Spencer Barbosa da Silva

Detecção de clusters irregulares através da Não Conectividade Ponderada de Grafos

Dissertação de Mestrado apresentado 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: Sabino José Ferreira Neto
Co-orientador: Anderson Ribeiro Duarte

Universidade Federal de Minas Gerais
Belo Horizonte, Maio de 2010

Dedicatória

Dedico este trabalho à minha esposa Dani, o grande amor da minha vida, que sempre esteve comigo. Também ao melhor presente que já me foi dado, a minha florzinha Gabi.

Agradecimentos

Em primeiro lugar, agradeço à Deus por estar provendo este momento tão desejado em minha vida, sem ele não somos nada.

Agradeço também aos meus pais Francisca e Geraldo, exemplos de vida que sempre se esforçaram para manter nossa família sempre unida.

Agradeço também a Francine, Ricardo e Matheus; Thaísa e Rodrigo; Etienne e Jauber; Lívia, German e Luana; Willian, Lucélia e Mariana e a minha querida sogra Shiyauko sempre solicitos a me ajudar.

Não posso deixar de falar também dos professores Sabino, Anderson e Luiz Duczmal, por acreditarem em mim à todo instante e serem exemplo de profissionalismo. Com eles somente tive evoluções em minha carreira docente.

Quanto aos amigos, como esquecer ao longo destes anos. Ampliei minha família, acredito que o maior bem que podemos conquistar é a amizade. Sei que serei injusto ao citar nomes, provavelmente esquecerei de alguém, se isto ocorrer, por favor me perdoem. Mas não posso deixar de citar pessoas como Anderson e Hélida; Flávio dos Reis e Eva; Fabrício e Luciana; Flávio Pereira, Bárbara e Luca; Alexandre e Petrusca; Ricardo Tavares e Elen; Cleide; Ronaldo e Angélica. Estas pessoas, direta ou indiretamente contribuíram para este momento. Agradeço também a todos do Departamento de Estatística da UFMG pela oportunidade dada.

Resumo

Conglomerados (clusters) espaciais de forma irregular são difíceis de delinear. O cluster mais verossímil pode se espalhar em grandes parcelas do mapa, impactando seu significado geográfico. Métodos empregando a estatística Scan Espacial de Kulldorff, associados a medidas de penalização, foram usados para controlar a liberdade excessiva da forma dos clusters. Funções de penalidade para a geometria e o grau de conexidade dos clusters foram propostas recentemente. A medida de Não Conectividade, se mostra bastante eficaz no auxílio para a detecção, entretanto apresenta dificuldades para interpretar a importância de conexões dentro de um possível cluster. Apresentamos uma estratégia de ponderação para os termos da medida de Não Conectividade que aumenta sua eficiência para a detecção de clusters irregulares. Foram executados experimentos através de dados simulados para comprovar a melhoria de desempenho quando utilizada a versão ponderada. Os resultados de nossas simulações apresentam uma significativa melhoria em relação ao método de detecção de clusters irregulares que utilizava a versão sem ponderação. Este método se mostra muito importante no estudo epidemiológico e de vigilância sindrômica. A proposta apresenta outra vantagem importante, o baixo tempo computacional necessário para sua utilização.

Palavras-chave: vigilância sindrômica; cluster espacial; estatística Espacial Scan; clusters irregulares; algoritmos multi-objetivo; compacidade geométrica; função de não-conectividade.

Abstract

Irregularly shaped spatial clusters are difficult to delineate. The most likely cluster often spreads in a great proportion of the map, playing a significant role in its geography. Methods employing the Kulldorff's scan statistics, associated to penalization procedures were used to control the over freedom of the clusters. Penalization functions for cluster geometry and the level of the cluster's connectivity are recent proposals. The non-connectivity measurement is efficient when guiding the detection, however, it shows problems when interpreting the important role of the connections inside a possible cluster. This study presents a weighing strategy for the non-connectivity terms which maximizes its efficiency when detecting irregular clusters. Experiments using simulated data were undertaken in order to check the improvement when using the weighing version. The results show a significant improvement when compared to experiments which do not use the ponder version. This method can be very important in epidemiology studies and disease surveillance. Another important advantage of this proposal is the fact that it requires low computational time.

Keywords: disease surveillance; spatial clusters; Spatial Scan statistic; irregularly shaped clusters; multi-object algorithms; geometric compactness regularity function; non-connectivity regularity function.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Apresentação

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.

Observa-se recentemente um crescente número de trabalhos sobre metodologias para a detecção e avaliação de conglomerados (clusters) espaciais e temporais. No enfoque deste texto, um cluster é um conjunto conexo de regiões, no qual existe uma quantidade de casos discrepante para o fenômeno de interesse. O processo de detecção 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.

Um dos objetivos deste trabalho é apresentar estratégias já existentes para a detecção de clusters espaciais. Também é objetivo, propor novas metodologias visando uma melhor adequação para os diferentes cenários em que o problema se apresenta. Buscando então métodos mais eficientes no processo de avaliação e detecção de clusters, um método bastante difundido e largamente utilizado se baseia na estatística Espacial Scan, tal método é proposto por Kulldorff (1997), Kulldorff e Nagarwalla (1995).

A proposta de Kulldorff se baseia na maximização da razão de verossimilhança e utiliza uma estatística de varredura multidimensional. Este método possui três propriedades básicas: varredura geométrica da área em estudo, distribuição de probabilidade para os casos sob a hipótese de completa aleatoriedade espacial, tamanho e forma para uma janela de varredura. Uma descrição mais detalhada do método pode ser observada em seções posteriores deste texto.

Na aplicação da estatística Espacial Scan através da proposta de Kulldorff é necessário a escolha de um formato de *janela de busca*. Uma possível escolha para a forma da janela de busca é o formato circular. Tal escolha leva ao método denominado Scan Circular. A utilização deste formato de janela de busca apresenta bons resultados, mas também revela algumas deficiências. Dentre as deficiências, podemos destacar a possibilidade do método identificar um conglomerado maior ou menor que o cluster real, nas situações em que o cluster real não apresenta formato regular (por exemplo conjuntos não circulares). Seriam casos de superestimação ou subestimação no processo de detecção de clusters.

Nesta área de estudo é muito frequente a existência de clusters com formatos bastante irregulares. Em diversos problemas os clusters não regulares podem ser observados: problemas de tráfego, poluição, vigilância sindrômica, etc. Em muitos destes casos, formatos não regulares se devem às características geográficas do mapa em estudo, tais como rios, regiões litorâneas, regiões montanhosas dentre outras.

Alguns métodos foram desenvolvidos recentemente para detectar clusters de formato irregular. Ao trabalharmos com estratégias para clusters irregulares, alguns problemas podem surgir. Um primeiro problema seria a avaliação de todos os possíveis candidatos (subconjuntos conexos de regiões do mapa), visto que o número destes candidatos cresce exponencialmente a medida que o número de regiões no mapa em estudo aumenta. Um segundo problema é que na possibilidade de avaliarmos todos os candidatos, se avaliando através da razão de verossimilhanças, decorrente da proposta da estatística Espacial Scan, a solução obtida nem sempre seria uma solução viável. Isto se deve ao fato de ser possível existir soluções com alta razão de verossimilhança, entretanto dentre estes candidatos, alguns deles podem ter sido obtidos através da junção de regiões com elevado risco no mapa em estudo. Seriam então, conjuntos de regiões que estão espalhadas ao longo do mapa, abarcando grandes áreas de estudo. Este formato de solução tende a não ser muito informativo e em geral não é uma solução de interesse para o problema na prática. Dada a possibilidade de existência de tais soluções, o poder de detecção destes métodos seria reduzido.

Neste sentido, existem algoritmos que propõem estratégias para a detecção de clusters com formatos irregulares. Muitos deles são heurísticas, portanto não vasculham todas as possíveis soluções. (São analisadas apenas algumas das soluções, que seriam as mais promissoras). Ainda assim, per-

sistiria o problema de soluções não viáveis. Existem propostas de funções penalizadoras que buscam coibir a possibilidade destas soluções. Dentre as funções penalizadoras já existentes podemos citar penalizações para a regularidade da forma geométrica do cluster ou a regularidade da estrutura de conexidade do possível cluster.

Todo o trabalho que foi realizado levou a uma ampla discussão sobre os diversos métodos já existentes e também a utilização de estratégias de penalização para a estatística Espacial Scan. Esta discussão levou a proposição de uma nova função de penalização, que descreveremos ao longo deste texto. A heurística utilizada para a implementação da nova proposta será um algoritmo genético. Este algoritmo tem se mostrado bastante eficiente na solução de problemas desta natureza como pode ser visto em Duarte et al. (2010), Duczmal et al. (2008) e Duczmal et al. (2007).

1.2 Revisão Bibliográfica

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 e Carpenter (1999), Lawson (2001), Glaz et al. (2001), Balakrishnan e Koutras (2002) e Buckeridge et al. (2005). A estatística Espacial Scan definida em Kulldorff e Nagarwalla(1995) e 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 e Na-

garwalla (1995). 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. Diversos métodos para a detecção de clusters com formatos irregulares já foram discutidos no artigos a seguir [1, 2, 3, 5, 8, 11, 14, 15, 19, 17, 20, 16, 24, 29, 30, 35, 36, 39, 41, 43, 44, 46, 47, 48, 50, 53, 51, 52]. Pode-se encontrar uma interessante revisão bibliográfica deste assunto e de todos os artigos citados anteriormenente em Duczmal et al. (2009).

Capítulo 2

Métodos para detecção de clusters espaciais

Quando pensamos em avaliar todos os possíveis subconjuntos de regiões no mapa em estudo, o número de possíveis candidatos aumenta exponencialmente. Para uma mapa dividido em m regiões, existem $2^m - 1$ possíveis subconjuntos de regiões, dos quais deveríamos verificar quais são conexos. Considerando a viável 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 do mapa, 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. (2008) e a penalização por Não Conectividade apresentada em Yiannakoulias et al. (2007). Nestes trabalhos, as funções de penalização, aparecem como expoente para a razão de verossimilhança.

Uma nova abordagem utiliza um algoritmo genético multi-objetivo para o problema de detecção de clusters. Esta abordagem é apresentada por Duczmal et al. (2008). Este método conduz a uma estratégia que busca maximizar dois objetivos, 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 inferiores às 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 5, 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) e em Fonseca et al. (2005). A utilização da função de aproveitamento no problema específico de detecção de clusters se encontra bem detalhada em Cançado (2009) e em Duarte(2009). O uso da função de apro-

veitamento permite que o significado do p-valor para o espaço bi-objetivo seja mais claramente compreendido, como será descrito na seção 5.2.

Comparamos aqui três métodos para detecção de clusters através de estratégia multi-objetivo, em todos eles um dos objetivos é a estatística Espacial Scan e como segundo objetivo: a Compacidade Geométrica proposta por Duczmal et al. (2008), a Não Conectividade proposta por Yiannakoulias et al. (2007) e a nova proposta de função de penalização que denominaremos Não Conectividade Ponderada.

Neste trabalho os conceitos da estatística Espacial Scan de Kulldorff são discutidos. Posteriormente a função de compacidade geométrica e a função baseada na Não Conectividade são descritas e são revistos o algoritmo genético mono-objetivo e o algoritmo genético multi-objetivo. Como principal contribuição, é introduzida a nova função de Não Conectividade Ponderada. Finalmente é avaliado o poder de detecção para clusters espaciais com formatos irregulares através da utilização da função de aproveitamento, bem como a sensibilidade e o valor preditivo positivo (PPV) através de simulações numéricas para os três métodos.

2.1 Estatística Scan Espacial de Kulldorff

Considere o mapa associado a área de estudo A, dividido em m regiões, com população total N 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 re-

giões é distribuído conforme uma Poisson com taxa proporcional à população da respectiva região no caso da hipótese nula. 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 $I(z) = \dfrac{c_z}{\mu_z}$ enquanto o risco relativo fora da zona z, ou seja, no complementar da zona z em relação ao mapa em estudo, é dado por $O(z) = \dfrac{C - c_z}{C - \mu_z}$.

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 Kuldorff (2007)) 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 sim 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 e Nagarwalla (2005). Se considerarmos Z como o conjunto de todas as

zonas conexas, o problema se tornaria impraticável, entretanto podemos utilizar heurísticas (algorítmos 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. De acordo com Dwass (1957), 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.

2.2 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 $LLR(z)$, mas utiliza também 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, o fecho convexo será 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 Penalização por Não Conectividade

Yiannakoulias et al. (2007) 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:

$$y(z) = \frac{a(z)}{3\,(v(z) - 2)} \tag{2.3}$$

O termo $3\,(v(z) - 2)$ no denominador da expressão anterior, 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.

O cálculo da medida de Não Conectividade pode ser ilustrado através da Figura 2.1

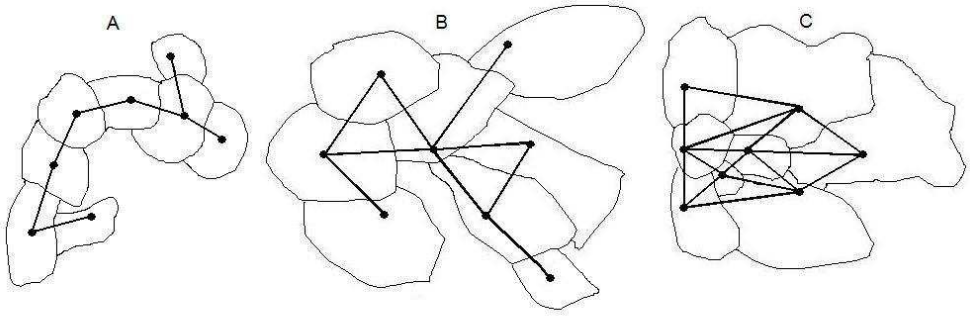

Figura 2.1: Penalização por Não Conectividade

Para cada um dos clusters hipotéticos teríamos:

Cluster A $\quad y(z) = \dfrac{7}{3(8-2)} = 0,389$

Cluster B $\quad y(z) = \dfrac{9}{3(8-2)} = 0,500$

Cluster C $\quad y(z) = \dfrac{15}{3(8-2)} = 0,833$

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. (2007) 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 indivíduo 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*, *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 caracteristicas de A e B escolhidos mantendo a conexidade nas zonas geradas, veja a Figura 3.1.

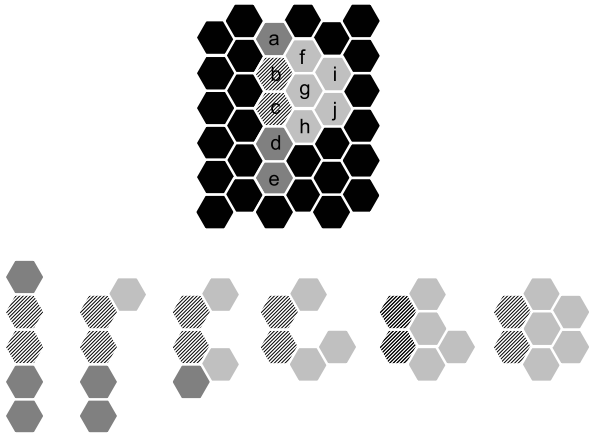

Figura 3.1: Cruzamento entre zonas $A = \{a, b, c, d, e\}$ e $B = \{b, c, f, g, h, i, j\}$

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 a 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 e Fleming (1995), Takahashi et al. (2003). Em Duczmal et al. (2008) é proposta a utilização da Penalização por Compacidade Geométrica através de uma proposição Multi-Objetivo. Nesta nova 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 por Duarte et al. (2009). 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 eficiente 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 os operadores de cruzamento e de mutação são idênticos àquelas usadas no algoritmo genético mono-objetivo (veja Duczmal et al. (2007) para uma descrição detalhada daqueles operadores).

Observe um descrição simples do algoritmo utilizado:

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

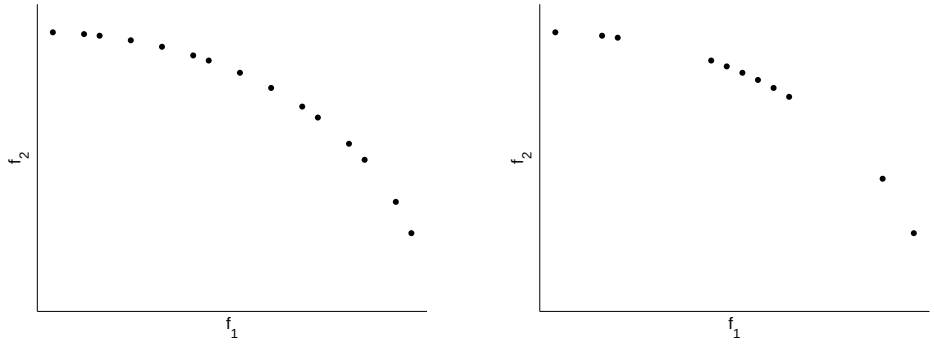

Figura 3.2: Distância de aglomeração

A Figura 3.2, apresenta à esquerda um conjunto de soluções não dominadas uniformente distribuido (alta distância de aglomeração em seus pontos) e à direita um conjunto de soluções não dominadas não uniformente distribuido (baixa 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

Penalização por Não Conectividade Ponderada

Neste trabalho, utilizaremos o Algoritmo Genético Multi-Objetivo citado anteriormente. Entretanto não mais com as propostas de funções penalizadoras descritas anteriormente, mas sim com uma nova formulação de função penalizadora, a Função de penalização por Não Conectividade ponderada de Grafos.

4.1 Ponderação de arestas e vértices associados ao grafo de um mapa em estudo

A função de penalização por Não Conectividade apresentada por Yiannakoulias et al. (2007) se mostra bastante eficiente na detecção e inferência de clusters, quando avaliadas medidas de poder, sensibilidade e PPV 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.

Em outras palavras, estamos interessados em perguntar se existem arestas mais ou menos relevantes para a conexidade do grafo.

Pensando apenas na análise do grafo é fato que tal relevância não precisa ser considerada. Quando observamos que estamos trabalhando com subgrafos associados à zonas em um mapa, lembramos que as arestas são conexões de vizinhança entre regiões que podem ser muito ou pouco populosas. Neste contexto, observamos 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, estabeleceremos uma ponderação para os vértices e arestas do subgrafo associado à zona z. Tal ponderação será construída pensando na estrutura da distribuição populacional ao longo das regiões da zona z.

A Figura 4.1 apresenta uma zona hipotética que pode ser utilizada para compreendermos melhor a estrutura de ponderação para os vértices e as arestas.

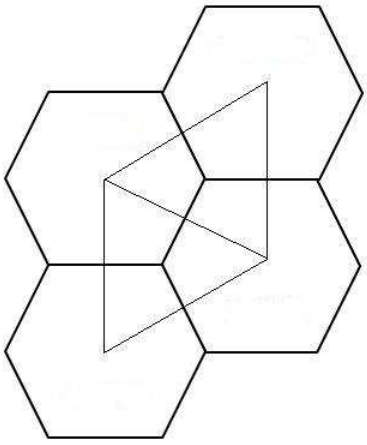

Figura 4.1: Zona hipotética

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

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

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

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

Se considerarmos na Figura 4.1 as regiões por $R1$, $R2$, $R3$ e $R4$, podemos construir diferentes cenários de distribuição populacional e então verificar a poderação das arestas e dos vértices para diferentes distribuições populacionais.

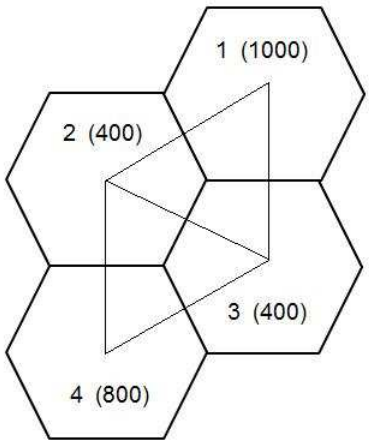

Figura 4.2: Ponderação de arestas e vértices

Um destes possíveis cenários é ilustrado pela Figura 4.2. No qual temos as populações das regiões sendo: $pop(R1) = 1000$, $pop(R2) = 400$, $pop(R3) = 400$ e $pop(R4) = 800$. Teríamos então os seguintes pesos para as arestas:

$$P(a_{1,2}) = \frac{1000 + 400}{2} = 700, \qquad P(a_{1,3}) = \frac{1000 + 400}{2} = 700,$$

$$P(a_{2,3}) = \frac{400 + 400}{2} = 400, \qquad P(a_{2,4}) = \frac{400 + 800}{2} = 600,$$

$$P(a_{3,4}) = \frac{400 + 800}{2} = 600.$$

Já para os vértices os seguintes pesos:

$$P(v_1) = 1000, \qquad P(v_2) = 400, \qquad P(v_3) = 400, \qquad P(v_4) = 800.$$

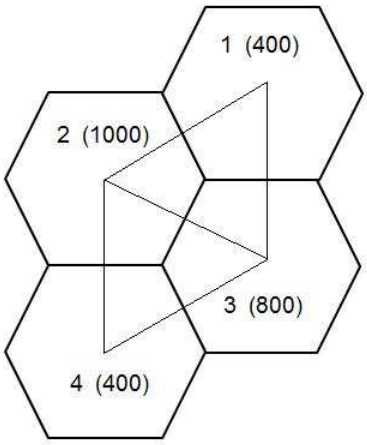

Figura 4.3: Ponderação de arestas e vértices

Utilizando o mesmo formato de zona, quando consideramos um outro cenário de distribuição populacional, os pesos associados às arestas ficam modificados, isto pode ser observado na Figura 4.3.

Considere neste novo cenário as populações $pop(R1) = 400$, $pop(R2) = 1000$, $pop(R3) = 800$ e $pop(R4) = 400$ teríamos então os seguintes pesos para as arestas:

$$P(a_{1,2}) = \frac{400 + 1000}{2} = 700, \qquad P(a_{1,3}) = \frac{400 + 800}{2} = 600,$$

$$P(a_{2,3}) = \frac{1000 + 800}{2} = 900, \qquad P(a_{2,4}) = \frac{1000 + 400}{2} = 700,$$

$$P(a_{3,4}) = \frac{800 + 400}{2} = 600.$$

Para os vértices os seguintes pesos:

$$P(v_1) = 400, \qquad P(v_2) = 1000, \qquad P(v_3) = 800, \qquad P(v_4) = 400.$$

4.2 Função de penalização por Não Conectividade ponderada de Grafos

A medida de penalização por Não Conectividade proposta por Yiannakoulias et al. (2007)[51] é dada por:

$$y(z) = \frac{a(z)}{3\left(v(z) - 2\right)}$$

Para reformular a função descrita, susbtituiremos as arestas e vértices por

seus respectivos pesos ponderadores da seguinte forma:

$$yp(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} P(a_{i,j})}{3\left[\sum_{i=1}^{k} P(v_i) - 2\left(\frac{\sum_{i=1}^{k} P(v_i)}{k}\right)\right]} \tag{4.1}$$

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

Alguma dúvida pode pairar sobre o termo $\frac{\sum_{i=1}^{k} P(v_i)}{k}$ associado ao valor 2 no denominador, entretanto se pensarmos na suposição de todas as populações identicas nas regiões da zona a ser avaliada, se faz necessário este termo para que tenhamos $y(z) = yp(z)$ para esta situação específica.

Com este novo formato estaremos levando 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.

Voltando aos exemplos descritos através das Figuras 4.2 e 4.3 podemos observar o efetivo efeito da utilização deste formato de função de penalização. No exemplo apresentado através da Figura 4.2 as regiões mais populosas (R1 e R4) não estão conectadas, ou seja, não são vizinhas. Já no exemplo da Figura 4.3 as regiões mais populosas (R2 e R3) estão conectadas, ou seja, são vizinhas. A motivação deste formato de penalização fica evidênciada nesta situação. Em outras palavras, nossa suposição se baseia no fato de acreditarmos que vizinhanças de regiões mais populosas devem gerar maior movimentação entre habitantes e portanto tendendo a reforçar a proliferação do fenômeno de interesse. Neste caso estaríamos reforçando o cluster em

questão.

Note que utilizando a medida de Não Conectividade proposta anteriormente, os exemplos das Figuras 4.2 e 4.3 apresentariam a mesma medida de Não Conectividade que neste caso seria:

$$y(z) = \frac{5}{3\,(4-2)} = 0.833$$

Agora utilizando a medida de Não Conectividade Ponderada, as medidas seriam diferentes em cada um dos exemplos.

No exemplo da Figura 4.2 teriamos:

$$yp(z) = \frac{700 + 700 + 400 + 600 + 600}{3\left(1000 + 400 + 400 + 800 - 2\left(\frac{2600}{4}\right)\right)}$$

$$yp(z) = 0.769$$

No exemplo da Figura 4.3 teriamos:

$$yp(z) = \frac{700 + 600 + 900 + 700 + 600}{3\left(400 + 1000 + 800 + 400 - 2\left(\frac{2600}{4}\right)\right)}$$

$$yp(z) = 0.897$$

confirmando então o objetivo da proposição da nova função de Não Conectividade Ponderada.

Capítulo 5

Inferência e Resultados

5.1 Significância Estatística

Construímos então um algoritmo genético multi-objetivo utilizando como funções objetivo a $LLR(z)$ e a nova medida de não Conectividade Ponderada $yp(z)$. 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 aprozimação do conjunto de Pareto. Buscamos então uma estratégia para verificar para cada solução deste conjunto de soluções não dominadas sua significância estatística.

De forma similar ao procedimento de Dwass (1957) já mencionado anteriormente, através de simulações de Monte Carlo. 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).

5.2 Função de Aproveitamento

Para cada uma 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 simultâneamente; já algum ponto que se situasse na região R_0 este seria 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 5.1).

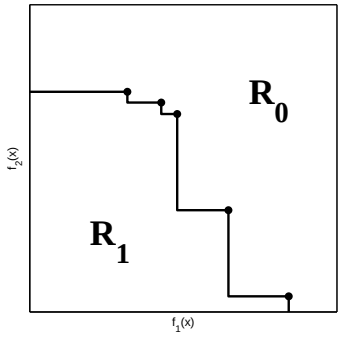

Figura 5.1: Superfície de aproveitamento dividindo o espaço de objetivos

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.

Podemos repetir a execução do algoritmo para n alocações distintas de casos no mapa, obtidas de cada uma 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 5.2).

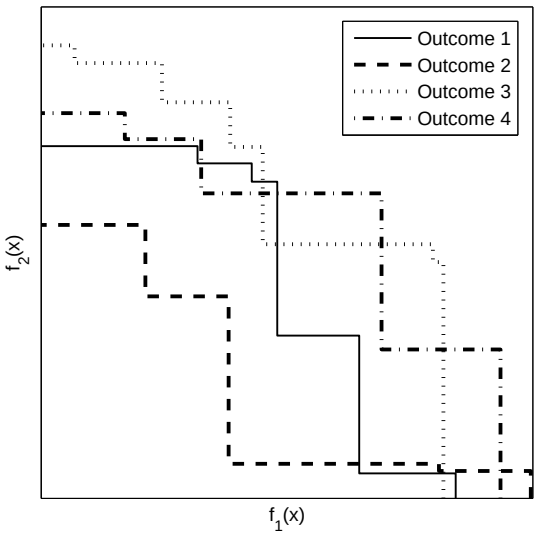

Figura 5.2: Limites obtidos para diferentes execuções do algoritmo

O conjunto dos n limites pode ser utilizado para dividir o espaço de objetivos em $n + 1$ regiões (veja Figura 5.3).

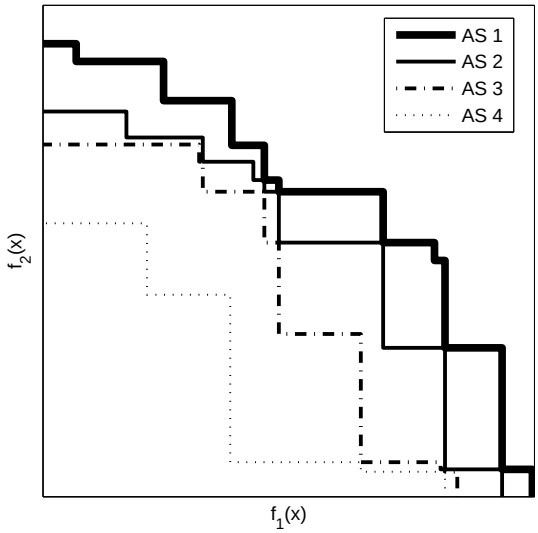

Figura 5.3: Superfícies de aproveitamento para n execuções do algoritmo

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 encontraremos 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.

5.3 Poder, Sensibilidade e PPV

Para avaliar a qualidade do método para detecção e inferência de clusters aqui proposto, precisamos de uma estratégia para avaliar o poder de detecção dos métodos citados. Serão produzidos clusters artificiais sobre o mapa, 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.

Inicialmente, uma quantidade pré-estabelecida de casos é distribuída no mapa de acordo com a distribuição de Poisson, considerando que o número de casos esperado em cada uma das regiões do mapa é proporcional à sua população. Esta distribuição satisfaz a hipótese nula de não existência de cluster no mapa em estudo.

Posteriormente, para cada uma das hipóteses alternativas, a mesma quantidade pré estabelecida de casos na hipótese nula é distribuída aleatoriamente no mapa de acordo com uma distribuição de Poisson. 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. 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 segundo Kulldorff et al. (2003).

Para o modelo da hipótese nula, 10000 execuções do algoritmo são realizadas. Podemos então utilizar o procedimento da *Função de Aproveitamento*

já citada anteriormente e produzir então uma superfície de aproveitamento para algum nível de significância específico, em geral utilizamos $\alpha = 0.05$.

Dado um modelo da hipótese alternativa, 5000 execuções do algoritmo são realizadas, produzindo então 5000 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) igualmente servem para avaliar a qualidade do processo da detecção de clusters. Estas medidas são probabilidades condicionais definidas a partir dos seguintes eventos:

V = Indivíduo escolhido ao acaso na população do mapa pertence a população do cluster verdadeiro;

D = Indivíduo escolhido ao acaso na população do mapa pertence a população do cluster detectado;

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

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

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

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

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 subestimação no processo de detecção é reduzida, enquanto altas medidas de sensibilidade significam que a chance de superestimação no processo de detecção é reduzida.

É importante verificar que considerando a prevalência para a doença em estudo, que é dada pela razão entre o número de casos observados e a população no mapa em estudo, a medida de PPV se altera. Aumento na prevâlencia acarreta em aumento na medida de PPV. Por outro lado a medida de sensibilidade para o teste não é impactada por alterações na prevalência da doença em estudo.

As medidas de PPV e Sensibilidade começaram a ser utilizadas para atestar a qualidade dos métodos de detecção de clusters espaciais com grande frequência nos últimos anos. Muitos trabalhos desta área apresentados no últimos quatro anos fazem uso de tais medidas.

5.4 Avaliações Numéricas

Utilizamos um benchmark (Duczmal et al. (2006)) com dados para uma população real com casos de câncer de mama no nordeste dos Estados Unidos. Este conjunto de dados consiste em 245 condados em 10 estados e no distrito de Columbia, com uma população de risco totalizando 29.535.210, referentes ao período de 1988 até 1992, considerando a população de 1990. A distribuição populacional no mapa pode ser observada na figura 5.4.

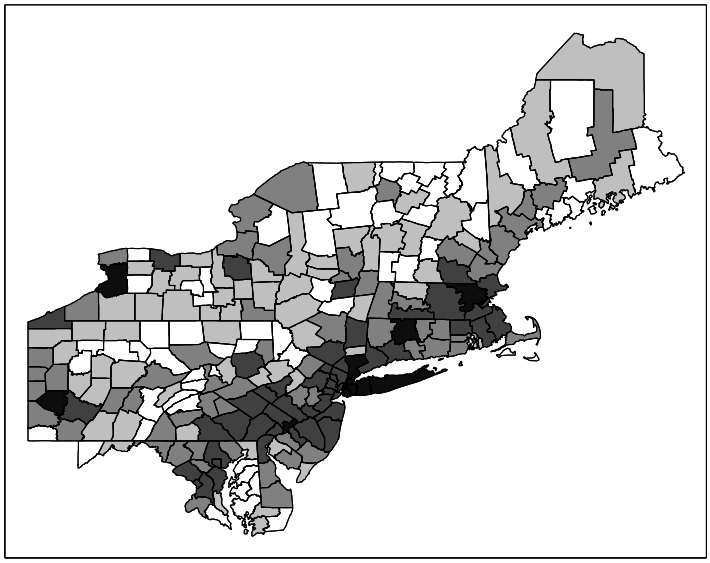

Figura 5.4: Distribuição populacional no nordeste dos EUA

Sobre este mapa foram construídos nove clusters artificiais A-F, NY, BOS e DC, seguindo os procedimentos já citados (veja Figuras 5.5, 5.6 e 5.7). Informações detalhadas sobre estes dados e os clusters artificiais podem ser obtidas em *http://www.est.ufmg.br/∼duczmal/(2010-05)* e em Duczmal et al. (2006).

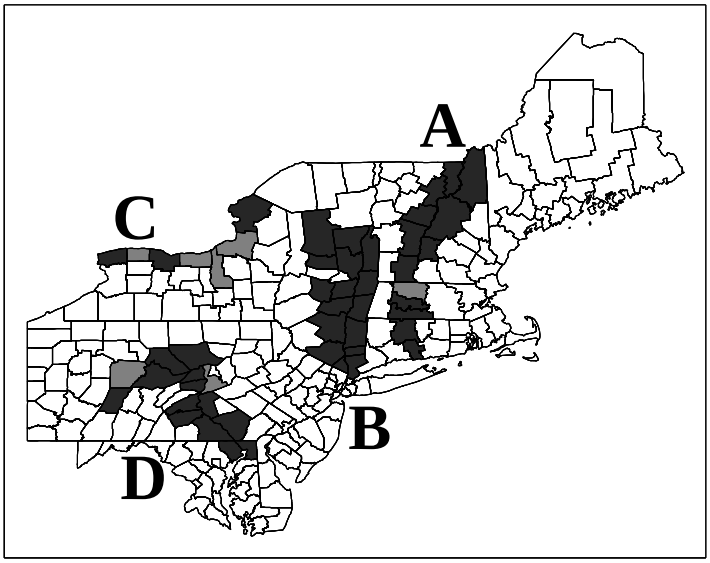

Figura 5.5: Clusters artificiais A–D no nordeste dos EUA

Estes clusters foram definidos também levando em conta características geográficas tais como rios, regiões litorâneas entre outras.

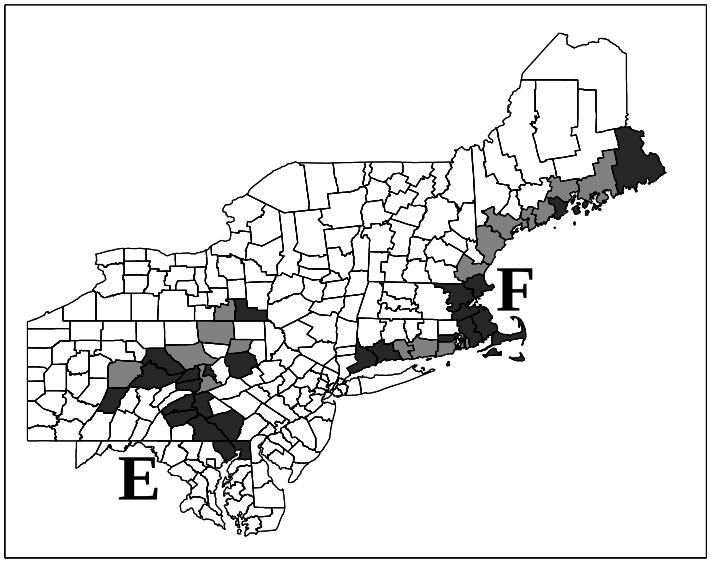

Figura 5.6: Clusters artificiais E, F no nordeste dos EUA

Estes clusters foram escolhidos com a finalidade de testar os limites dos algoritmos para algumas formas muito irregulares de possíveis clusters.

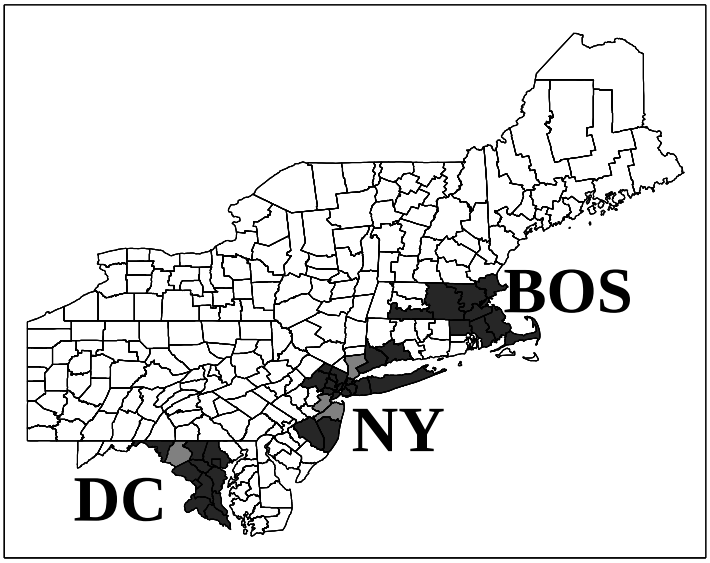

Figura 5.7: Clusters artificiais NYC, BOS e DC no nordeste dos EUA

Os clusters NY, BOS e DC estão situados em áreas altamente povoadas, contrastando com os clusters restantes, que estão situados em áreas rurais ou então misturando áreas rurais com áreas mais povoadas.

Deste momento em diante, os nove clusters artificiais, já mencionados, serão tratados como *clusters reais*. Iremos comparar inicialmente os algoritmos genéticos multi-objetivo ora utilizando a medida de Não Conectividade, ora utilizando a medida de Não Conectividade Ponderada.

Para cada simulação dos dados inicialmente sob a hipótese nula, 600 casos são distribuídos aleatoriamente de acordo com um modelo de Poisson, visando obter nossa superfície de aproveitamento. Posteriormente, consideramos cada um dos clusters reais como uma hipótese alternativa, novamente

600 casos são distribuídos aleatoriamente mas agora com um ajuste para considerar o risco relativo igual a um nas regiões não pertencentes ao cluster real, e maior que um, mas idêntico, para todas as regiões pertencentes ao cluster real, de forma que sendo conhecida a posição exata do cluster real, o poder de detecção seja de 0.999 segundo Kulldorff et al. (2003). A comparação entre os algoritmos forneceu os resultados apresentados na tabela 5.1.

Tabela 5.1: Comparação entre Não Conectividade e Não Conectividade Ponderada

	Power		PPV		Sensitivity	
cluster	AG1	AG2	AG1	AG2	AG1	AG2
A	0.939	**0.957**	**0.806**	0.775	0.843	**0.875**
B	**0.967**	0.959	0.906	0.906	**0.857**	0.847
C	0.910	**0.935**	0.778	**0.807**	0.855	**0.884**
D	0.963	**0.969**	0.891	**0.895**	0.769	**0.820**
E	0.942	**0.954**	0.876	**0.883**	0.596	**0.632**
F	0.733	**0.851**	**0.777**	0.767	0.593	**0.619**
NY	**0.906**	0.898	0.973	**0.976**	0.743	**0.758**
BOS	0.924	**0.952**	0.896	**0.900**	0.873	**0.919**
DC	0.933	**0.937**	0.927	**0.948**	0.874	**0.892**

- AG 1 - AG multi-objetivo utilizando $LLR(z)$ e $y(z)$ - Não Conectividade;
- AG 2 - AG multi-objetivo utilizando $LLR(z)$ e $yp(z)$ - Não Conectividade Ponderada.

Em praticamente todos os clusters analisados, verificou-se um melhor poder de detecção quando utilizada a nova metodologia. As medidas de PPV e sensibilidade também apresentaram ganhos na maioria dos casos.

Este resultado vem confirmar a suposição inicial de que realmente os vértices e as arestas de um subgrafo associado a uma zona do mapa em estudo tem efeitos distintos na estrutura de conectividade da zona.

Outra comparação também importante pode ser feita se levarmos em conta o algoritmo multi-objetivo que utiliza a medida de Compacidade Geométrica e o algoritmo genético utilizando a Não Conectividade Ponderada. A comparação destes algoritmos forneceu os resultados apresentados na tabela 5.2.

Tabela 5.2: Comparação entre Compacidade Geométrica e Não Conectividade Ponderada

	Power		PPV		Sensitivity	
cluster	AG1	AG2	AG1	AG2	AG1	AG2
A	0.950	**0.957**	**0.902**	0.775	0.827	**0.875**
B	0.954	**0.959**	0.895	**0.906**	0.801	**0.847**
C	0.934	**0.935**	**0.813**	0.807	0.860	**0.884**
D	0.962	**0.969**	0.860	**0.895**	0.740	**0.820**
E	0.947	**0.954**	0.868	**0.883**	0.609	**0.632**
F	0.752	**0.851**	**0.796**	0.767	0.583	**0.619**
NY	0.891	**0.898**	0.961	**0.976**	0.689	**0.758**
BOS	0.918	**0.952**	**0.939**	0.900	0.837	**0.919**
DC	**0.955**	0.937	**0.977**	0.948	0.880	**0.892**

- AG 1 - AG multi-objetivo utilizando $LLR(z)$ e $k(z)$ - Compacidade Geométrica;
- AG 2 - AG multi-objetivo utilizando $LLR(z)$ e $yp(z)$ - Não Conectividade Ponderada;

Novamente na maioria dos clusters analisados, verificou-se um ganho em poder de detecção quando utilizada a nova metodologia (houve perda somente no cluster DC). O caso particular do cluster F, que se mostra bastante irregular quanto à forma geométrica, apresentou um ganho expressivo no poder de detecção. Avaliando a medida de sensibilidade, o algoritmo genético multiobjetivo utilizando a Não Conectividade Ponderada foi superior em todos os casos avaliados. Já quanto a medida de PPV foi verificado um equilíbrio, em alguns clusters artificiais a nova metodologia foi superior enquanto em outros se mostrou inferior quando comparada à Compacidade Geométrica. Este resultado serve para confirmar as contribuições obtidas através da utilização da nova metodologia de detecção de clusters.

As comparações feitas até aqui buscam avaliar a qualidade no trabalho de detecção para cada um destes algoritmos, entretanto não menos importante são considerações a respeito da velocidade de execução de cada um destes algoritmos. A estrutura da penalização por Não conectividade já sugere que o tempo de execução deste algoritmo seja inferior aos demais. Uma constatação importante vem do fato de que não ocorre perda excessiva em tempo de processamento quando utilizada a função de Não Conectividade Ponderada.

Utilizando um processador *Intel(R) Core(TM)2 Duo* com dois núcleos de *1.66 GHz* em um equipamento com *2 GB* de memória *RAM*, mas apenas um núcleo estava sendo utilizado. Nas execuçõs sob a hipótese nula para o mapa do nordeste dos Estados Unidos já citado, para 100 execuções, o algoritmo com a função de Não Conectividade realizou as execuções em 54 segundos, já com a função de Não Conectividade Ponderada as execuções levaram 60 segundos, enquanto com a função de Compacidade Geométrica as execuções levaram 220 segundos.

Capítulo 6

Conclusões

Comparamos as estratégias multi-objetivo que utilizam a Penalização por Compacidade Geométrica, a Penalização por Não Conectividade e a nova Penalização por Não Conectividade Ponderada, para a detecção e a inferência de clusters espaciais. Empregamos como primeiro objetivo a estatística espacial Scan de Kulldorff e como segundo objetivo alguma das funções de penalização citadas. Foi então utilizado um algoritmo genético multi-objetivo. Estas funções de regularidade avaliam um cluster potencial, nos termos de sua forma geométrica, ou através de sua estrutura topológica, visando controlar a liberdade excessiva na forma dos clusters. A nova função de regularidade é comparada com duas funções precedentes para regularidade, a Compacidade Geométrica e a Não Conectividade. A nova função avalia a estrutura de conectividade do grafo associado ao cluster, mas leva em consideração também a estrutura populacional deste cluster.

Os algoritmos genéticos multi-objetivo maximizam dois objetivos: a estatística espacial Scan e uma função de regularidade para o formato do cluster. Para as três funções de regularidade aqui utilizadas foram avaliados seu poder de detecção, a sensibilidade e o PPV. Estas medidas foram comparadas para

os três formatos. Nossas simulações sugerem que dentre os três algoritmos, o poder da deteção para a Não Conectividade Ponderada é mais elevado para a maioria das situações; esta estratégia também apresenta melhor desempenho quanto a sensibilidade; a Compacidade Geométrica e a nova proposta se mostram equilibradas quanto à avaliação de PPV. A estratégia utilizando a Não Conectividade é significativamente mais rápida que quando utilizamos a Compacidade Geométrica, e a nova estratégia utilizando a Não Conectividade Ponderada não apresenta perda efetiva quanto ao tempo de execução, ou seja, preserva a rapidez já observada na utilização da medida de Não Conectividade.

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.

É objetivo futuro avaliar a nova penalização utilizando como pesos ponderadores a densidade demográfica associada a cada uma das regiões, bem como um estudo com a utilização da nova metodologia para uma base de dados reais. Também uma proposta de continuidade dos estudos produzindo novas funções de regularidade ou utilizando outras estratégias de otimização.

Referências Bibliográficas

[1] Agarwal, D., McGregor, A., Venkatasubramanian, S. and Zhu, Z (2006). Spatial Scan Statistics Approximations and Performance Study, *Conference on Knowledge Discovery in Data Mining 2006*.

[2] Aldstadt, J. and Getis, A. (2006). Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters, *Geographical Analysis*, **38**, 327–343.

[3] Assunção, R.M., Costa, M.A., Tavares, A., Neto, S.J.F. (2006). Fast detection of arbitrarily shaped disease clusters, *Statistics in Medicine*, **25**, 723–742.

[4] Balakrishnan, N. and Koutras, M.V. (2002). *Runs and Scans with Applications*, John Wiley & Sons, New York.

[5] Boscoe, F.P. (2003). Visualization of the spatial scan statistic using nested circles, *Health & Place*, **9**, 273–277.

[6] 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.

[7] Cançado, A.L.F. (2009). Doctor thesis: *Detecção de Clusters Espaciais*

Através de Otimização Multiobjetivo, *Department of Electric Engineering - UFMG, Brasil*.

[8] Conley, J., Gahegan, M. and Macgill, J. (2005). A Genetic Approach to Detecting Clusters in Point Data Sets, *Geographical Analysis*, **37**, 286–314.

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

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

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

[12] Duarte, A.R., Duczmal, L., Ferreira, S.J. and Cançado, A.L.F. (2008). Optimizing Simultaneously the Geometry and the Internal Cohesion of Clusters, *Advances in Disease Surveillance*, **5**, 27.

[13] Duarte, A.R., Duczmal, L., Ferreira, S.J. and Cançado, A.L.F. (2010). Internal cohesion and geometric shape of spatial clusters, 1–19 (to appear in *Environmental and Ecological Statistics*).

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

[15] Duczmal, L. and Buckeridge, D.L. (2006). A Workflow Spatial Scan Statistic, *Statistics in Medicine*, **25**, 743–754.

[16] Duczmal, L., Cançado, A.L.F. and Takahashi, R.H.C. (2008). Geographic Delineation of Disease Clusters through Multi-Objective Optimization, *Journal of Computational & Graphical Statistics*, **17**, 243–262.

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

[18] Duczmal, L., Duarte, A.R. and Tavares, R. (2009). Extensions of the scan statistic for the detection and inference of spatial clusters, In *Scan Statistics,*Glaz J., Pozydnyakov V., and Wallestein S. (eds). Birkhäuser, **157–182** *Birkhäuser, Boston*, **2009**.

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

[20] Duczmal, L., Moreira, G.J.P., Ferreira, S.J. and Takahashi, R.H.C. (2007). Dual Graph Spatial Cluster Detection for Syndromic Surveillance in Networks, *Advances in Disease Surveillance*, **4**, 88.

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

[22] Fonseca, C.M., and Fleming, P. (1995). An Overview of Evolutionary Algorithms in Multiobjective Optimization, *Evolutionary Computation*, **3**: 1–16.

[23] Fonseca, C. M., da Fonseca, V. G. and Paquete, L. (2005).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,* Lecture Notes In Computer Science, vol. 3410. Berlin: Springer-Verlag; 250–264.

[24] Gaudart, J., Poudiougou, B., Ranque, S. and Doumbo, O. (2005). Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk, *BMC Medical Research Methodology*, **5**, 22.

[25] Glaz, J. and Zhang, Z. (2006). Maximum scan score-type statistics, *Statistics & Probability Letters*, **76**, 1316–1322.

[26] Glaz, J., Naus, J., and Wallestein, S. (2001). *Scan Statistics In Springer Series in Statistics*, Springer, Berlin Heidelberg New York.

[27] Huang, L., Kulldorff M. and Gregorio D. (2007). A Spatial Scan Statistic for Survival Data, *Biometrics*, **63**, 109–118.

[28] Huang, L., Pickle, L.W., Stinchcomb, D. and Feuer, E.J. (2007). Detection of Spatial Clusters: Application to Cancer Survival as a Continuous Outcome, *Epidemiology*, **18**, 73–87.

[29] Iyengar, V.S. (2004). Space-time Clusters with flexible shapes, *IBM Research Report RC23398 (W0408-068).*

[30] Jacquez, G.M., Kaufmann, A. and Goovaerts, P. (2007). Boundaries, links and clusters: a new paradigm in spatial analysis?, *Environmental and Ecological Statistics*, (Published online).

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

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

[33] Kulldorff, M. (1999). Spatial Scan Statistics: Models, Calculations, and Applications, In *Scan Statistics and Applications* (Ed., N. Balakrishnan and J. Glaz), pp. 303–322, Birkhäuser.

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

[35] Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic, *Statistics in Medicine*, **25**, 3929–3943.

[36] Kulldorff, M., Mostashari, F., Duczmal, L., Yih, K., Kleinman, K. and Platt, R. (2007). Multivariate Scan Statistics for Disease Surveillance, *Statistics in Medicine*, **26**, 1824–1833.

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

[38] Lawson, A. (2001). Statistical methods in spatial epidemiology, In *Large scale: surveillance* (Ed., A. Lawson), pp. 197–206, Wiley.

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

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

[41] Moura, F.R., Duczmal, L., Tavares, R. and Takahashi, R.H.C. (2007). Exploring Multi-cluster structures with the Multi-objevtive Circular Scan, *Advances in Disease Surveillance*, **2**, 48.

[42] Naus, J.I.(1965). Clustering of Random Points in Two Dimensions, *Biometrika*, **52**, 263–267.

[43] Neill, D.B., Moore, A.W., Pereira, F. and Mitchell, T. (2005). Detecting Significant Multidimensional Spatial Clusters, *Advances in Neural Information Processing Systems*, **17** 969–976.

[44] Neill, D.B., Moore, A.W. and Cooper, G.E. (2007). A multivariate Bayesian scan statistic, *Advances in Disease Surveillance*, **2**, 60.

[45] Neill, D.B. and Lingwall, J. (2007). A Nonparametric Scan Statistic for Multivariate Disease Surveillance, *Advances in Disease Surveillance*, **4**, 106.

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

[47] Patil, G.P., Modarres, R., Myers, W.L. and Patankar, P. (2006). Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics, *Environmental and Ecological Statistics*, **13**, 365–377.

[48] Sahajpal, R., Ramaraju, G.V. and Bhatt, V. (2004). Applying niching genetic algorithms for multiple cluster discovery in spatial analysis, *International Conference on Intelligent Sensing and Information Processing.*

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

[50] Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11.

[51] Yiannakoulias, N., Rosychuk, R.J. and Hodgson, J. (2007). Adaptations for finding irregularly shaped disease clusters, *International Journal of Health Geographics*, **6**, 28.

[52] Yiannakoulias, N., Karosas, A., Schopflocher, D.P., Svenson, L.W. and Hodgson, M.J. (2007). Using quad trees to generate grid points for application in geographic disease surveillance, *Advances in Disease Surveillance*, **3**.

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