalla [1995]. A metodologia proposta através do Scan Circular é aplicável tanto para dados agrupados por regiões quanto para dados pontuais. Este formato é bastante eficiente, mas apresenta algumas deficiências quando os clusters a serem detectados não apresentam formato regular (por exemplo conjuntos não circulares) o que ocorre com frequência nesta área de estudo.

Uma estratégia bastante utilizada em casos de dados pontuais pode ser encontrada em Bessag and Newell [1990]. Neste caso, o objetivo é identificar clusters mais prováveis de forma circular. A utilização do método requer a especificação do número mínimo k de casos que devem estar dentro de uma janela antes da mesma configurar um cluster. Com círculos centrados em cada ponto (*caso*), o método calcula o raio necessário para que o círculo contenha pelo menos k casos. Tal procedimento é implementado através de um

algoritmo que começa a partir de um único ponto (considerando raio igual a zero). O raio é aumentado até incluir o ponto (*caso*) mais próximo, este ponto é então incluído no candidato a cluster. O procedimento é paralisado quando a janela circular contém k casos. A estatística de teste é então calculada baseada no círculo formado ao redor de cada caso. O método portanto, desenha no mapa círculos estatisticamente significantes. Isso leva a necessidade de um parâmetro adicional, o nível de significância. A principal vantagem desta proposta é a estimativa da localização do cluster, uma importante questão para a prática de intervenção na saúde pública. O método requer o conjunto de dois parâmetros, k e α. Uma alternativa para superar a influência do parâmetro k é executar o procedimento para várias escolhas diferentes de k e combinar os resultados de alguma forma. Pode-se encontrar uma interessante revisão bibliográfica sobre metodologias para detecção de clusters e dos artigos citados anteriormenente em Duczmal et al. [2009].

1.3 Escopo da dissertação

Buscamos discutir um procedimento de detecção de clusters para dados pontuais. Acreditando na eficiência de alguns métodos para detecção de clusters para dados por regiões, estabeleceremos uma estratégia de associação entre o problema de dados pontuais com o problema de dados por regiões.

A associação será construída através do Diagrama de Voronoi. Este Diagrama divide o espaço em estudo através de polígonos, cada um destes polígonos está associado a um dos pontos que deu origem ao Diagrama de Voronoi. Um polígono de Voronoi associado a um ponto é o conjunto dos pontos mais próximos ao ponto originário dentre todos os pontos do plano no qual estão inseridos os pontos originais (casos e controles).

A inserção do Diagrama de Voronoi no problema será importante para a constituição de uma estrutura de vizinhança entre os pontos (casos e controles). Será então possível determinar se dois pontos são vizinhos ou não.

Outro item não menos relevante será a definição de uma estratégia de estimação de densidade populacional associada a cada um dos pontos (casos e controles). Vale ressaltar que estamos interessados em utilizar métodos para dados por regiões em um problema de dados pontuais. As estratégias usuais para dados por regiões são bastante dependentes da distribuição populacional no mapa.

Utilizaremos a estatística Scan e alguma função penalizadora para uma abordagem multi-objetivo de detecção de cluster. Portanto a estrutura de vizinhança entre pontos e as estimativas populacionais serão preponderantes para o desenvolvimento deste trabalho.

Este texto se encontra organizado da seguinte forma: o capítulo 2 descreve os métodos usuais para detecção de clusters espaciais com dados por regiões; o capítulo 3 apresenta detalhes do algoritmo genético que será utilizado no procedimento de detecção; os capítulos 4 e 5 descrevem a nova proposta para detecção de clusters espaciais com dados pontuais; o capítulo 6 discute assuntos de inferência visando uma avaliação da significância estatítica das soluções obtidas; o capítulo 7 apresenta resultados para testes simulados e um estudo de dados reais; o capítulo 8 relata as conclusões obtidas através deste trabalho.

Capítulo 2

Métodos para detecção de clusters espaciais

Dado que visamos analisar um conjunto de dados pontuais através de técnicas para dados por regiões, será importante descrever de forma resumida alguns métodos para dados por regiões.

2.1 Uma apresentação para o problema para dados por regiões

Suponha que tenhamos um mapa dividido em regiões, cada uma delas com uma população conhecida e um número de casos observados para a ocorrência de um determinado fenômeno de interesse. Assim, cada caso pode ser, por exemplo, um indivíduo infectado por uma certa doença ou uma vítima de um determinado tipo de crime. Neste mapa um *cluster* é um aglomerado de regiões vizinhas onde o risco de ocorrência do fenômeno de interesse é muito elevado ou muito baixo se comparado com o risco das demais regiões, e ao mesmo tempo significativo do ponto de vista estatístico.

Para cada região definimos um centróide, que é um ponto arbitrário em seu interior. Chamaremos de *zona* qualquer subconjunto conexo de regiões do mapa. A figura 2.1 mostra uma zona no mapa do estado de São Paulo dividido em 72 microrregiões.

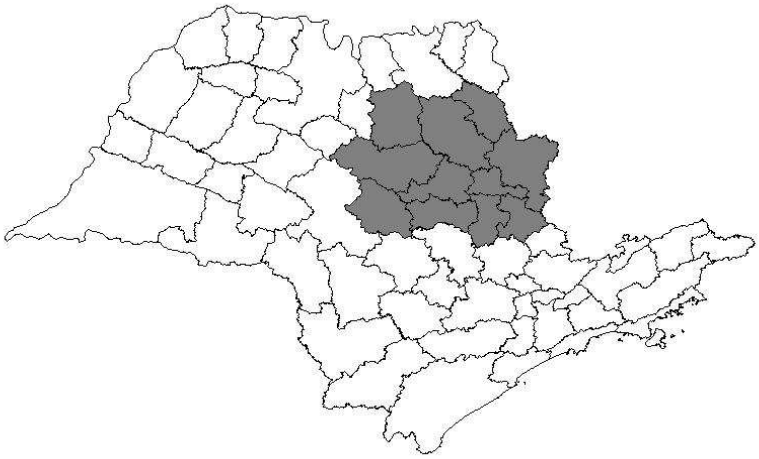

Figura 2.1: Mapa do estado de São Paulo dividido em microrregiões.

Kulldorff [1997] propõe uma metodologia baseada em um teste de razão de verossimilhança. Kulldorff and Nagarwalla [1995] apresentam o Scan Circular, um teste que encontra o cluster mais verossímil dentre todas as zonas circunscritas por círculos de raios variados centrados em cada região do mapa.

Uma janela circular sobre a área em estudo define uma zona formada pelas regiões cujos centróides estão dentro da janela. Note pela figura 2.2 que, embora parte de uma região possa estar dentro da janela circular, se seu centróide está fora dessa janela, essa região não fará parte da zona. Do mesmo modo, mesmo que uma região não esteja totalmente inserida na janela, se seu centróide está, então essa região fará parte da zona definida pela janela.

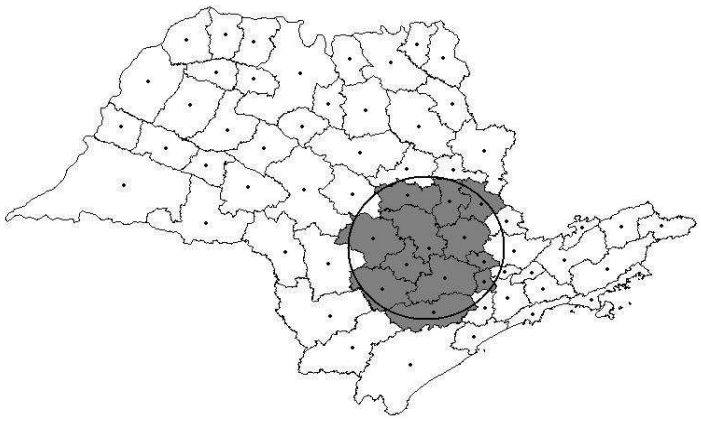

Figura 2.2: Uma possível zona obtida para uma dada janela circular.

Denotaremos por Z o conjunto de todas as zonas obtidas por janelas centradas em cada centróide e de raios variando entre zero e um valor máximo. A busca por soluções eficientes seria feita então dentro do conjunto Z. Um grande problema desses métodos é a forma fixa dos clusters detectados, tipicamente circulares ou quadrados, dependendo do método.

Essa restrição vem do fato de que seria computacionalmente inviável testar todas as zonas possíveis. No entanto, em situações reais frequentemente encontramos clusters em formatos bastante diferentes. A incidência de uma doença pode ser maior ao longo de um rio, por exemplo, o que daria uma forma mais alongada ao cluster.

Muitos algoritmos para detecção de clusters espaciais não têm procedimentos adequados para controlar as formas dos clusters encontrados. A solução pode às vezes se espalhar através de diversas regiões do mapa, fazendo com que se torne difícil a avaliação de seu significado geográfico.

Uma primeira ideia para tentar detectar um cluster poderia levar em conta simplesmente a incidência de casos em cada zona, isto é, o número de

casos observados dividido pela população, ou ainda o risco relativo que é o número observado de casos dividido pelo número esperado de casos. Apesar de parecer razoável, essa análise não resolve o problema de detecção de clusters, porque é possível que clusters com populações muito discrepantes possam apresentar uma mesma proporção de casos. Neste caso, estes candidatos seriam comparados em situação de igualdade, quando na verdade são bastante diferentes devido à discrepância entre as populações. Um aumento no risco relativo é tão mais significativo quanto maior é a população de risco dentro do cluster candidato. Isso significa que, embora uma região ou uma zona, possa apresentar um alto risco relativo, se sua população é pequena, ela se torna pouco significativa.

Para contornar este problema, precisamos encontrar um método que nos permita analisar somente as zonas mais promissoras e descartar as que não parecem muito interessantes. Uma vez que não analisam todas as zonas, esses métodos não garantem que encontraremos a solução ótima, mas um bom método deve encontrar uma boa solução na maioria das vezes. A estatística Scan proposta em Kulldorff [1997] prevê a possibilidade de clusters de formato arbitrário, porém não propõe algoritmos para a detecção de clusters de formato irregular.

Neste sentido, existem alguns algoritmos que propõem estratégias para a detecção de clusters com formatos irregulares. Uma técnica bastante razoável e já utilizada, é a incorporação de alguma função de penalização para o formato geométrico ou topologia do grafo associado ao cluster.

2.2 Estatística Scan Espacial de Kulldorff para dados por regiões

Considere o mapa associado a área de estudo A, dividido em m regiões, com população total P e um total de casos C. Um grafo não orientado G_A é associado à área em estudo A. No grafo associado G_A os vértices representam as regiões do mapa e as arestas conectam vértices associados à regiões adjacentes. É construído um teste de hipóteses no qual a hipótese nula será de não existência de cluster no mapa em estudo.

O número de ocorrências do fenômeno de interesse em cada uma das regiões é distribuído conforme uma Poisson com taxa proporcional à população da respectiva região no caso da hipótese nula, neste caso, para a $i-$ésima região, se considerarmos uma população p_i, a taxa para a distribuição Poisson seria $\mu_i = p_i\left(\frac{C}{P}\right)$. Qualquer subconjunto conexo de regiões no mapa será denominado como uma zona (possível cluster). Teremos então para cada zona z, o número observado de ocorrências do fenômeno de interesse c_z e o número esperado de ocorrências do fenômeno de interesse sob a hipótese nula μ_z. O risco relativo de uma zona z é $I(z) = \frac{c_z}{\mu_z}$ enquanto o risco relativo fora da zona z, ou seja, no complementar da zona z em relação ao mapa em estudo, é dado por $O(z) = \frac{C - c_z}{C - \mu_z}$.

Assumindo o modelo Poisson, considerando L_0 como a função de verossimilhança sob a hipótese nula e $L(z)$ como a função de verossimilhança sob a hipótese alternativa de que existe um cluster no mapa em estudo. Pode-se mostrar (veja em Kulldorff [1997]) que o logaritmo da razão de verossimilhança é dado por:

$$LLR(z) = \log\left(\frac{L(z)}{L_0}\right)$$

$$LLR(z)=\begin{cases} c_z \log(I(z)) + (C-c_z)\log(O(z)) & \text{se } I(z) > 1 \\ \\ 0 & \text{caso contrário} \end{cases} \tag{2.1}$$

A razão de verossimilhança é maximizada em um conjunto Z formado por zonas z definidas no mapa em estudo. O conjunto Z é definido segundo algum critério restritivo visando não realizar uma busca exaustiva em todas as possíveis zonas z, mas apenas em um conjunto de zonas mais promissoras. Um exemplo de construção do conjunto Z, bastante utilizado, seria o conjunto das zonas definidas por janelas circulares de raios e centros variados. Tal estratégia é conhecida como Scan Circular proposta em Kulldorff and Nagarwalla [1995]. Se considerarmos Z como o conjunto de todas as zonas conexas, o problema se tornaria impraticável, entretanto podemos utilizar heurísticas (algoritmos estocásticos) avaliando somente candidatos em potencial e não todo o conjunto Z.

A significância estatística de uma solução, obtida através da distribuição dos casos observados, pode ser verificada através de simulações de Monte Carlo. Sob a hipótese nula, os casos simulados são distribuídos nas regiões do mapa em estudo. Usando tal distribuição será obtido o cluster mais verossímil. Este procedimento é repetido por diversas vezes, e a distribuição empírica dos valores para a razão de verossimilhança pode ser obtida. Esta distribuição empírica é comparada então com a razão de verossimilhança da solução obtida para os casos observados, produzindo então uma estimativa de p-valor para esta solução (Dwass [1957]).

2.3 Estratégias com penalizações

Quando pensamos em avaliar todos os possíveis subconjuntos de regiões no mapa em estudo, o número de possíveis candidatos aumenta exponencialmente. Para um mapa dividido em m regiões, existem $2^m - 1$ possíveis subconjuntos de regiões, dos quais deveríamos verificar quais são conexos. Considerando viável a possibilidade de investigar todos os possíveis subconjuntos de regiões, ainda assim o problema persistiria, pois poderíamos encontrar possíveis clusters que são formados conectando regiões levando em conta apenas elevados riscos de ocorrência do fenômeno em estudo. Seriam construídos então, subconjuntos bastante irregulares, formando então soluções não viáveis. Possíveis clusters muito irregulares que se espalham através do mapa, tomando grandes regiões da área em estudo, tendem a não apresentar informações úteis. É plausível esperar que os verdadeiros clusters sejam conjuntos menores, mais regulares, mesmo que com razão de verossimilhança um pouco menor. Este fato leva a motivação da utilização de funções de penalização para evitar a possibilidade de qualquer formato para uma possível solução. Tais funções de penalização podem ser associadas à estatística espacial Scan. Dentre estas funções de penalização, podemos mencionar a penalização por Compacidade Geométrica apresentada em Duczmal et al. [2006]; a penalização por Coesão Topológica descrita em Cançado et al. [2010] e as penalizações por Não-conectividade e Não-conectividade Ponderada apresentadas em Yiannakoulias et al. [2005], Silva [2010], Duarte et al. [2010b]. Nestes trabalhos, as funções de penalização, aparecem como expoente para a razão de verossimilhança ou então, em abordagens multi-objetivo.

A abordagem que utiliza um algoritmo genético multi-objetivo para o problema de detecção de clusters foi apresentada originalmente em Duczmal et al. [2008]. Este método conduz a uma estratégia que busca maximizar dois obje-

tivos, sendo eles: a estatística Espacial Scan e a Compacidade Geométrica que avalia a regularidade do formato geométrico do possível cluster. Não é apresentada uma única solução, mas sim um conjunto de soluções não-dominadas, ou seja, que não são piores que outras soluções nos dois objetivos simultaneamente. O algoritmo multi-objetivo apresenta uma importante vantagem: todos os clusters potenciais são considerados sem uma classificação de acordo com os valores da penalização. Assim a classificação quanto à qualidade das possíveis soluções é executada somente depois que todos os candidatos são avaliados.

A avaliação quanto à significância estatística é realizada paralelamente para todos os clusters do conjunto de soluções não-dominadas usando simulações de Monte Carlo, quebrando o laço de dependência entre elas, como será explicado no capítulo 6, e determinando a melhor solução no conjunto de soluções não-dominadas. Utilizamos para a avaliação da significância estatística a teoria de funções de aproveitamento apresentada em da Fonseca et al. [2001], Fonseca et al. [2005]. A utilização da função de aproveitamento no problema específico de detecção de clusters se encontra bem detalhada em Cançado [2009], Duarte [2009], Cançado et al. [2010]. O uso da função de aproveitamento permite que o significado do p-valor para o espaço bi-objetivo seja mais claramente compreendido. Vamos descrever então de forma sucinta tais funções de penalização.

2.3.1 Penalização por Compacidade Geométrica

Como já citado anteriormente, os algoritmos para detecção de clusters espaciais podem encontrar soluções em forma de árvore que se espalham ao longo do mapa conectando as regiões com elevada incidência. Uma forma de evitar tais soluções seria a utilização de um algoritmo que busca soluções

através da $LLR(z)$, juntamente com alguma estrutura de penalização para o formato do possível cluster. Neste caso, não estaríamos presos a um formato de janela de busca, mas sim avaliaríamos os candidatos em potencial segundo a $LLR(z)$ e alguma medida de penalização. Um destes possíveis formatos de penalização é a medida de Compacidade Geométrica.

Esta penalização foi apresentada em Duczmal et al. [2006] e seu objetivo é penalizar as zonas do mapa que possuem formato muito irregular. A Compacidade Geométrica $k(z)$ de uma zona z é dada pela área da zona z definida por $A(z)$ dividida pela área do círculo com o mesmo perímetro que o fecho convexo da zona z, sendo este aqui definido por $H(z)$.

A expressão para $k(z)$ é dada por:

$$k(z) = \frac{A(z)}{\pi \left(\frac{H(z)}{2\pi} \right)^2} \tag{2.2}$$

A Estatística Scan penalizada pela Compacidade Geométrica será definida por $max_{z \in Z} LLR(z).k(z)$.

2.3.2 Penalização por Coesão Topológica

Cançado et al. [2010] apresentam uma proposta para penalização denominada Fórmula da Coesão. Este formato de penalização tenta avaliar a estrutura topológica de cada possível cluster. O objetivo é penalizar zonas candidatas ao possível cluster através da existência de "elos fracos". São candidatos a elos fracos, regiões que, ao serem removidas do cluster, desconectam o cluster, quebrando este em partes.

Para qualquer zona z, com regiões $v_1, v_2, \ldots, v_n$, definida no mapa, o seguinte procedimento é executado:

- Inicialmente retira-se da zona z a região v_1, construindo a nova zona $\bar{z}$;

- Então é verificado se a nova zona $\bar{z}$ é conexa;

- Se a nova zona $\bar{z}$ é conexa, considera-se a retirada da próxima região v_2 da zona original z, caso contrário incluí-se v_1 ao conjunto D.

- O procedimento é repetido para todas as regiões até obter o conjunto $D = \{x_1, x_2, \ldots, x_d\}$ que é denotado conjunto dos elos de desconexão;

- Ao retirar-se da zona original z o conjunto D, serão geradas L partes desconexas $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_L$ (cada parte com uma ou mais regiões);

- Ordena-se as partes desconectadas por suas respectivas populações;

  - Seja $\hat{z}_{(1)}$ a parte 1 de maior população entre as partes;
  - Seja $\hat{z}_{(2)}$ a parte 2 com a segunda maior população entre as partes;
  - Até $\hat{z}_{(L)}$ a parte L com a menor população entre as partes.

Quando o conjunto D não é vazio, avalia-se a quantidade

$$c(G) = \left(\prod_{i=1}^{d} \left(1 - e^{-\mu_{x_i}}\right) \right) \prod_{i=1}^{L} \frac{Pop(\hat{z}_{(i)})}{\sum_{j=i}^{L} Pop(\hat{z}_{(j)})} \tag{2.3}$$

Caso contrário $c(G) = 1$

A função definida anteriormente avalia na verdade a relevância de cada região da zona z como um elo que pode desconectar ou não as regiões da zona z, bem como o desequilíbrio entre as populações nas partes desconectadas.

Em outras palavras, é avaliado se os clusters possuem elos fracos, que tendem a se desconectar facilmente.

2.3.3 Penalização por Não-conectividade

Yiannakoulias et al. [2005] propõe um algoritmo guloso que avalia algumas possíveis zonas z. A função de penalização para a Não-conectividade se baseia em uma relação do número de vértices $v(z)$ e de arestas $a(z)$ do subgrafo associado à zona z. De forma análoga a que foi citada na penalização por Compacidade Geométrica, a penalização por Não-conectividade foi utilizada como um multiplicador para a $LLR(z)$. A penalização por Não-conectividade para uma zona z é definida por:

$$y(z) = \frac{a(z)}{3\,(v(z) - 2)} \tag{2.4}$$

O termo $3\,(v(z) - 2)$ no denominador da expressão (2.4), representa o número máximo de arestas para um grafo planar, ou seja, para o grafo planar mais conexo possível teriamos $y(z) = 1$.

Apesar de existir alguma similaridade entre a Penalização por Não-conectividade e a Penalização por Compacidade Geométrica, uma diferença importante é o fato de buscar zonas sem uma associação direta ao formato, mas sim ao grau de conexidade do subgrafo associado à zona z.

2.3.4 Penalização por Não-conectividade Ponderada

A proposta de função de penalização através da Não-conectividade apresentada em Yiannakoulias et al. [2005] se mostra bastante eficiente na detecção e inferência de clusters, quando avaliadas medidas de poder, sensibilidade e PPV (valor preditivo positivo) do teste. Entretanto o formato desta penalização leva em conta apenas a contagem das arestas do subgrafo associado à zona z. Não existe uma consideração quanto ao grau de importância de uma aresta na conexidade do subgrafo.

Pensando apenas na análise do grafo é fato que tal relevância não precisa ser considerada. Quando se observa que os subgrafos estão associados à zonas em um mapa, lembra-se de que as arestas são conexões de vizinhança entre regiões que podem ser muito ou pouco populosas. Neste contexto, observa-se que existem sim arestas mais e menos importantes para a conexidade do subgrafo associado a uma zona z. A mesma análise pode ser realizada para o grau de importância de cada um dos vértices do subgrafo em estudo.

Para tanto, é estabelecida uma ponderação para os vértices e arestas do subgrafo associado à zona z. Tal ponderação é construída pensando na estrutura da distribuição populacional ao longo das regiões da zona z.

A ponderação das arestas do subgrafo associado a zona z é definida pela média entre as populações das regiões cujos vértices são conectados pela aresta em questão. Portanto uma aresta $a_{i,j}$ conectando os vértices v_i e v_j associados às regiões Ri e Rj com populações $pop(Ri)$ e $pop(Rj)$, terá o seguinte peso ponderador:

$$P(a_{i,j}) = \frac{pop(Ri) + pop(Rj)}{2}.$$

Já a ponderação dos vértices é dada pela população da região associada ao respectivo vértice, ou seja, um vértice v_i associado à região Ri cuja população é $pop(Ri)$, terá o seguinte peso ponderador:

$$P(v_i) = pop(Ri).$$

Para reformular a função descrita em Yiannakoulias et al. [2005], substitui-se as arestas e vértices por seus respectivos pesos ponderadores da seguinte forma (veja Silva [2010], Duarte et al. [2010b]):

$$WNC(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} P(a_{i,j})}{3 \left[\sum_{i=1}^{k} P(v_i) - 2 \left(\frac{\sum_{i=1}^{k} P(v_i)}{k} \right) \right]} \tag{2.5}$$

em que k é a quantidade de regiões na zona z.

Com este novo formato a medida leva em conta não somente a estrutura do subgrafo associado à zona z, mas também informações inerentes a estrutura da distribuição populacional dentro da zona z e o grau de relevância das vizinhanças entre regiões quanto às suas populações.

Capítulo 3

Algoritmos Genéticos

3.1 Algoritmo Genético Mono-Objetivo

Para utilizar os procedimentos citados anteriormente, se faz necessária a utilização de alguma heurística de otimização. Dentre as possíveis heurísticas para serem utilizadas no problema de detecção de clusters, o algoritmo genético foi implementado para detecção e inferência de clusters em Duczmal et al. [2007a] usando como objetivo a ser maximizado a estatística de teste Scan de Kulldorff (2.1).

O algoritmo genético utiliza o princípio da evolução biológica para procurar as melhores soluções de um problema de otimização. São simulados os mecanismos de variação aleatória e de seleção adaptativa da evolução natural. Os mecanismos (operadores genéticos) que constituem a base de um algoritmo genético são:

1. Um operador de cruzamento que gera novos indivíduos a partir da combinação da informação contida em dois ou mais indivíduos;

2. Um operador de mutação que utiliza a informação contida em um in-

divíduo para estocasticamente gerar outro indivíduo.

3. Um operador de seleção que decide se um indivíduo terá a oportunidade de gerar descendentes para a próxima geração, baseado em sua aptidão.

O algoritmo parte de uma população inicial de possíveis soluções para construir uma sequência de gerações. Nas gerações, são utilizados os três operadores: *cruzamento* e *mutação* que servem para aumentar a variabilidade da população de soluções, e *seleção*, que escolhe as soluções que passarão à próxima geração, direcionando a busca e mantendo fixo o tamanho populacional dentro de uma geração.

Para o nosso problema específico, a população inicial deve ser capaz de captar as informações do mapa como um todo. Não há razão para iniciarmos o algoritmo com os indivíduos concentrados em apenas uma parte do mapa, mesmo porque um cluster somente pode ser identificado se possuir valor de LLR discrepante das demais zonas, o que nos obriga a ter um mínimo de conhecimento sobre zonas espalhadas pelo mapa. Para tanto utilizamos uma estratégia gulosa (*algoritmo guloso*) visando obter zonas com alta LLR, construindo as zonas para a população partindo de cada uma das regiões do mapa em estudo, através da estratégia gulosa. Já entre os operadores temos:

1. O operador de cruzamento cria novos indivíduos, ou seja, novas zonas, misturando as características de dois indivíduos (zonas) aleatoriamente escolhidos e denominados por A e B. Diversos novos indivíduos são produzidos assim, sendo eles, zonas intermediárias entre as duas zonas extremas A e B. No formato de implementação que foi utilizado, um cruzamento somente é possível entre duas zonas cuja interseção de regiões entre as zonas A e B seja não vazia. As novas zonas geradas por um cruzamento representam uma transição entre as características

de A e B escolhidas mantendo a conexidade nas zonas geradas, veja a Figura 3.1.

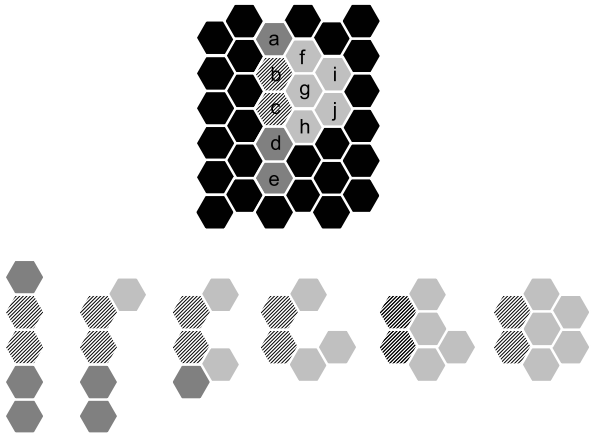

Figura 3.1: Cruzamento entre zonas $A = \{a, b, c, d, e\}$ e $B = \{b, c, f, g, h, i, j\}$.

2. O operador de mutação introduz uma perturbação aleatória nas características de uma zona individual, (adicionando ou removendo uma região ao acaso) assim aumentando a variabilidade da população. Do ponto de vista computacional, a operação de mutação tem custo elevado dada a necessidade de verificação de conexidade a cada operação.

3. O operador de seleção classifica as zonas de acordo com o valor da função objetivo, no caso a Estatística Espacial Scan, escolhendo então aqueles que farão parte da geração seguinte. Esperamos encontrar os indivíduos (zonas) com valores cada vez maiores para a função objetivo à medida que as gerações vão evoluindo. Uma função de penalização como as descritas anteriormente pode ser empregada para evitar a irregularidade excessiva da possível solução.

3.2 Algoritmo Genético Multi-Objetivo

Os algoritmos genéticos são bastante utilizados para problemas de otimização multi-objetivo, analisando a evolução de possíveis soluções avaliando paralelamente dois ou mais objetivos como em Fonseca and Fleming [1995], Takahashi et al. [2003]. Duczmal et al. [2007b] propõem a utilização da Penalização por Compacidade Geométrica através de uma proposição Multi-Objetivo. Nesta proposta a penalização seria uma das funções objetivo, enquanto o logaritmo da razão de verossimilhança $LLR(z)$ seria a outra função objetivo. Tal proposta é extendida para as outras medidas de penalização propostas em Duarte et al. [2010a]. A penalização escolhida, aqui definida por $Pen(z)$, é usada não mais como uma correção mas sim como uma função objetivo nova. Em Duarte [2009] tal estratégia é discutida e pode-se ver que ocorre melhora significativa para a busca pela solução mais eficientes para o problema de detecção de clusters irregulares com Algoritmos Genéticos Multi-objetivo em comparação aos Algoritmos Genéticos Mono-objetivo.

A construção da população inicial e também os operadores de cruzamento e de mutação são idênticos àqueles usados no algoritmo genético mono-objetivo (veja Duczmal et al. [2007a] para uma descrição detalhada daqueles operadores).

- No início de cada geração, construímos a lista da geração atual, que consiste no conjunto dos indivíduos da geração anterior que foram selecionados.

- Esta lista é completada com a adição do resultado dos cruzamentos realizados para esta geração através do operador do cruzamento.

- A lista de geração seguinte, inicialmente vazia, armazena os indivíduos que sobreviverão para a geração seguinte. Obteremos o conjunto das

soluções não-dominadas P_0 da lista da geração atual, que será transferida à lista da geração seguinte inicialmente vazia;

- O mesmo conjunto P_0 é removido igualmente da lista de geração atual. Um conjunto novo P_1 dos indivíduos restantes é obtido da mesma forma;

- O procedimento é repetido até que a lista da geração nova contenha m indivíduos, em que m é o número de regiões do mapa original e corresponde ao tamanho da população que será constante ao longo das gerações.

- Após um número de etapas, o conjunto P_l não será adicionado eventualmente por completo à lista de geração seguinte, porque isto faria com que a lista contivesse mais do que m indivíduos. Nesses casos, os indivíduos de P_l serão transferidos segundo algum critério de desempate entre eles.

Existem alguns possíveis critérios, dentre estes, escolhemos neste trabalho, a distância de aglomeração (*crowding distance*) que se encontra bem detalhada em Deb et al. [2002]. A distância de aglomeração é a maior distância dentre as distâncias entre uma das soluções não dominadas e as soluções não dominadas vizinhas imediatamente à direita e imediatamente à esquerda. Soluções com alta distância de aglomeração, em geral, são provenientes de regiões menos "povoadas" no conjunto das soluções não dominadas. Soluções com baixa distância de aglomeração, em geral, são provenientes de regiões muito "povoadas" no conjunto das soluções não dominadas. A implementação utilizada é o algoritmo genético NSGA-II apresentado em Deb et al. [2002].

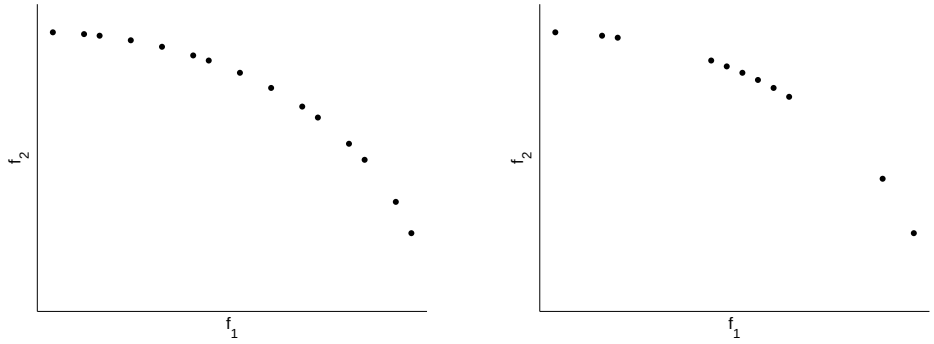

Figura 3.2: Distância de aglomeração.

A Figura 3.2 apresenta à esquerda um conjunto de soluções não dominadas uniformente distribuído (baixa distância de aglomeração em seus pontos) e à direita um conjunto de soluções não dominadas não uniformente distribuído (alta distância de aglomeração em seus pontos). Espera-se então que soluções com baixa distância de aglomeração já tenham sido representadas anteriormente devido ao alto "povoamento" nestas regiões do conjunto de soluções não dominadas, portanto podendo ser excluídas através deste critério de desempate.

Capítulo 4

Métodos para detecção de clusters para dados pontuais

Buscamos construir uma analogia entre os problemas de dados pontuais e os problemas de dados por regiões para então utilizar uma abordagem multi-objetivo. Esta analogia será baseada no Diagrama de Voronoi associado aos pontos (casos-controle). Outro fato relevante será a proposição de uma estimativa de densidade populacional associada a cada ponto (caso-controle). Apresentaremos o Diagrama de Voronoi, a estrutura de vizinhança constituída por este Diagrama. Finalizaremos este capítulo apresentando estratégias de estimação da densidade populacional associada a cada ponto (caso-controle) para o cálculo da função de verossimilhança.

4.1 Diagrama de Voronoi

Em um mapa com dados pontuais, não existem regiões associadas aos pontos. Estamos interessados em associar uma subregião do mapa para cada um dos pontos, para tanto definiremos o Diagrama de Voronoi.

Definição 4.1 *Dado um conjunto finito de pontos distintos em um espaço contínuo, podemos associar a cada ponto desse conjunto o lugar geométrico no espaço que está mais próximo deste ponto do que de todos os outros pontos desse conjunto.*

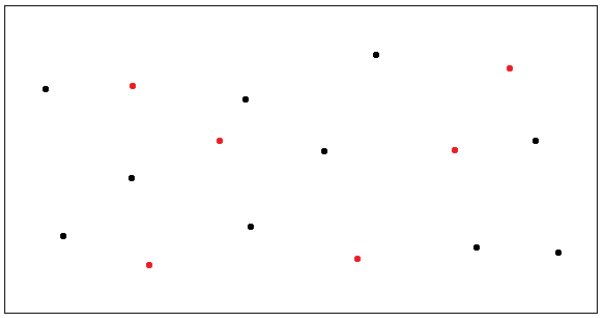

Figura 4.1: Conjunto hipotético de pontos (casos e controles).

Dado o conjunto de pontos e o plano no qual estão representados tais pontos, dividiremos o plano em subregiões. Tais regiões representam o lugar geométrico descrito na Definição 4.1

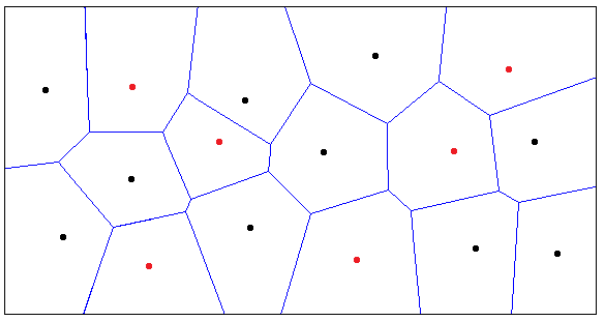

Figura 4.2: Diagrama de Voronoi associado aos pontos da Figura 4.1.

Para cada polígono no entorno de um dos pontos, temos a coleção dos pontos do plano que estão mais próximos deste ponto que de qualquer outro dos pontos geradores do Diagrama de Voronoi.

4.2 Estrutura de Vizinhança

De posse do Diagrama de Voronoi se torna possível estabelecer uma estrutura de vizinhança entre os pontos. Isto será feito através da construção do grafo associado aos pontos. O grafo será construído considerando os pontos como vértices e construindo arestas entre vértices que possuem polígonos de Voronoi que sejam adjacentes.

A importância do grafo associado está no fato de podermos determinar agora se um subconjunto de pontos constitui um subgrafo conexo ou não.

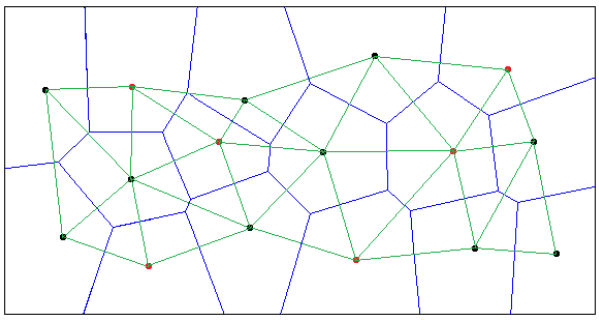

Figura 4.3: Grafo associado ao Diagrama de Voronoi da Figura 4.1.

Agora, dados os polígonos de Voronoi e o grafo associado que constitui uma estrutura de vizinhança entre os pontos, será possível pensar em uma estimativa para a densidade populacional associada a cada polígono e então utilizar técnicas de detecção de clusters para dados por regiões.

4.3 Estimativas populacionais para dados pontuais

A estratégia de detecção que será utilizada depende do cálculo do logarítmo da razão das verossimilhanças já descrita na seção 2.2 como sendo:

$$LLR(z) = \log\left(\frac{L(z)}{L_0}\right)$$

$$LLR(z)=\begin{cases} c_z \log(I(z)) + (C-c_z)\log(O(z)) & \text{se } I(z) > 1 \\ \\ 0 & \text{caso contrário} \end{cases} \tag{4.1}$$

Note que dependemos do número esperado de casos na zona z, dado por $\mu_z = C\frac{pop(z)}{P}$, ou seja, depende-se da população na zona z (subconjunto de regiões) bem como da população P em todo o mapa. Para tanto apresentaremos algumas propostas de estimativa para as densidade populacional associadas a cada polígono de Voronoi.

4.3.1 População 1 para cada polígono

Uma proposta, inicial e simples, seria desconsiderar possíveis diferenças de densidade populacional entre os polígonos de Voronoi. Neste caso estaríamos considerando população 1 para cada polígono, ou seja, o único ponto gerador do polígono de Voronoi seria a população deste polígono. Apesar de parecer simples em demasia tal proposta pode se tornar eficiente como será visto no capítulos subsequentes deste texto.

4.3.2 População associada a uma medida linear.

Uma segunda proposta visa associar a cada polígono um efeito populacional além de seu próprio ponto gerador.

Acreditamos que quanto mais acumulados estejam os pontos, mais populosa será esta vizinhança. Portanto a densidade populacional em um polígono de Voronoi associado a um ponto deveria ser inversamente proporcional às distância entre os pontos.

Dado um polígono de Voronoi construiremos uma medida linear associada aos apótemas dos polígonos. O apótema é o segmento perpendicular ao lado de um polígono que liga este lado ao seu baricentro.

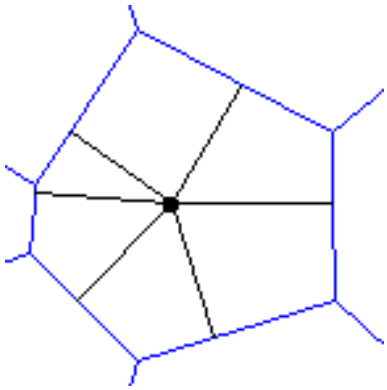

Figura 4.4: Apótemas para um dos polígonos de Voronoi.

Como já citamos, acreditamos em uma porporcionalidade inversa entre a medida linear associada ao polígono e sua densidade populacional. Portanto para cada ponto definiremos a medida fator populacional dada pela média dos inversos dos apótemas. Dado um dos pontos p dentre os pontos em estudo temos:

$$fp(p) = \frac{\sum_{i=1}^{k} \left(\frac{1}{a_{p,i}} \right)}{k}$$

em que p gera um polígono de Voronoi de k lados e $a_{p,i}$ é o apótema associado ao i-ésimo lado do polígono.

Considerando o mapa em estudo com m pontos, a densidade populacional em um polígono de Voronoi gerado pelo ponto p será estimada por:

$$\frac{\hat{\text{Pop}}(p)}{\sum_{p=1}^{m} \hat{\text{Pop}}(p)} \quad \text{em que} \quad \hat{\text{Pop}}(p) = \text{Arredondamento}\left(m * \frac{fp(p)}{\sum fp(i)}\right) + 1$$

O arredondamento utilizado é para o inteiro mais próximo fazendo com que as populações sejam inteiras. O termo $m * \frac{fp(p)}{\sum fp(i)}$ considera a proporcionalidade inversa, já citada e o termo $+1$ considera o ponto utilizado na construção do polígono de Voronoi.

Capítulo 5

O problema dos dados pontuais

Como já descrevemos anteriormente, para possíveis clusters irregulares, a avaliação somente da $LLR(z)$ poderá não ser completamente eficiente. Utilizaremos então uma técnica multi-objetivo em que um dos objetivos a ser maximizado será a $LLR(z)$ e o outro objetivo será a medida de Não-conectividade Ponderada, entretanto apresentaremos uma ponderação específica para o tratamento de dados pontuais.

5.1 Função de penalização por Não-conectividade Ponderada de Grafos para dados pontuais

A primeira proposta de utilização da medida de Não-conectividade Ponderada é apresentada para dados por regiões em Silva [2010], Duarte et al. [2010b]. As ponderações neste caso são associadas às populações. Para dados pontuais as ponderações serão associadas à alocação de casos e controles ao longo do mapa em estudo.

Um possível cluster detectado nesta abordagem pode ser um subgrafo composto não somente por casos, mas também com a inclusão de controles. A ponderação será construída com o intuito de diferenciar clusters que incluem mais ou menos pontos que sejam controles. Para a construção da nova medida penalizadora descreveremos novamente a medida de Não-conectividade.

Dado um subgrafo z a penalização por Não-conectividade originalmente proposta em Yiannakoulias et al. [2005] é dada por:

$$NC(z) = \frac{a(z)}{3(v(z) - 2)},$$

em que $a(z)$ é o número de arestas no subgrafo z e $v(z)$ é o número de vértices no subgrafo z.

É importante observar que dentre as possíveis soluções de nosso algoritmo existem subgrafos z que podem ser construídos não apenas por casos, mas também por controles. Portanto a medida de ponderação que apresentaremos leva em conta o fato de cada ponto do subgrafo z ser um caso ou um controle.

Para cada vértice do grafo consideraremos a proporção de vizinhos-caso dentre os vizinhos pertencentes ao subgrafo z, conforme o exemplo a seguir:

Exemplo 5.1 *Ponderação através da proporção de vizinho-caso.*

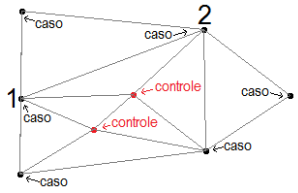

Figura 5.1: Subgrafo hipotético com casos e controles.

O vértice 1 *possui* 5 *vizinhos dos quais* 3 *são casos, logo sua proporção de vizinho-caso é* $\frac{3}{5}$. *Já o vértice* 2 *possui* 5 *vizinhos dos quais* 4 *são casos, logo*

sua proporção de vizinho-caso é $\frac{4}{5}$ *Cada aresta do subgrafo* z *será ponderada pela média das proporções de vizinho-caso dos vértices que tal aresta liga.*

Note que devido a natureza dos pesos construídos, o peso de cada aresta será um número no intervalo $[0, 1]$.

Considerando $p(a_{i,j})$ sendo o peso associado à aresta $a_{i,j}$ em um subgrafo z composto por k pontos teríamos então a nova medida de Não-conectividade Ponderada:

$$WNC(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} p(a_{i,j})}{3(v-2)}. \tag{5.1}$$

Apresentaremos três exemplos para a melhor elucidação das diferenças geradas por tal estratégia de ponderação:

Exemplo 5.2 *Neste caso, como existem somente casos no subgrafo* z *a penalização coincide com a penalização original de Não-conectividade, pois todas as arestas terão peso* 1*.*

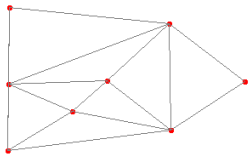

Figura 5.2: Subgrafo hipotético somente com casos.

$$WNC(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} p(a_{i,j})}{3(v-2)} = \frac{15}{3(8-2)} = 0,833.$$

Exemplo 5.3 *Agora com a inclusão de um controle no mesmo subgrafo analisado no exemplo anterior, a nova medida de Não-conectividade ponderada tem seu valor reduzido.*

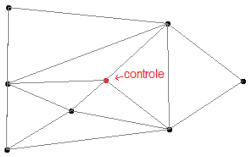

Figura 5.3: Subgrafo hipotético com casos e um controle.

$$WNC(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} p(a_{i,j})}{3(v-2)} = \frac{11,225}{3(8-2)} = 0,624.$$

Exemplo 5.4 *Agora com a inclusão de um segundo controle no mesmo subgrafo analisado no exemplo anterior, a nova medida de Não-conectividade ponderada tem seu valor ainda mais reduzido.*

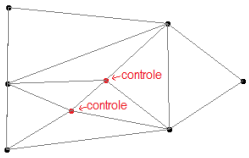

Figura 5.4: Subgrafo hipotético com casos e dois controles.

$$WNC(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} p(a_{i,j})}{3(v-2)} = \frac{8,792}{3(8-2)} = 0,488.$$

Dada a estrutura do grafo associado aos pontos e a nova medida de Não-conectividade ponderada, o novo método de detecção de clusters para dados pontuais está posto. Precisamos então avaliar do ponto de vista estatístico se as soluções apresentadas realmente constituem clusters quanto a sua significância estatística.

Capítulo 6

Inferência em problemas multi-objetivo

A abordagem multi-objetivo será a mesma já utilizada para os problemas de dados por regiões. A heurística de otimização utilizada será o algoritmo genético já descrito anteriormente. Descreveremos agora em detalhes o conceito de dominância através do Conjunto de Pareto para a determinação das melhores soluções em um problema de múltiplos objetivos.

6.1 Conjunto de Pareto

O conjunto de Pareto é o conjunto das soluções não dominadas. Neste contexto, uma solução não dominada é uma solução que nunca é pior que as demais nos dois objetivos simultaneamente.

Definido o espaço de objetivos, para este problema temos $\mathbb{R} \times [0, 1]$, podemos separar as soluções entre não dominadas e dominadas.

Na figura 6.1 podemos observar os pontos circulados representando soluções não dominadas, ou seja, nunca inferiores às demais nos dois objetivos

simultaneamente. Já os pontos representados por cruzes são soluções dominadas, ou seja, inferiores a pelo menos uma das demais soluções em dois objetivos simultaneamente.

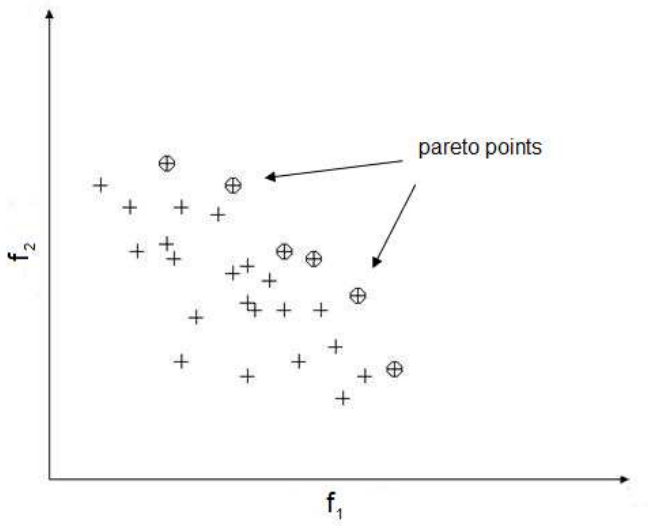

Figura 6.1: Soluções não dominadas

Além do conceito de dominância, para a utilização destas técnicas de otimização será importante uma descrição do critério de avaliação da significância estatística de uma solução através da função de aproveitamento.

6.2 Significância Estatística

Construímos então um algoritmo genético multi-objetivo utilizando como funções objetivo a $LLR(z)$ e a nova medida de não Conectividade Ponderada $WNC(z)$. Precisamos portanto definir uma estratégia para a verificação da significância estatística de uma solução obtida.

Devemos observar que a execução deste algoritmo não nos fornece uma

única solução, mas sim um conjunto de soluções não dominadas, ou seja, uma aproximação do conjunto de Pareto. Buscamos então uma estratégia para verificar para cada solução deste conjunto de soluções não dominadas sua significância estatística.

Através de simulações de Monte-Carlo, de forma similar ao procedimento de Dwass [1957] já mencionado anteriormente na Seção 2.2, podemos executar o algoritmo para diversas distribuições de casos sob a hipótese nula de não existência de clusters no mapa em estudo. Cada uma destas execuções fornece um conjunto de soluções não dominadas. O conjunto destas diversas execuções pode ser utilizado para mensurar a significância estatística de uma solução pertencente ao conjunto de soluções não dominadas obtido através de uma execução do algoritmo no mapa com distribuição original de casos observados. Para tal tarefa será importante definir a função de aproveitamento já citada em da Fonseca et al. [2001], Fonseca et al. [2005], Cançado [2009], Duarte [2009].

6.3 Função de Aproveitamento

Para cada execução do nosso algoritmo, obtemos um conjunto de soluções eficientes. Este conjunto particiona o espaço de objetivos em duas regiões R_1 e R_0: R_1 é a região dos pontos dominados por nosso conjunto de soluções eficientes, ou seja, qualquer ponto de R_1 nunca é superior a qualquer dos pontos do conjunto de soluções eficientes se considerando os dois objetivos simultaneamente; já qualquer ponto que se situasse na região R_0 corresponde a um ponto não dominado pelos pontos do conjunto de soluções eficientes, ou seja, pontos sempre superiores aos pontos do conjunto de soluções eficientes em pelo menos um dos objetivos (veja Figura 6.2).

Para alguma solução x dominada por algum ponto do conjunto de soluções eficientes, ou seja, pertencente a R_1, dizemos que x foi superada por nosso conjunto de soluções eficientes, construindo então um limite para avaliar a significância estatística da solução x.

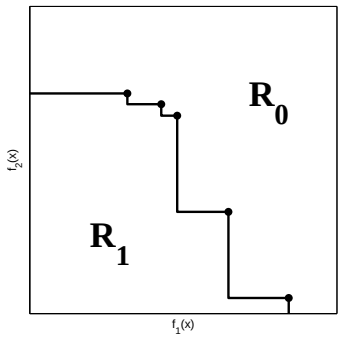

Figura 6.2: Superfície de aproveitamento dividindo o espaço de objetivos.

Podemos repetir a execução do algoritmo para n alocações distintas de casos no mapa, obtidas de cada réplica de Monte Carlo, sob a hipótese nula de não existência de cluster, obtendo então n conjuntos de soluções eficientes, produzindo n limites distintos (veja Figura 6.3).

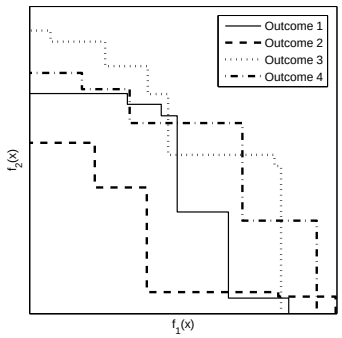

Figura 6.3: Limites obtidos para diferentes execuções do algoritmo.

O conjunto dos n limites pode ser utilizado para dividir o espaço de objetivos em $n+1$ regiões (veja Figura 6.4).

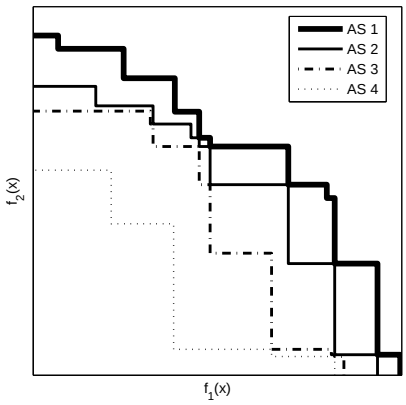

Figura 6.4: Superfícies de aproveitamento para n execuções do algoritmo.

Uma solução que apresenta um ponto no espaço de objetivos à direita de todas as superfícies de aproveitamento, não foi superada em nenhuma das execuções. Ao passo que uma solução que apresente um ponto à esquerda de alguma das superfícies de aproveitamento, foi superado em algumas das execuções. Um ponto à esquerda de todas as superfícies de aproveitamento foi superado em todas as execuções.

Estamos então dividindo o espaço de objetivos em $n+1$ regiões. Podemos com um grande número de execuções sob a hipótese nula de não existência de cluster no mapa, mensurar a significância estatística de uma solução obtida através dos casos originais distribuídos no mapa, através da proporção de regiões não alcançadas no espaço de objetivos.

Lembrando que o método em questão é estocástico, nem todas as possíveis soluções estão sendo avaliadas, portanto não existe garantia que en-

contraremos a solução ótima. Portanto poderíamos ter uma avaliação que subestimasse os p-valores. De fato os p-valores são um pouco menores que os p-valores teóricos.

Capítulo 7

Avaliação de Resultados

7.1 Testes com clusters artificiais

Para avaliar a qualidade do método para detecção e inferência de clusters aqui proposto, precisamos de uma estratégia para avaliar o seu poder de detecção. Para tanto, serão produzidos clusters artificiais sobre um mapa hipotético. Denotaremos estes clusters por *clusters reais*, enquanto os clusters encontrados pelo algoritmo serão denominados *clusters detectados*. Para cada cluster real temos então uma possível construção de hipótese alternativa de existência de um cluster no mapa, ou seja, a existência de um cluster artificial no mapa em estudo.

Inicialmente, construiremos um mapa hipotético, que será um quadrado de lado 1. Dentro deste quadrado distribuiremos aleatoriamente P pontos com suas coordenadas seguindo distribuição uniforme. Uma quantidade pré-estabelecida de casos C é distribuída no mapa de forma que alguns dos P pontos se tornarão casos enquanto os outros serão ditos controles. Em particular, neste estudo foram utilizados os valores $P = 1000$ e $C = 50$. Considerando que cada ponto tenha igual probabilidade de se tornar caso, esta

distribuição satisfaz a hipótese nula de não existência de cluster no mapa em estudo.

O procedimento aleatório para geração da hipótese nula pode ser executado diversas vezes, preservando-se a distribuição dos pontos, mas alterando a cada execução a configuração dos casos. Estas execuções serão utilizadas para a produção de uma superfície crítica através da função de aproveitamento.

Em um segundo momento, construiremos nossas hipóteses alternativas. Para tanto fixaremos uma janela interior ao mapa, que será o cluster real. Tal janela terá seu risco elevado, ou seja, a probabilidade de um ponto se tornar caso será superior se comparada a probabilidade para os pontos exteriores à janela. Portanto, para cada uma das hipóteses alternativas, a mesma quantidade pré-estabelecida de casos na hipótese nula é distribuída aleatoriamente no mapa de acordo com a nova ditribuição de probabilidades. Para esta distribuição o risco relativo para cada uma das regiões é ajustado de forma que fora do cluster real seja igual a um, enquanto nas regiões pertencentes ao cluster real o risco relativo seja idêntico e maior que um. Conforme Kulldorff et al. [2003], apresentamos brevemente como o risco relativo de um cluster é calculado.

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

$$m_0 = C\frac{p_z}{P} \qquad \text{e} \qquad v_0 = C\frac{p_z(P - p_z)}{P^2}$$

Usando a aproximação normal para a distribuição binomial, o número

crítico de casos k para que o teste unilateral rejeite a hipótese nula com o nível de signicância $0 < \alpha < 1$ é tal que:

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

em que $\Phi(.)$ é a função de distribuição acumulada da Normal padrão. Se $\alpha = 0.05$ e $\theta = 1 - \alpha$ temos que $\Phi^{-1}(1 - \alpha) = 1.645$, daí o valor crítico k é tal que $\frac{k-m_0}{\sqrt{v_0}} = 1.645$. Sob a hipótese alternativa, com o risco relativo ρ_z para a região do cluster, o número de casos nesta região tem distribuição Binomial com média $m_a = \frac{Cp_z\rho_z}{(P-p_z+p_z\rho_z)}$ e variância $v_a = \frac{Cp_z\rho_z(P-p_z)}{(P-p_z+p_z\rho_z)^2}$. Observe, neste caso, que $\tau_z = \frac{p_z\rho_z}{(P-p_z+p_z\rho_z)}$. Usando novamente a aproximação Normal, selecionamos o risco relativo ρ_z tal que $\frac{k-m_a}{\sqrt{v_a}} = \Phi^{-1}(\theta)$. Desta forma o risco relativo é escolhido de modo que o poder atingido por qualquer teste para cluster espacial tem um limite superior igual a θ. Neste trabalho a medida para este risco relativo é tal que se a **posição exata** do cluster real **for conhecida**, o poder de detecção deve ser de 0.999.

Para o modelo da hipótese nula, 10000 execuções do algoritmo foram realizadas. Então, através do procedimento da *Função de Aproveitamento* já citada anteriormente, foi produzida uma superfície de aproveitamento para algum nível de significância específico, neste caso utilizamos $\alpha = 0.05$.

Dado um modelo da hipótese alternativa, diversas execuções do algoritmo são realizadas, produzindo então conjuntos de soluções eficientes. Estes conjuntos de soluções eficientes são comparados com a superfície de aproveitamento para $\alpha = 0.05$, obtida anteriormente. O *poder de detecção* é estimado através da proporção de conjuntos de soluções eficientes que possuam pelo menos um ponto à direita da superfície de aproveitamento, ou seja, pelo menos um ponto não dominado em relação a superfície de aproveitamento.

As medidas de sensibilidade e de PPV (valor de predição positivo) igual-

mente servem para avaliar a qualidade do processo da detecção de clusters. Estas medidas são probabilidades condicionais definidas a partir dos seguintes eventos:

V = Indivíduo(caso) escolhido ao acaso na população do mapa pertence a população de casos do cluster verdadeiro;

D = Indivíduo(caso) escolhido ao acaso na população do mapa pertence a população de casos do cluster detectado;

$$Sens = P(D|V) = \frac{P(D \cap V)}{P(V)} = \frac{\left(\frac{Casos(Cluster\ \ Detectado\ \cap\ Cluster\ \ Real)}{Casos(Mapa\ \ em\ \ estudo)}\right)}{\left(\frac{Casos(Cluster\ \ Real)}{Casos(Mapa\ \ em\ \ estudo)}\right)}$$

$$Sens = \frac{Casos(\text{Cluster Detectado} \cap \text{Cluster Real})}{Casos(\text{Cluster Real})}$$

$$PPV = P(V|D) = \frac{P(D \cap V)}{P(D)} = \frac{\left(\frac{Casos(Cluster\ \ Detectado\ \cap\ Cluster\ \ Real)}{Casos(Mapa\ \ em\ \ estudo)}\right)}{\left(\frac{Casos(Cluster\ \ Detectado)}{Casos(Mapa\ \ em\ \ estudo)}\right)}$$

$$PPV = \frac{Casos(\text{Cluster Detectado} \cap \text{Cluster Real})}{Casos(\text{Cluster Detectado})}$$

Neste sentido, um método de detecção de clusters que apresente altas medidas para PPV detecta uma grande porção do cluster verdadeiro, enquanto um método de detecção de clusters que apresente altas medidas para Sensibilidade tem grande parte do cluster detectado pertencente ao cluster verdadeiro. Em outras palavras, para métodos de detecção de clusters, altas medidas para PPV significam que a chance de superestimação no processo de detecção é reduzida, enquanto altas medidas de sensibilidade significam que a chance de subestimação no processo de detecção é reduzida.

Utilizando o mesmo mapa inicial foi escolhida uma janela cuja área era de 1% do mapa. Dado que os pontos no mapa foram distribuídos de forma

uniforme, espera-se que 10 pontos sejam interiores à janela. Entretanto a taxa de variabilidade também é significativa, ou seja, o número de pontos na janela pode variar bastante. Uma única escolha de janela poderia então não ser suficiente para a avaliação dos resultados.

Optou-se então pela escolha aleatória de dez janelas distintas todas de área 1% da área total do mapa (veja Figura 7.1).

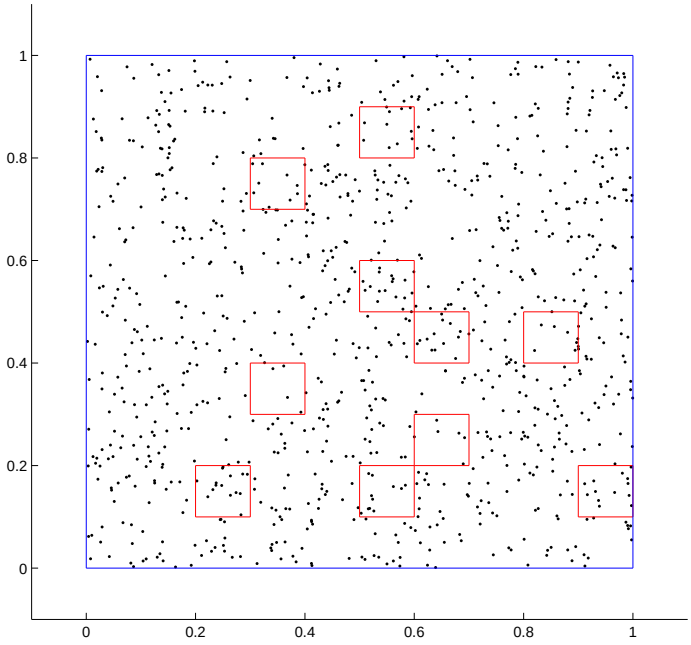

Figura 7.1: Clusters artificiais com 1% da área do mapa

Executou-se o procedimento em 1000 rodadas de Monte-Carlo sob a Hipótese Alternativa para cada uma dessas janelas e considerou-se o resultado médio dos dez procedimentos, como pode ser visto na Tabela 7.1.

Tabela 7.1: Clusters artificiais com 1% da área do mapa

		População 1			Estimativa linear		
cluster	pontos	poder	PPV	Sens	poder	PPV	Sens
A1	4	0.906	0.889	0.791	0.891	0.927	0.858
A2	10	0.947	0.966	0.732	0.930	0.971	0.682
A3	11	0.943	0.959	0.632	0.884	0.958	0.638
A4	11	0.939	0.972	0.604	0.936	0.987	0.663
A5	7	0.953	0.967	0.819	0.941	0.975	0.824
A6	6	0.934	0.888	0.616	0.965	0.978	0.851
A7	14	0.916	0.953	0.517	0.897	0.979	0.538
A8	15	0.912	0.956	0.527	0.897	0.925	0.534
A9	13	0.935	0.958	0.611	0.898	0.932	0.582
A10	18	0.900	0.951	0.425	0.853	0.956	0.437
Média	10.9	0.929	0.946	0.627	0.909	0.956	0.661

Um segundo teste foi realizado com janelas cujas áreas escolhidas foram de $2\%, 3\%$ e 5% da área total do mapa, neste caso as janelas escolhidas não foram mais quadradas. O cluster artificial B é retangular e tem a altura sendo 4 vezes maior que o comprimento de sua base, enquanto o cluster artificial C é retangular e tem a altura sendo 16 vezes maior que o comprimento de sua base (veja Figura 7.2), estes clusters artificiais tem sua área sendo 5% da área total do mapa. Já o cluster articial D é retangular e tem a altura 2 vezes maior que o comprimento de sua base e ocupa 2% da área total do mapa, enquanto os clusters artificiais E e F ocupam 3% da área total do mapa. Neste caso, o cluster E é um retângulo com altura 3 vezes maior que o comprimento de sua base e o cluster F tem formato L (veja Figura 7.3). Agora não foram conduzidos experimentos com escolhas de múltiplas

janelas. Apenas uma janela foi escolhida de forma aleatória e o algoritmo foi executado 5000 vezes sob a hipótese alternativa. Os resultados podem ser observados na Tabela 7.2.

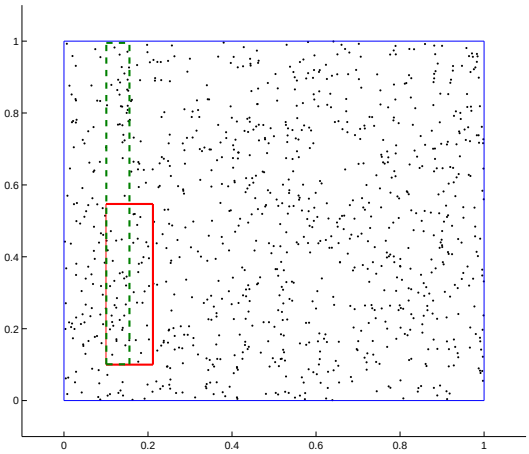

Figura 7.2: Clusters artificiais com 5% da área do mapa.

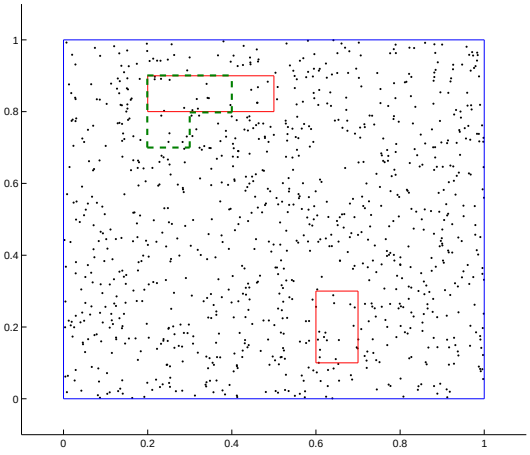

Figura 7.3: Clusters artificiais com 2% e 3% da área do mapa.

Tabela 7.2: Clusters artificiais com $2\%, 3\%$ e 5% da área do mapa

		População 1			População linear		
cluster	pontos	poder	PPV	Sens	poder	PPV	Sens
B (4×1)	53	0.992	0.984	0.347	0.991	0.987	0.325
C (16×1)	60	0.970	0.920	0.192	0.952	0.944	0.178
D (2×1)	18	0.875	0.837	0.580	0.793	0.843	0.521
E (3×1)	18	0.835	0.811	0.569	0.727	0.863	0.496
F (L)	18	0.841	0.784	0.554	0.722	0.867	0.488

Pode-se constatar que a avaliação das medidas de poder e de PPV apresentam resultados bastante significativos. através da Tabela 7.1 nota-se que os melhores resultados são observados em janelas cujo número de pontos é próximo do esperado pela distribuição uniforme. Nota-se também que o algoritmo se comporta um pouco melhor se o número de pontos fica abaixo do esperado se comparado quando fica acima do esperado.

Não detectou-se na Tabela 7.1 diferenças significativas quando comparadas a abordagem que considera população 1 para cada ponto e a abordagen de estimativa linear para densidade populacional (ver Sec.4.3). Em especial, as medidas de sensibilidade parecem estar um pouco abaixo da expectativa.

Para os clusters artificiais B, C, D, E e F apresentados na Tabela 7.2 novamente não detecta-se diferenças significativas entre as duas abordagens para a população. Observa-se resultados promissores quando avaliados o poder e o PPV. Entretanto a avaliação da medida de sensibilidade, principalmente nos clusters B e C, parece não ser satisfatória.

Realmente podemos notar que estes são testes em condições extremas (principalmente os clusters B e C). O formato excessivamente irregular destes clusters faz com que o algoritmo detecte em muitas das execuções de

Monte Carlo, apenas um sub-aglomerado de casos dentro da janela.

É fácil observar que mesmo com o risco elevado dentro de uma janela altamente irregular, mas com uma área grande em relação ao mapa em estudo, como os clusters B e C, o número de casos distribuídos no interior da janela tende a ser ainda bem inferior ao total de pontos na janela. A natureza do algoritmo proposto leva a busca de um aglomerado que inclua muitos casos, ou seja, o algoritmo tende a detectar o melhor sub-aglomerado de casos dentro da janela.

Em outras palavras, dizemos que para janelas muito irregulares (estreitas como o caso do cluster C) existem alguns pontos que se tornam casos dentro da janela, mas se comportam como pontos isolados, devido aos controles que estão no interior da janela cluster. Este fato leva a um decréscimo no valor das medidas de Sensibilidade.

7.2 Estudo com dados reais

Diggle [1990] apresenta um benchmark de dados reais em um mapa de Chorley-Rilble/Inglaterra, com casos de câncer de pulmão e câncer de laringe, registrados entre 1973 e 1984. 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.

Foi investigada a suspeita de que uma incineradora localizada na região pode 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. O mapa de Chorley-Rilble/Inglaterra pode ser observado na Figura 7.4.

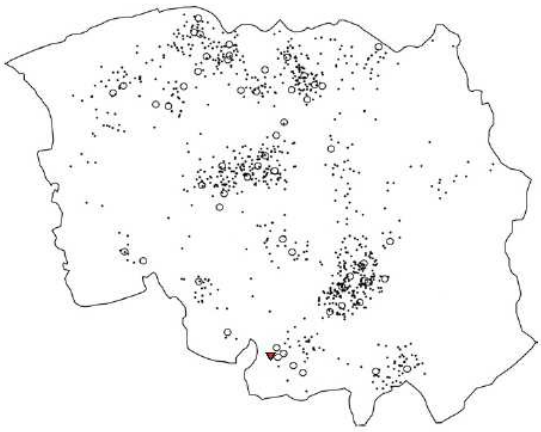

Figura 7.4: Mapa de Chorley-Rilble/Inglaterra, com casos de câncer de pulmão e câncer de laringe, registrados entre 1973 e 1984.

Cucala [2009] estudou o mesmo benchmark de dados e utilizou a estatística Scan com janelas elípticas para detecção de clusters neste mapa. Utilizando um p-valor estimado através de 999 simulações de Monte-Carlo, foi detectado um cluster de cinco pontos (CASOS) utilizando nível de significância $\alpha = 0,05$. Nesta solução o p-valor obtido foi de $0,012$.

Voltamos a utilizar este benchmark de dados agora com a nova proposta de algoritmo multi-objetivo e verificamos os resultados com os dois formatos de estimativa populacional já citados.

7.2.1 Soluções utilizando população 1 em cada ponto

A figura 7.5 apresenta superfícies de aproveitamento para 999 simulações de Monte-Carlo sob a Hipótese Nula de não existência de cluster e também as soluções significativas obtidas dos casos observados em asteriscos. Além disso estão representados os pontos destas soluções bem como a estimativa do p-valor observada.

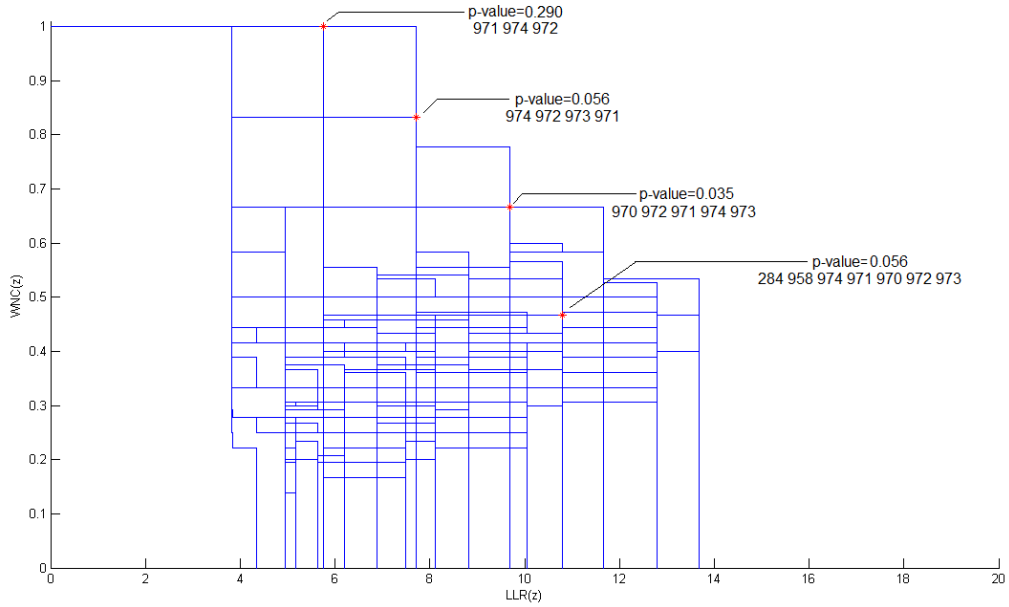

Figura 7.5: Superfícies de aproveitamento para 999 execuções do algoritmo sob Hipótese Nula e soluções para os dados observados.

Em uma análise superficial pode-se pensar que o número de superfícies plotadas parece bem inferior às 999 execuções de Monte-Carlo. Isto se deve ao fato de que várias superfícies estão sobrepostas, afinal considerando população 1 associada a cada polígono de Voronoi teríamos diversas soluções distintas podendo ter exatamente a mesma medida tanto para o objetivo LLR quanto para o objetivo WNC.

7.2.2 Soluções utilizando população estimada pela medida linear

A figura 7.6 apresenta superfícies de aproveitamento para 999 simulações de Monte-Carlo sob a Hipótese Nula de não existência de cluster plotadas e soluções significativas obtidas da distribuição observada em asteriscos, estão representados os pontos destas soluções e a estimativa do p-valor observada.

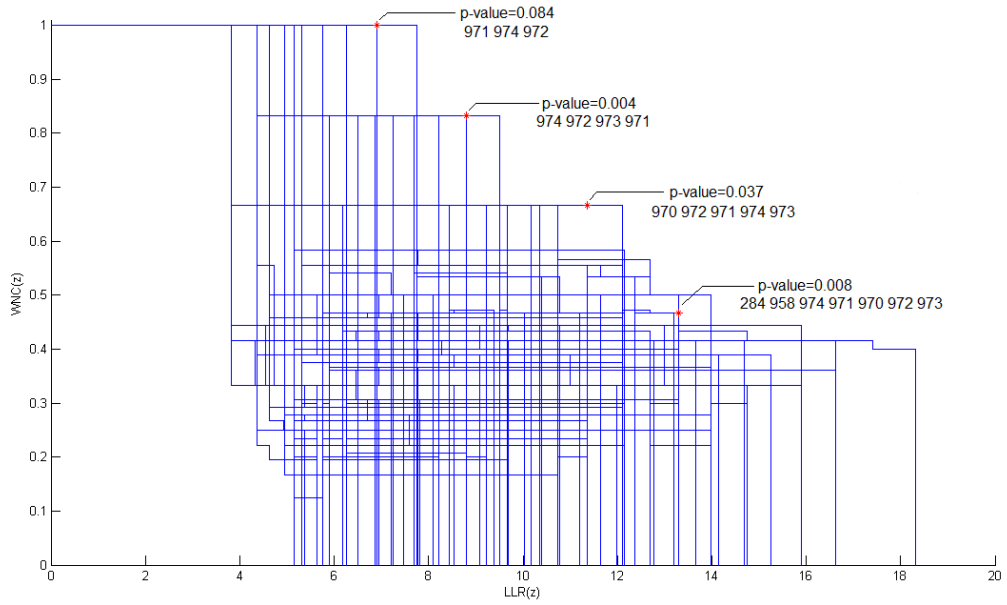

Figura 7.6: Superfícies de aproveitamento para 999 execuções do algoritmo sob Hipótese Nula e soluções para os dados observados.

Neste segundo gráfico, o número de superfícies de aproveitamento parece bem superior ao da Figura 7.5. Ainda existem algumas superfícies sobrepostas mas em número bem inferior.

Uma surpresa observada foi o fato das duas metodologias fornecerem exatamente as mesmas soluções. As Figuras 7.5 e 7.6 apresentam as mesmas soluções, mas com p-valores diferentes nos dois casos. Acredita-se que este seja um forte indicativo de que as estratégias mais complexas de estimativa populacional talvez sejam desnecessárias.

Dentre as soluções obtidas, duas aparentemente não têm grande importância prática, pois são subconjuntos de uma outra solução também significativa composta somente por casos. A solução aparentemente mais interessante dentre as encontradas, foi um conjunto de cinco pontos que são todos casos (veja Figura 7.7). Tal solução é coincidente com a solução apresentada para o mesmo benchmark de dados em Cucala [2009].

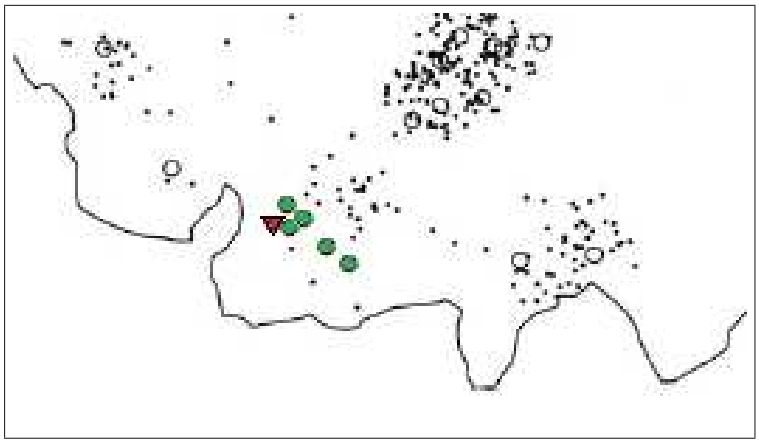

Figura 7.7: Círculos em cinza representam o cluster e o triângulo representa a incineradora.

Uma outra solução bastante peculiar utiliza os cinco pontos da solução anterior e acrescenta mais dois pontos (um caso e um controle). Tal controle aparece nesta solução como uma ponte de conexão para o sexto caso incluso na solução (veja Figura 7.8).

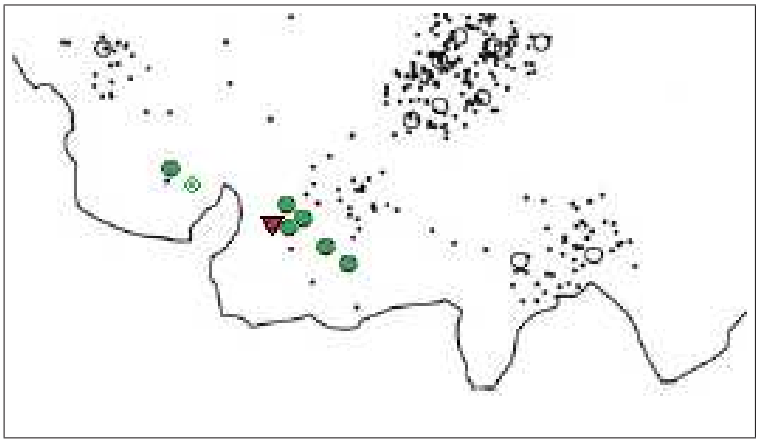

Figura 7.8: Círculos em cinza e o ponto circulado representam o cluster e o triângulo representa a incineradora.

A priori, parece difícil optar entre estas duas soluções apesar da diferença de p-valores (veja Figuras 7.5 e 7.6), afinal a segunda proposta de solução

inclui um controle. Entretanto vale ressaltar que a solução que inclui o controle é mais verossímil segundo as superfícies de aproveitamento e também apresenta um valor para a LLR superior a outra solução obtida.

A apresentação da solução coincidente com a obtida em Cucala [2009] se mostra como um indício de que a metodologia multi-objetivo anteriormente proposta para detecção de clusters em casos distribuídos por regiões parece também ser eficiente para estudo de dados caso-controle. Por outro lado a solução composta de 7 pontos (incluindo 1 controle) parece inovadora neste problema, dado que não seria proposta pelos métodos usuais.

Outra constatação importante para a confirmação da validade do método proposto pode ser obtida ao observarmos as mesmas soluções para as duas estimativas de densidades populacionais. Além disso estamos usando uma metodologia que já se encontra amplamente justificada se considerarmos os bem sucedidos resultados já observados para estudos com dados por regiões.

Também é importante fazer considerações sobre o esforço computacional para aplicação da técnica aqui apresentada. Vale ressaltar que, em um primeiro momento, o esforço necessário para a utilização de metodologias como o Scan Elíptico não são diretamente comparáveis com a técnica multi-objetivo proposta. Isto se deve ao fato de que a quantidade de candidatos a solução vasculhados pela nova metodologia é em muito superior a quantidade vasculhada pelo Scan Circular. Ainda assim, mencionamos que para o mapa de Chorley-Rilble/Inglaterra, o Scan Elíptico executa 999 rodadas de Monte-Carlo em aproximadamente 330 segundos, enquanto o método proposto executa as mesmas 999 rodadas de Monte-Carlo em aproximadamente 2200 segundos. Nos dois casos utilizando um processador *Intel Core 2 Duo* $1.86GHz$ trabalhando nos dois núcleos.

Capítulo 8

Conclusões

Os resultados já bem sucedidos de abordagens multi-objetivo, e a utilização de algoritmos genéticos para o problema de detecção de clusters espaciais, bem como todo este estudo levam a um grande arcabouço de possibilidades de medidas para avaliar a regularidade de clusters detectados. Isto faz com que diversas metodologias sejam propostas, entretanto a maioria delas para estudos de dados por regiões.

Este trabalho propõe a adequação e utilização de uma metodologia específica para o problema de detecção de clusters espaciais para o estudo com dados pontuais. Dada a natureza dos algoritmos genéticos quando aplicados a este tipo de problema, não se espera um método competitivo quanto a tempo de execução se comparado aos métodos usuais. Por razões óbvias, esta metodologia vasculha uma quantidade muito maior de candidatos a cluster solução do problema. Além disso, os métodos usuais se adaptam muito bem para problemas em que o verdadeiro cluster tenha forma muito regular, mas perdem poder de detecção em situações cujo verdadeiro cluster tenha forma irregular.

O método proposto se adapta bem a problemas nos quais o cluster ver-

dadeiro é regular e também quando é irregular. Tal conclusão se deve à natureza da medida de penalização por Não-conectividade Ponderada utilizada como segundo objetivo na avaliação das soluções. Discutindo especificamente a medida de penalização, pode-se elencar duas contribuições relevantes. A proposição de uma estrutura de vizinhança e conexidade entre os pontos do mapa através do Diagrama de Voronoi e também a discussão de duas estratégias de avaliação da estrutura de densidade populacional no mapa.

Os resultados em testes simulados com clusters artificiais servem como confirmação da aplicabilidade da metodologia proposta e levam a acreditar que estratégias mais complexas para a análise da densidade populacional sejam desnecessárias. De outra forma acredita-se que a melhor e mais simples forma de utilização do método seja considerar população 1 associada a cada ponto no mapa.

As simulações sugerem que o método tem alto poder de detecção; apresentando também ótimos resultados na avaliação de PPV. Os resultados levam ainda a uma discussão da validade da avaliação da medida de sensibilidade para este problema específico. Foi aplicada a metodologia da função de aproveitamento para estender o significado do p-valor no espaço bi-objetivo, preservando a dependência entre pontos dentro do mesmo conjunto de soluções não-dominadas. Esta aproximação dá uma definição mais robusta para o significado do conjunto de soluções obtido através da estratégia multi-objetivo.

Um estudo com dados reais também contribui para a qualidade deste trabalho e para avalizar a metodologia proposta. Isto pode ser observado através da comparação das soluções obtidas com os métodos clássicos já bastante difundidos.

Referências Bibliográficas

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

J. Bessag and J. Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society*, 154(1):143–155, 1990.

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

A L F Cançado. *Detecção de Clusters Espaciais Através de Otimização Multiobjetivo*. PhD thesis, UFMG-Brasil, Department of Electric Engineering, 2009.

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

L. Cucala. A flexible spatial scan test for case event data. *Computational Statistics and Data Analysis*, 53:2843–2850, 2009.

V G da Fonseca, C M Fonseca, and A O Hall. Inferential performance assessment of stochastic optimisers and the attainment function. In *Proceedings*

of the First International Conference on Evolutionary Multi-Criterion Optimization, volume 1993, pages 213–225, Berlin: Springer-Verlag, 2001. Lecture Notes In Computer Science.

K Deb, A Pratap, S Agrawal, and T Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6:(2):182–197, 2002.

P.J. Diggle. 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, 1990.

A R Duarte. *Geometria e Topologia de Conglomerados Espaciais Baseados em Grafos*. PhD thesis, UFMG-Brasil, Department of Statistics, 2009.

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

A R Duarte, S B Silva, L Duczmal, S J Ferreira, and A L F Cançado. Weighted non-connectivity for detection of irregular clusters. In *9th Annual Conference International Society for Disease Surveillance*, 2010b.

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

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

L Duczmal, G J P Moreira, S J Ferreira, and R H C Takahashi. Dual graph spatial cluster detection for syndromic surveillance in networks. *Advances in Disease Surveillance*, 4:88, 2007b.

L Duczmal, A L F Cançado, and R H C Takahashi. Geographic delineation of disease clusters through multi-objective optimization. *Journal of Computational & Graphical Statistics*, 17:243–262, 2008.

L Duczmal, A R Duarte, and R Tavares. Extensions of the scan statistic for the detection and inference of spatial clusters. In N Balakrishnan and J Glaz, editors, *Scan Statistics*, pages 157–182. Birkhäuser, Boston, Basel and Berlin, 2009.

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

C M Fonseca and P Fleming. An overview of evolutionary algorithms in multi-objective optimization. *Evolutionary Computation*, 3:1–16, 1995.

C M Fonseca, V G da Fonseca, and L Paquete. Exploring the performance of stochastic multiobjective optimisers with the second-order attainment function. In *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization*, volume 3410, pages 250–264, Berlin: Springer-Verlag, 2005. Lecture Notes In Computer Science.

J Glaz, J Naus, and S Wallestein. Disease mapping and risk assessment for public health. In *Springer Series in Statistics*. Springer, Berlin Heidelberg New York, 2001.

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

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

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

A Lawson. Statistical methods in spatial epidemiology. In A Lawson, editor, *Large scale: surveillance*, pages 197–206. Wiley, London, 2001.

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

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

S B Silva. *Penalização por Não-Conectividade Ponderada de Grafos*, Master thesis, UFMG-Brasil, Department of Statistics, 2010.

R H C Takahashi, J A Vasconcelos, J A Ramirez, and L Krahenbuhl. A multi-objective methodology for evaluating genetic operators. *IEEE Transactions on Magnetics*, 39:(3):1321–1324, 2003.

N Yiannakoulias, R J Rosychuk, and J Hodgson. Adaptations for finding irregularly shaped disease clusters. *International Journal of Health Geographics*, 6:28, 2005.