Anderson Ribeiro Duarte

Geometria e Topologia de Conglomerados Espaciais Baseados em Grafos

Tese de doutorado apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Estatística.

Orientador: Luiz Henrique Duczmal
Co-orientador: Sabino José Ferreira Neto

Universidade Federal de Minas Gerais
Belo Horizonte, Setembro de 2009

Anderson Ribeiro Duarte

Geometry and Topology of Graph Based Spatial Clusters

Agradecimentos

Inicialmente eu agradeço à Deus, pois ele é a razão superior de estarmos aqui e de conseguirmos sucesso em qualquer realização proposta.

Apesar de não existirem palavras suficientes para agradecer a minha amada esposa Hélida, deixo aqui também um enorme sinal de satisfação e alegria para a contribuição e compreensão sem tamanho que ela sempre prestou.

Agradeço também à meus pais Marcos e Ana, meu irmão Christian e sua esposa pela compreesão nos grandes momentos de ausência para realização deste trabalho. Faço menção também aos meus sobrinhos Bárbara e Bernardo que sempre foram uma grande fonte de motivação.

Também fico sem palavras para tentar agradecer a enorme ajuda e os conhecimentos sem limites oferecidos pelos meus orientadores Luiz Duczmal e Sabino Ferreira, além dos agradecimentos quanto a contribuição científica, agradeço a amizade dos dois e também as boas indicações de vinhos para as horas de descanso prestadas por Sabino Ferreira.

Quanto aos amigos que também contribuiram muito nesta etapa da minha vida, sei que ao tentar citar nomes, certamente serei injusto e me esquecerei de alguns deles, portanto espero que os não citados também se sintam agradecidos. Entretanto não posso deixar de citar pessoas como: André Cançado; Fábio Demarqui; Cristiano, Érika e Anne; Glaysson e Alessandra; Flávio dos

Reis e Eva; Cristiano dos Reis; Ricardo Tavares e Elen; Spencer Barbosa, Daniela e Gabriela; Ricardo e Carol; Léo e Cláudia; Airton, Vane, Flávia, Ronaldo, Ester e Rodrigo; os amigos do Clube dos Anjos, pois todos estes, com certeza, sempre foram grandes fontes de incentivo e ajuda.

Agradeço também a todos do Departamento de Estatística da UFMG que de alguma forma contribuiram para o sucesso desta empreitada.

Resumo

Conglomerados (clusters) espaciais de forma irregular são difíceis de delinear. 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. Funções de penalidade baseadas na geometria dos clusters e na não-conectividade foram propostas recentemente. Uma outra estratégia envolveu o uso de um algoritmo multi-objetivo para maximizar dois objetivos: a estatística Scan espacial de Kulldorff e a função de penalização geométrica. São apresentados dois novos algoritmos multi-objetivo utilizando a estatística Scan espacial de Kulldorff: o primeiro algoritmo emprega uma função baseada na topologia do gráfico, visando penalizar a presença de nodos de desconexão com população baixa no cluster candidato. Um *nodo de desconexão* é definido como uma região dentro de um cluster, tal que sua remoção desconecta o cluster; o segundo algoritmo maximiza, simultaneamente, a *função de penalização geométrica* e a *função coesão* para nodos de desconexão. 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. Nosso método distingue claramente aqueles clusters geograficamente inadequados que são piores do ponto de vista geométrico e os que são piores do ponto de vista topológico. Adicionalmente, a irregularidade da forma geométrica é permitida desde que não impacte a regularidade topológica. Nosso método tem o melhor poder da deteção para os conjuntos que satisfazem àquelas exigências. Propomos uma definição mais robusta do cluster espacial usando estes conceitos. Uma teoria estatística é apresentada para avaliar o significado estatístico das soluções obtidas através do algoritmo multi-objetivo que emprega o conceito de *funções de aproveitamento*. Neste trabalho nós comparamos diferentes métodos com a estatística espacial Scan, nos quais empregamos a penalização geométrica, a penalização por não-conectividade e a penalização dos nodos de desconexão. Também construímos algoritmos multi-objetivo que empregam funções de penalização e os comparamos com os algoritmos mono-objetivo anteriores (penalizados). Mostramos que os algoritmos multi-objetivo apresentam melhor desempenho, comparados aos algoritmos mono-objetivo, considerando o poder, a sensibilidade e o valor preditivo positivo. Uma aplicação dos algoritmos é apresentada usando dados reais para doença de Chagas em mulheres parturientes no estado de Minas Gerais, Brasil.

Palavras-chave: vigilância sindrômica; cluster espacial; estatística espacial Scan de Kulldorff; clusters espaciais de formato irregular; algoritmos multi-objetivo; compacidade geométrica; função de regularidade para não-conectividade; função coesão para nodos de desconexão; doença de Chagas; testes de poder.

Abstract

Irregularly shaped spatial clusters are difficult to delineate. The most likely cluster often spreads through large portions of the map, impacting its geographical meaning. Penalized likelihood methods for Kulldorff's spatial scan statistics have been used to control the excessive freedom of the shape of clusters. Penalty functions based on cluster geometry and non-connectivity have been proposed recently. Another approach involves the use of a multi-objective algorithm to maximize two objectives: the spatial scan statistics and the geometric penalty function. We present a two novel scan statistic algorithm: an algorithm employing a function based on the graph topology to penalize the presence of under-populated disconnection nodes in candidate clusters, the *disconnection nodes cohesion function.* A disconnection node is defined as a region within a cluster, such that its removal disconnects the cluster. By applying this function, the most geographically meaningful clusters are sifted through the immense set of possible irregularly shaped candidate cluster solutions; and an algorithm maximizing simultaneously compactness and disconnection nodes regularity function of clusters. The solution is a Pareto-set, consisting of all clusters not simultaneously worse on both objectives. Significance evaluation through Monte Carlo simulations determines the best cluster solution. Our method distinguishes clearly those geographically inadequate clusters which are worse from both geometric and disconnection nodes cohesion function viewpoints. Besides, irregularity of shape is allowed provided that it does not impact topological regularity. Our method has better power of detection for clusters satisfying those requirements. We propose a more robust definition of spatial cluster using these concepts. We evaluate the statistical significance of solutions for multi-objective scans, a statistical approach based on the concept of attainment function is used. In this work we compare different penalized likelihoods employing the (1) geometric and (2) non-connectivity regularity functions and introduces a novel penalty function namely (3) disconnection nodes cohesion function. We also build multi-objective scans using those three functions and compare them with the previous penalized likelihood scans. We show that the multi-objective scans present better performance, compared to the other algorithms, regarding to power, sensitivity and positive predicted value. An application is presented using comprehensive data for Chagas' disease in puerperal women in Minas Gerais state, Brazil.

Keywords: disease surveillance; spatial cluster; Kulldorff's spatial scan statistic; irregularly shaped clusters; multi-objective; geometric compactness regularity function; non-connectivity regularity function; disconnection nodes cohesion function; Chagas' disease; power tests.

Summary

Chapter 1

Apresentação

1.1 Motivação

Observa-se, recentemente, um crescente número de trabalhos sobre metodologias para detecção e avaliação de conglomerados (clusters) espaciais e temporais. No enfoque deste texto, um cluster é um conjunto conexo de regiões onde existe a ocorrência discrepante de casos localizados para algum 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)*.

O problema de detecção de clusters espaciais encontra-se presente em diversas situações, tais como problemas associados à saúde pública (epidemiologia e vigilância sindrômica), criminologia, pesquisa de mercados, entre outros. É importante determinar modelos satisfatórios para a execução de procedimentos para detecção e avaliação destes clusters.

1.2 Objetivos e escopo da tese

Um dos objetivos deste trabalho é determinar e desenvolver estratégias para a detecção de clusters espaciais. 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, os limites geográficos de possíveis clusters espaciais, apresentando justificativas estatísticas para o funcionamento adequado do método.

Apesar da tese restringir-se ao estudo para detecção de clusters espaciais, as propostas aqui discutidas podem ser estendidas para a busca de clusters espaço-temporais.

1.3 Principais contribuições

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 geometria e topologia de clusters baseado em grafos;
- utilização de algoritmos genéticos mono e 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.

- estudo de casos de incidência da doença de chagas no estado de Minas Gerais;

1.4 Uma apresentação para o problema proposto

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 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 1.1 mostra uma zona no mapa do estado de São Paulo dividido em 72 microrregiões.

Kulldorff (1997) [47] propõe uma metodologia baseada em um teste de razão de verossimilhança. Kulldorff e Nagarwalla (1995) [46] 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.

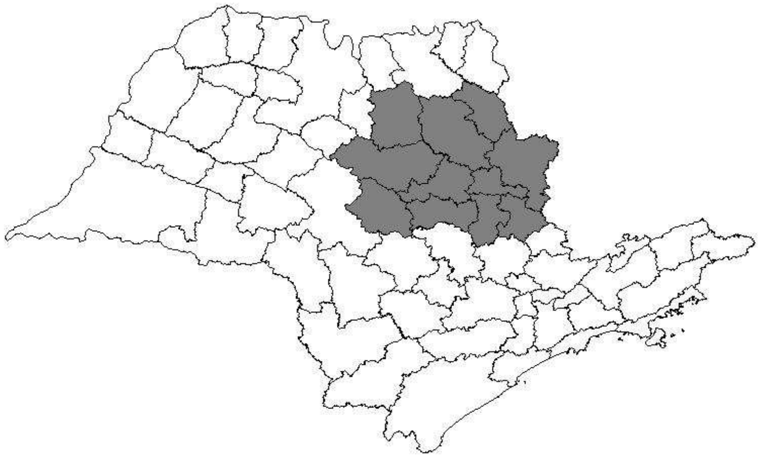


Figure 1.1: Mapa do estado de São Paulo dividido em micro regiões.

Uma janela circular sobre a área em estudo define uma zona formada pelas regiões cujos centróides estão dentro da janela. Note pela figura 1.2 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.

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 invíavel 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.

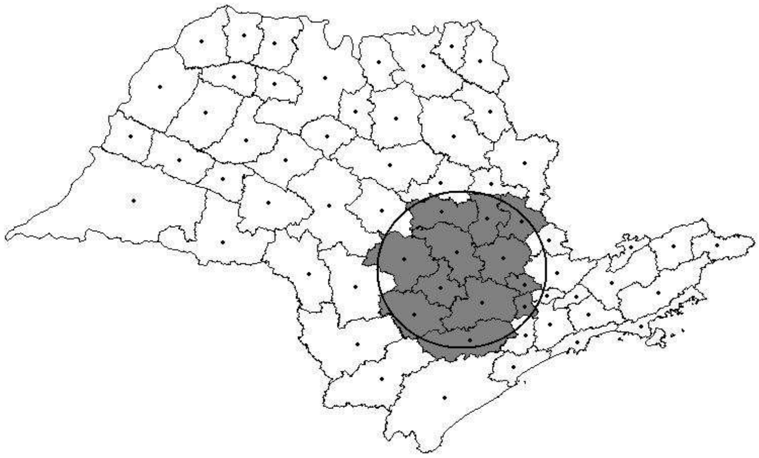


Figure 1.2: Uma possível zona obtida para uma dada janela circular.

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 idéia 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) [47] prevê a possibilidade de clusters de formato arbitrário, porém não propõe algoritmos para a detecção de clusters de formato irregular.

Neste sentido, existem alguns algorítmos 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.

Neste trabalho iremos propor uma nova forma de penalização que visa suprir as deficiências das penalizações já existentes, bem como a utilização de algumas formas de penalização já existentes. Estudaremos também algoritmos que utilizam combinações da estatística Scan espacial com estas funções de penalização. Iremos utilizar uma técnica que tem se mostrado bastante eficaz em problemas de otimização, os algoritmos genéticos multi-objetivo. Comparamos os diversos métodos propostos com os métodos já existentes em um estudo com casos reais de doença de Chagas.

1.5 Organização da tese

Apresentamos uma tese que, essencialmente, mostra os resultados obtidos e submetidos à publicação durante o programa de doutorado em Estatística. Assim, parte deste texto reproduz, na integra, a estrutura de artigos já submetidos (inclusive no idioma inglês) e apresenta a seguinte organização. No capítulo 2 é apresentada uma introdução sobre os assuntos tratados nesta tese. No capítulo 3 apresentamos uma revisão detalhada da literatura sobre tratamentos a problemas similares. Esta revisão se transformou em um capítulo de livro contendo revisão bibliográfica apresentado em Duczmal et al. (2009) [31] sobre o tema. No capítulo 4 apresentamos, de uma forma mais detalhada, algumas propostas já utilizadas para o tratamento do problema proposto. No capítulo 5 apresentamos uma nova estratégia de penalização para auxiliar a detecção de clusters espaciais irregulares. Esta nova estratégia é o tema central de Duczmal et al. (2009) [30] submetido em um periódico internacional nesta área de pesquisa. No capítulo 6 apresentamos uma descrição do algoritmo genético que foi utilizado, bem como a formulação dos problemas de otimização na forma mono-objetivo e na forma multi-objetivo. Também apresentamos uma contribuição de Duarte et al. (2009) [22] recentemente aceito para publicação em um periódico internacional nesta área de pesquisa, uma abordagem que avalia simultaneamente duas funções de penalização para o cluster, sendo uma para sua forma geométrica e outra para sua estrutura topológica. No capítulo 7 apresentamos análises numéricas que mostram a qualidade dos métodos aqui propostos na detecção de clusters em um estudo com casos simulados de câncer de mama no nordeste dos Estados Unidos e em outro onde os dados são casos reais de incidência de doenças de Chagas no estado de Minas Gerais. No capítulo 8 concluímos a tese com observações finais e tópicos para futuros trabalhos.

Chapter 2

Introduction

Algorithms for the detection and inference of irregularly shaped spatial clusters have attracted considerable attention recently. The geographic delineation of spatial clusters in a map is important to assess the causal mechanisms for the occurrence of diseases in Lawson et al. (1999) [54] and Lawson (2001) [55]. The circular scan, presented in Kulldorff and Nagarwalla (1995) [46], a particular case of the spatial scan statistic, is the most popular method for the detection and inference of disease clusters. Nevertheless, situations where spatial disease clusters do not have a regular shape (e.g. non-circular or non-square shaped clusters) are fairly common. Clusters with arbitrary shape are found along traffic ways, plumes of air pollution, or geographical features such as rivers, shores and valleys. Many heuristics were developed recently to find arbitrarily shaped clusters.

Two problems arise when detecting irregularly shaped clusters. First, the set of possible (connected) cluster candidates increases exponentially with the number of regions in the map. Second, even if this immense set could be listed and the candidate clusters were analyzed one by one, the selection of the best cluster solution based solely on the maximization of the likelihood ratio scan leads to poor solutions. High likelihood ratio clusters can be easily assembled by adjoining the highest risk regions of the map, glued together using lower risk regions, or “bridges”, forming very irregularly shaped clusters. Those clusters spread through large portions of the study area and do not bring useful information about the location of potentially interesting clusters within the map, which are generally smaller and have somewhat lower likelihood ratio values. As a result, the power of detection is reduced. Other measures for the strength of a cluster must be taken into consideration, such as geometric (Kulldorff et al. (2006) [52], Duczmal et al. (2006) [25], Duczmal et al. (2007) [27], Duczmal et al. (2008) [26]) or non-connectivity (graph-based) presents in Yiannakoulias et al. (2007) [84] regularity function or disconnection nodes cohesion function in Duczmal et al. (2009) [30]. Most cluster finding methods do not address this problem, however. Based on the fact that the circle is the most compact geometric shape, the *geometric compactness regularity function* acts like a low-pass filter, reducing the value of clusters which are very different from a round shape. The *non-connectivity regularity function* penalizes more a cluster whose associated adjacency graph has fewer edges, given its number of nodes. In other words, the most penalized clusters are those whose graphs are trees, which are loosely connected by definition. In Duczmal et al. (2006) [25], the concept of *disconnection node* was briefly discussed. A disconnection node is a region within a cluster which disconnects it when removed, splitting the cluster into two or more connected pieces. In Duczmal et al. (2009) [30] we argue that the presence of under-populated disconnection nodes impacts the power of detection of clusters. It happens because it is more difficult to aggregate loosely connected pieces which are glued through small population regions. A novel regularity function, the *disconnection nodes cohesion function*, is defined in order to measure the strength or cohesion of a cluster, based on the presence or absence of under-populated disconnection nodes.

Multi-objective genetic algorithms was developed elsewhere (Duczmal et al. (2009) [26]) to identify irregularly shaped clusters. Those methods conduct a search aiming to maximize two objectives, namely the scan statistic and the regularity of shape (using either the geometric compactness regularity function presented in Duczmal et al. (2008) [26], the non-connectivity regularity function presented in Yiannakoulias et al. (2007) [84], the disconnection nodes regularity function presented in Duczmal et al. (2009) [30] and the simultaneous compactness and disconnection nodes multi-objective scan where Kulldorff's likelihood ratio scan statistic is combined with both the geometric compactness function and the disconnection nodes cohesion function presented in Duarte et al. (2009) [22], and used as the two objectives of a multi-objective scan genetic algorithm. The solution presented is a Pareto-set, consisting of all the clusters found which are not simultaneously worse in both objectives. The multi-objective approach has an advantage over penalized likelihood methods: all potential clusters are considered for comparison without altering their ranking due to penalty modifications. Thus ranking decision is executed only after all the candidates are evaluated. Penalized methods otherwise decide beforehand the amount of applied penalty, being prone to distortions in the process of choosing the most likely cluster. Multi-objective methods eliminate all but a small set of potential *non-dominated solutions*, the candidate clusters which are not worse than any other candidate in both objectives simultaneously. The significance evaluation is conducted in parallel for all the clusters in the Pareto-set through a Monte Carlo simulation, determining the best cluster solution. In this work, we also employ a more efficient multi-objective genetic algorithm based on the NSGA-II, described in Deb et al. (2002) [19].

We also present the statistical basis for the evaluation of the significance of solutions for our multi-objective scans, through the idea of the attainment function described in da Fonseca et al. (2001) [18] and Fonseca et al. (2005) [35]. Previous approaches, based on the union of the Pareto-sets into a set of independent points, lead to a loss of information about the distribution of the Pareto sets among the objective space under the null hypothesis. The natural extension of the p-value concept to the bi-objective space is achieved with the use of the attainment function, due to the preservation of the dependence between points within the same Non-dominated sets, for all Pareto-sets obtained by the Monte Carlo simulation.

In this work we compare four multi-objective scan methods using the geometrical compactness, the non-connectivity, the disconnection node cohesion function as one objective and the spatial scan statistics as the second objective and the simultaneous compactness and disconnection nodes multi-objective scan. Those methods are compared with the corresponding single-objective likelihood penalized methods. Their power to detect irregularly shaped spatial clusters, sensitivity and positive predicted value are studied through numerical simulations and in a real data study.

Chapter 3

Review - Extensions of the scan statistic for the detection and inference of spatial clusters

3.1 Introduction

Algorithms for the detection and evaluation of the statistical significance of spatial clusters are important geographic tools in epidemiology, syndromic and disease surveillance, crime prevention and environmental sciences. The elucidation of the etiology of diseases, the availability of reliable alarms for detecting intentional and non-intentional outbreaks, the study of spatial patterns of criminal activities, and the geographic monitoring of environmental changes are current topics of intense research. Methods for finding spatial clusters were reviewed in Elliott, Martuzzi and Shaddick (1995), Waller and Jacquez (2000), Kulldorff (1999), Lawson et al. (1999), Moore and Carpenter

(1999), Glaz, Naus and Wallestein (2001), Lawson (2001), Balakrishnan and Koutras (2002) and Buckeridge et al. (2005) [33, 86, 48, 54, 59, 38, 55, 6, 8].

A descendant of Naus' pioneering spatial scan statistic, Kulldorff's spatial scan statistic Kulldorff (1997) and Kulldorff (1999) [47, 48] is currently the most popular method for finding spatial clusters. The significance of the most likely cluster is estimated through a Monte Carlo simulation (Dwass (1957) [32]). It can be used for data with exact point locations or for aggregated data, where a study region is partitioned into cells. The circular scan (Kulldorff and Nargawalla (1995) [46]), the most commonly used spatial scan statistic, sweeps completely the configuration space of circularly shaped clusters, but in many situations we would like to recognize spatial clusters in a much more general geometric setting. Several proposals for finding arbitrarily shaped spatial clusters are reviewed in section 3.2. Section 3.3 examines a number of recent data-driven algorithms for cluster detection that have been developed to include spatial mobility, survival time, multiple data streams, alternative parametric models, and non-parametric and learning models.

3.2 Irregularly shaped spatial clusters

When searching for clusters with unlimited freedom of geometric shape, the power of detection is reduced. This happens because the collection of all connected zones, irrespective of shape, is very large; the maximum value of the objective function is likely to be associated with 'tree-shaped' clusters, which merely link the highest likelihood ratio cells of the map, without contributing to the discovery of geographically meaningful solutions that delineate correctly the 'true' cluster. In other words, there is much 'noise', against which the legitimate solutions cannot be distinguished. That problem occurs in every irregularly shaped cluster detector. In this section several proposed solutions for this issue are reviewed.

The Upper Level Sets (ULS) scan statistic (Patil and Taillie (2004) [73]) controled the excessive freedom of shape exploring a very small collection of graph connected candidate zones z, evaluated according to their rate (number of cases divided by the population at risk) in the study area of n regions. The ULS-tree is constructed such that selected zones with the highest rates consisting of single region, which are local maxima for the rate, form the leaves of the ULS-tree. Neighboring regions in the study area are successively joined to the individual regions represented by the leaves, forming larger zones with lower rates which are then identified with the lower inner nodes of the ULS-tree. Eventually, those aggregated zones coalesce creating even larger, lower rated zones, represented as inner nodes closer to the root. The root itself represents the entire study area. The collection of zones represented by the ULS-tree nodes constitutes the ULS reduced parameter space, its cardinality being at most n. The ULS-tree needs to be calculated again for each new Monte Carlo replication. This procedure is fast, but possibly many interesting clusters are overlooked in this procedure, due to the small cardinality of the ULS-tree. This issue is tackled in Patil et al. (2006) [74] where an extension of the original ULS set is constructed. In Modarres and Patil (2007) [58] discussed an extension of the ULS scan statistic to bivariate data. The sensitivity of the joint hotspots to the degree of association between the variables is studied.

Duczmal and Assunção (2004) [23] proposed a simulated annealing (SA) algorithm. The collection of connected irregularly shaped zones consists of all those zones for which the corresponding subgraphs are connected. This collection is very large, and it is impractical to calculate the likelihood ratio

(LR) statistic for all of them. Instead the SA tries to visit only the most promising zones, as follows. Two zones are neighbors when they differ by a single region. For each individual region of the study area, the circular scan is used to define a starting cluster z_0. The algorithm chooses some neighbor z_1 among all the neighbors of z_0. In the next step, another neighbor z_2 is chosen among the neighbors of z_1, and so on, until a pre-defined threshold in the number of regions is attained. Thus, at each step a new zone is built adding or excluding one cell from the zone in the previous step. Instead of behaving all the time like a greedy algorithm, always choosing the highest LR neighbor at every step, the SA algorithm evaluates if there has been little or no LR improvement during the latest steps; in that case, the SA algorithm opts for choosing a random neighbor. This is done while trying to avoid getting stuck at LR local maxima. The search is restarted many times, each time using each individual cell of the map as the initial zone. Thus, the effect of this strategy is to keep the program openly exploring the most promising zones in the configuration space and abandoning the directions that seems uninteresting. The best solution found by the program, which maximizes the LR is the most likely cluster. It is called a quasi-optimal solution, and is a compromise due to computer time restraints for the identification of the geographical location of the clusters.

The Flexibly Shaped (FS) spatial scan statistic (Tango and Takahashi (2005) [82]) made an exhaustive search of all possible first-order connected clusters contained within a set encompassing the nearest K neighbors of a given region. For each region i, the flexibly shaped scan considers K concentric circles plus all the sets of connected regions whose centroids are located within the $K-th$ largest concentric circle. The procedure is repeated for each region of the map, enabling that all connected clusters are enumerated up to a size limit K. The set of potential clusters is stored in memory, so the runs under null hypothesis are executed without rebuilding them every time. For computational reasons, the search is restricted to relatively small clusters. The authors consider that a practical value for K is about 30 - finding clusters larger than that should take more than one week of computation on a desktop PC. Compared to the SA without bounds on cluster size, the FS algorithm founds more compact clusters, but when the SA pre-defined number of regions threshold is set to the same size limit K, both algorithms give similar results. (Takahashi et al. (2007) [80]) further extended the FS scan to detect space-time irregularly shaped clusters.

The Static Minimum Spanning Tree (SMST) proposed by Assunção et al. (2006) [5] used a greedy algorithm to aggregate regions. Starting with a zone consisting of one individual region, the algorithm selects the adjacent region that maximizes the likelihood ratio scan statistic and aggregates it to the zone, successively until a maximum population proportion is attained, or all regions are used. The procedure is repeated for each region of the study area. The paper describes this algorithm as the growth of a minimum spanning tree; it minimizes the sum of edge weights, defined as the difference in rates between vertices within the tree. Each step of tree growth represents a new candidate cluster. The most likely cluster is defined as the cluster that maximizes the LR.

The Density-Equalizing Euclidean Minimum Spanning Tree (DEEMST) method (Wieland et al. (2007) [87]) was an improvement of the SMST idea. A study region is provided with n points in the data set of cases and controls. Neighboring points are connected through edges, forming the complete graph T of the whole study area. Initially a Voronoi diagram of the control locations is built, subdividing the study area into regions, satisfying the property that the density, or the number of controls in each region divided by the region's area, is kept constant. This constitutes the density-equalizing cartogram, a distorted map in which the regions are magnified or demagnified according to their local density. Next, the method finds all the potential clusters, here defined as the subset of points S such that each subset of S is closer to at least one other point in S than to any other point outside of S. The authors prove that it is not needed to consider all connected subgraphs of T: aside from the trivial n individual points, there are only $n-1$ non-trivial potential clusters. Those are found from the Euclidean minimum spanning tree solution using greedy edge deletion algorithm. This method does not use the likelihood ratio statistic, but the sum of the euclidean distances of the minimum spanning tree. This method was compared with the circular SaTScan. It was found that the EMST has higher power to detect irregularly shaped clusters, but the circular scan has higher power to detect large circular clusters. Compared with the circular SaTScan, for non-circular clusters, the EMST gains in average fraction of the true cluster detected, but loses in average fraction of the most likely cluster coinciding with the true cluster.

Dematteï et al. (2007) [20] proposed a method based on the construction of a trajectory for multiple cluster detection using the spatial scan statistic in point data sets. It begins by determining a certain trajectory linking the data set points. The general idea of the method is based on the assumption that the consecutive points inside a cluster have lower associated distances than those of points outside the cluster, because the density of points is higher within the cluster. Potential clusters are located by modelling the multiple structural changes of the distances on the selection order and the best model (containing one or several potential clusters) is selected. Finally a p-value is obtained for each potential cluster. The authors discuss the possibility that the trajectory leaves the cluster before going through all the cluster points. They conclude that the remaining cluster points will be detected as a second component cluster and the proximity analysis of these two component clusters by specialists could allow them to build a new bigger cluster as the union of the two clusters detected. It is not clear, however, how a fast automatic procedure could be devised to construct these unions, particularly when there are more than just a few components.

Kulldorff et al. (2006) [52] presented an elliptic version of the spatial scan statistic, generalizing the circular shape of the scanning window. It uses an elliptic scanning window of variable location, shape (eccentricity), angle and size, with and without an eccentricity penalty. The elliptic scan has more power to detect elongated clusters, compared to the circular scan statistic.

Duczmal et al. (2006) [25] developed a geometric penalty for irregularly shaped clusters. Many algorithms frequently end up with a solution that is nothing more than the collection of the highest incidence cells in the map, linked together forming a "tree-shaped" cluster spread through the map; the associated subgraph resembles a tree, except possibly for some few additional edges. This kind of cluster does not add new information with regard to its special geographical significance in the map. One easy way to avoid that problem is simply to set a smaller upper bound to the maximum number of cells within a zone. This approach is only effective when cluster size is rather small (i.e., for detecting those clusters occupying roughly up to 10% of the cells of the map). For larger upper bounds in size, the increased geometric freedom favors the occurrence of very irregularly shaped tree-like clusters, thus impacting the power of detection. Another way to deal with this problem is to have some shape control for the zones that are being analyzed, penalizing the zones in the map that are highly irregularly shaped. For this purpose the geometric compactness of a zone is defined as the area of z divided by the circle with the perimeter of the convex hull of z. Compactness is dependent on the shape of the object, but not on its size. Compactness also penalizes a shape that has small area compared to the area of its convex hull. A user defined exponent α is attached to the penalty to control its strength; larger values of α increases the effect of the penalty, allowing the presence of more compact clusters. Similarly, lower α values allows more freedom of shape. The idea of using a penalty function for spatial cluster detection, based on the irregularity of its shape, was first used for ellipses (Kulldorff et al. (2006) [52]), although a different formula was employed.

The greedy algorithm idea was used by Yiannakoulias et al. (2007) [84] to explore the space of all possible configurations. A new penalty function is now defined as the ratio of the number of edges $e(Z)$ to the total possible number of edges in the candidate cluster Z. The total possible number of edges is computed as $3(v(Z)-2)$ based solely on the number of vertices $v(Z)$ in the candidate cluster. The non-connectivity penalty is employed as an exponent to the LR, analogously to the geometric compactness penalty. In the same way, a user defined exponent α is attached to the non-connectivity penalty to control its strength. Instead of stopping the candidate clusters' aggregation process before reaching a pre-specified population proportion limit, another criterion is used, based on the failure to increase the LR to a higher value after a certain number u of steps. The parameter u is set by the user; larger values of u relaxes the search constraint, and making $u = 0$ halts the search when no vertices can be added that increases the LR. Although the non-connectivity penalty is in many ways similar to the geometric compactness penalty, it has an important difference: it does not rely on the geometric shape of the candidate cluster, which could be an interesting advantage when searching for real clusters that are highly irregularly shaped, but present good connectivity properties.

Conley et al. (2005) [15] proposed a genetic algorithm to explore a configuration space of multiple agglomerations of ellipses for point data sets. The method employed a strategy to “clean-up” the best configuration found in order to simplify geometrically the cluster.

Sahajpal et al. (2004) [77] also used a genetic algorithm to find clusters shaped as intersections of circles of different sizes and centers in point data sets.

Duczmal et al. (2007) [27] described a genetic algorithm scan for the detection and inference of irregularly shaped spatial clusters. Assuming a map divided into regions with given populations at risk and cases, the graph-related operations are minimized by means of a fast offspring generation and evaluation of Kuldorff´s spatial scan statistic. The penalty function of (Duczmal et al. (2006) [25]), based on the geometric non-compactness concept, is employed to avoid excessive irregularity of cluster geometric shape. This algorithm is an order of magnitude faster and exhibits less variance compared to the simulated annealing scan, and is more flexible than the elliptic scan. It has about the same power of detection as the simulated annealing scan for mildly irregular clusters and is superior for the very irregular ones.

Gaudart et al. (2005) [36] oblique decision tree (ODT) was a modification of the classification and regression tree (CART) strategy to obtain an optimal partitioning procedure in order to detect spatial patterns and find the candidate clusters without prior specifications. Instead of using rectangular partitions of the covariate space as in CART, ODT provides oblique partitions maximizing the interclass variance of the independent variable, providing polygonal candidate clusters. Classical ODT algorithms in R^n relies on evolutionary algorithms or heuristics, but in this work an optimal ODT algorithm is developed in R^2, based on the directions defined by each couple of point locations. The procedure consists on finding several partitions of the plane. The first step finds the best oblique split of the plane between two adjacent classes, maximizing the interclass variance. Going recursively, this algorithm will split the plane into several partitions, until a specific stopping criterion is reached. Monte Carlo replications are used to test significance.

Multi-Resolution methods (MR) (Neill and Moore (2003) [64] and Neill and Moore (2004) [65]) maximized Kulldorff's scan statistic over the square regions S of a grid of $g \times g$ squares, each one with an assigned number of cases and controls. Instead of using a naïve approach which would require $O(g^3)$ calculations (multiplied by R Monte Carlo replications), the MR algorithm partitions the grid into overlapping regions, bounds the maximum score of sub-regions contained in each region, and prunes regions which cannot contain the maximum density region. The maximum density region is found using $O(g^2)$, for sufficiently dense regions. (Neill et al. (2005) [66]) later introduced another algorithm, the fast spatial scan (FS), generalizing the original bi-dimensional MR to arbitrary dimensions and using rectangles instead of squares. Applications include multiple data streams in syndromic surveillance (emergency department visits and over-the-counter drug sales), and discovery of regions of increased brain activity corresponding to given cognitive tasks (from fMRI data).

Given n baseline and case points, Agarwal et al. (2006) [3] presented an algorithm to compute exactly the maximum discrepancy rectangle in time $O(n^4)$. If the points lie in a $g \times g$ grid, the algorithm runs in time $O(g^4)$. This algorithm has the same asymptotic running time as the MR algorithm.

A much better performance is achieved for the general family of *discrepancy functions* (including Kulldorff's scan), through the Approx-Linear Algorithm (AL) by representing the discrepancy function as the upper envelope of a collection of linear functions. It is shown that a thoroughly linear approximation of the discrepancy function, which would require many linear functions, is not strictly necessary, because the approximation needs only to preserve the ordering of points along the direction of the search. As a result, a much better algorithm can maximize the discrepancy function over axis parallel rectangles in time $O(n^2 \log n)$. The algorithm is also extended to aggregate data sets using a regular $g \times g$ grid. A further technique is presented, using sampling to compute an approximation to the maximum linear discrepancy.

Aldstadt and Getis (2006) [4] proposed the AMOEBA (Multidirectional Optimum Ecotope-Based Algorithm). An *ecotope* or *habitat* is defined in the literature as a specialized region within a larger region. A local spatial autocorrelation statistic is employed to construct a spatial weights matrix, used to describe the association between contiguous spatial units. The weights matrix is used in the determination of geometric form of spatial clusters. It searches for spatial association in all specified directions, starting from a selected collection of 'seed' spatial units. The main objective is to identify the ecotopes, the spatially homogeneous subregions within the study area. AMOEBA is compared with SaTScan.

Duczmal et al. (2008) [26] proposed an approach to the geographic delineation of irregularly shaped disease clusters, treating it as a multi-objective optimization problem. Irregularly shaped spatial disease clusters occur commonly in epidemiological studies, but their geographic delineation is poorly defined. Most current spatial scan software usually displays only one of the many possible cluster solutions with different shapes, from the most compact round cluster to the most irregularly shaped one, corresponding to varying degrees of penalization parameters imposed to the freedom of shape. Even when a fairly complete set of solutions is available, the choice of the most appropriate parameter settings is left to the practitioner, whose decision is often subjective. A quantitative criteria for choosing the best cluster solution is presented, maximizing simultaneously two competing objectives: regularity of shape ($K(z)$), and scan statistic value (LLR). The Pareto set is defined as the set of all clusters candidates z such that no other cluster has both higher LLR and higher regularity than z. For each value of $K(z)$, a separate empirical distribution of LLR under the null-hypothesis is computed, constituting a two-dimensional p-value surface. The cluster with lowest p-value is considered the most likely cluster. Instead of running a cluster finding algorithm with varying degrees of penalization, the set of non-dominated solutions is found in parallel, through a genetic algorithm. The p-value surface is computed using Gumbel approximations (Abrams et al. (2006) [2]).The introduction of the concept of Pareto-set in this problem, followed by the choice of the most significant solution, is shown to allow a rigorous statement about what is such “best solution”, without the need of arbitrary parameters.

Maps with irregularly shaped or multiple clustering, when there is not a clearly dominating primary cluster, occur frequently. Moura et al. (2007) [62] developed a method to analyze more thoroughly the several levels of clustering that arise naturally in a disease map divided into m regions. Instead of using a genetic algorithm, this method incorporates the simplicity and speed of the circular scan, being able to detect and evaluate irregularly shaped clusters. The circular occupation (CO) of a cluster candidate is defined roughly as its population divided by the population inside the smallest circle containing it. The CO concept, computationally faster and relying on familiar concepts, substitutes here the compactness definition as the measure of regularity of shape. A multi-objective modification of the circular scan algorithm is applied, using CO and LLR as the objectives. The comparison of Pareto-sets for observed cases with those computed under the null-hypothesis provides valuable hints for the spatial occurrence of diseases. The potential for monitoring incipient spatial-temporal clusters at several geographic scales simultaneously is a promising tool in syndromic surveillance, especially for contagious diseases when there is a mix of short and long range spatial interactions. The presence of "knees" in the Pareto-sets indicates sudden transitions in the clusters structure, corresponding to rearrangements due to the coalescence of loosely knitted (usually disconnected) clusters.

Yiannakoulias et al. (2007) [85] employed Quad trees to generate nonuniform grid points in order to detect spatial clusters in study areas provided with a large number of points. This strategy is compared with another scheme, which uses uniform grid points. The quad tree approach is more sensitive to high-resolution spatial clusters and is also more flexible, compared with the uniform grid approach.

Boscoe (2003) [10] proposed a tool to visualize relative risk and statistical significance simultaneously. Given a map of n regions, with their respective centroids, the procedure builds a grid of equidistant points between all combinations of two, three and four adjacent region centroids. For each grid point the distances to the regions centroids are computed and sorted. These distances are used to define almost circular groupings of regions, with their respective cumulative numbers of observed and expected cases. The relative risk and the LLR are then calculated for each circular grouping. The LLR values are compared to the results of a Monte Carlo simulation under the null hypothesis. Groupings with LLR values exceeding 95% of those obtained from the simulation are stored and stratified into ten levels of relative risk. Within each risk level, the grouping with largest LLR is then mapped. Circular groupings with lower LLR are also mapped if they did not overlap any grouping previously mapped. The final result is a ten color shaded map of regions with statistically significant relative risks, providing a very effective visualization tool to grasp these two concepts.

There exist many methods to detect boundaries and to detect clusters; Jacquez et al. (2007) [45] proposed the b-statistic as a tool for the simultaneous detection of boundaries and clusters. It evaluates boundaries between adjacent areas with different values, and also the existing links between adjacent areas with similar values. Clusters are constructed by joining similarly high valued areas, which are then connected through a link. Unlike the local Moran and other statistics, which describe local spatial variation in the immediate local neighborhood about a central location, the b-statistic describes properties of the edge between two areas. The b-statistic was compared with Polygon wombling cf.Womble (1951) [88] for detecting boundaries and the local Moran test (Moran (1948) [60] and Moran (1950) [61]).

Haiman and Preda (2002) [39] derived approximations for the estimation of the distribution of scan statistics for a two-dimensional Poisson process. Through extensive numerical tests, Abrams et al. (2006) [2] showed that, under the null hypothesis, the empirical distribution of values of Kulldorff's scan statistic for circular clusters is approximated by the well-known Gumbel distribution. The authors calculated that using this semi-parametric approach, 100 Monte Carlo replications suffice to provide the same accuracy in significance estimation as 10,000 replications using the usual empirical distribution.

Kulldorff et al. (2003) [50] presented a large collection of simulated benchmark data sets generated under different cluster models and the null hypothesis, to be used for power evaluations. These data sets are used to compare the power of the spatial scan statistic, the maximized excess events test and the nonparametric M statistic.

Duczmal et al. (2008) [28] described a graph based model for cluster detection and inference on networks based on the scan statistic. Nodes, associated to cities, are linked by means of edges, which represent routes between cities. Instead of forming clusters candidates by grouping neighboring nodes of the original graph, the cluster candidates are chosen among the connected subgraphs of the dual graph. The objective is to find collections of plausible pathways by which the disease could be transmitted. The most likely cluster is naturally the most structurally stable connected subgraph, or arrangement of pathways, meaning that adding or subtracting pathways to it should decrease the observed signal-to-noise proportion. In this model, traffic between cities is the analogous of population in the usual scan, and the number of syndromic individuals traveling between cities corresponds to the number of cases.

The Prospective Time Periodic Scan (Kulldorff (2001) [49]) was a space-time scan statistic for regular time periodic disease surveillance to detect any active geographical clusters of disease. The statistical significance of such clusters is adjusted for multiple testing, taking account of all possible geographical locations and sizes, time intervals and time periodic analyses.

The pyramidal flexible shape space-time scan for point data sets proposed by Iyengar (2004) [42], instead of building space-time cylinders, adopted the more flexible pyramid or cone shapes with its axis perpendicular to the space plane. It represents an advance over the usual cylindrical approach, because it is now possible to model emerging spatially growing or shrinking clusters over time.

Kulldorff et al. (2005) [51] presented the Space-Time Permutation Scan Statistics (STPSS) for outbreak detection in syndromic surveillance systems. Emerging clusters are detected using cylinder with variable radius and height which are used to scan the space-time region in order to select the candidate cluster with maximum likelihood. A data permutation procedure is executed through Monte Carlo simulation in order to estimate the p-value of the most likely cluster. This method does not require the previous knowledge of the population at risk. Costa et al. (2008) [17] extended the STPSS to detecting irregular space-time clusters.

3.3 Data-Driven Spatial Cluster Detection Models

In this section we review data-tailored algorithms for spatial cluster detection including censored survival data, spatial mobility, multiple data streams, parametric models different from the usual Poisson or Bernoulli distributions, non-parametric and learning models.

Cook et al. (2007) [16] considered a Spatial scan statistic for censored outcome data. In contrast to the traditional scan statistics, which usually requires a complete specification of the model, this paper uses a statistic score of the model of proportional risks to allow more flexibility. Cluster significance is estimated through permutation tests.

Huang et al. (2007) [40] proposed a spatial scan statistic based on an exponential model to include uncensored or censored continuous survival data. The method achieves good power and sensitivity, for several survival distribution functions including the exponential, gamma, and log-normal distributions. Huang et al. (2007) [41] applied the previous methodology to investigate possible relationships between the cluster locations and social and health conditions using nonparametric methods, and compare socioeconomic factors inside and outside of the detected clusters and evaluate the effect of related covariates on significant long and short-survival detected clusters.

Kulldorff et al. (2007) [53] proposed the Multivariate Scan Statistic. Frequently more than one data stream may be available in disease surveillance systems. When analyzed separately instead of combined, the power of detection of an outbreak signal that is present in all data streams may decrease due to low counts in each. Besides, the simple summation of all data stream counts may obliterate a signal that is primarily present in just one data stream, due to random noise present in the other data sets. These two problems are tackled by defining an extension of the space-time scan statistic as the sum of the individual log likelihoods for those data sets for which the observed case count is more than the expected.

The multivariate Bayesian scan statistic (MBSS) of Neill et al. (2007) [68] proposed modeling different outbreak types employing multiple data streams. However, this approach uses fixed methods and models for analysis, and cannot improve their performance over time. Neill and Makatchev (2008) [70] incorporated machine learning algorithms in the MBSS system. Two methods were devised for overcoming this limitation, learning a prior over outbreak regions and learning outbreak models from user feedback. They demonstrate through simulations that learning can enable systems to improve detection performance over time.

Motivated by the fact that the regions inside a cluster candidate are not homogeneous, Takahashi and Tango (2007) [79] proposed an alternative scan statistic that can take the variability of the relative risks of regions included in Z into account, employing Anscombe's variance stabilization transformation.

Tango (2007) [83] proposed a modified likelihood ratio test statistic which accounts for each individual region's risk. This modified scan includes an indicator variable based on the p-value for the zone consisting of the individual region i. Given a pre-specified α_1 significance level, and if p_i the p-value of the zone consisting of the individual region i, then the modified LR scan for a cluster including i is taken as zero when $p_i > \alpha_1$.

Neill and Moore (2006) [67] presented the Expectation-Based Scan Statistics (EBSS) as an extension of the usual spatial and space-time scan statistics by inferring expected counts for each location from past data and detecting regions where recent counts are higher than expected. Neill and Lingwall (2008) [69] presented the Nonparametric Scan Statistic (NPSS), a general detector of space-time clusters in syndromic surveillance using multiple data streams. It does not assume a parametric model, but instead combines empirical p-values across multiple locations, days, and data streams to detect anomalies.

A discrete event model was used by Beeker et al. (2007) [9] to simulate the spread of infectious diseases through an agent-based, stochastic model of transmission dynamics. The objective is to generate a benchmark from a network of individual contacts in an urban environment using publicly available population data. Such benchmark can be used to test the performance of various temporal and spatio-temporal detection algorithms when real data are not available or cannot be used due to confidentiality issues.

Duczmal and Buckeridge (2006) [24] have derived an extension to the spatial scan statistic that accounts for the mobility of individuals between home address and workplace. An analyst can use the workflow scan statistic to search for disease clusters due to workplace exposure when health records contain only residential address. The effect of the workflow scan statistic is to 'pull back' the scattered workers that were contaminated in the workplace. Simulation studies demonstrate that in most scenarios, the workflow scan statistic has greater power than the usual scan statistic for detecting disease outbreaks due to workplace exposures. The workflow scan statistic is particularly useful when clusters are not circularly symmetrical, and thus more easily recognized by the workflow scan than by the usual spatial scan algorithm.

Cami et al. (2007) [11] presented a refinement of a Bayesian algorithm used for aerosol detection (BARD) incorporating a model that includes the mobility of the individuals. The population is subdivided into groups based on the residential and workplace information.

Local, global and focused tests were developed by Jacquez et al. (2005) [43] to evaluate clustering in case-control data that take into account individual mobility. Matrices of nearest neighbor relationships are employed to represent the changing topology of cases and controls. The model includes the latency between exposure and disease manifestation. Jacquez et al. (2006) [44] analysed case-control clustering with individual mobility accounting for risk factors and covariates. Meliker and Jacquez (2007) [57] extended those previous ideas to space-time clustering of case-control data with individual mobility. Using the Q-statistic, a statistic that includes time-dependent nearest-neighbors, the authors evaluate empirical induction periods, age-specific susceptibility, and calendar year-specific effects.

Zhang and Lin (2007) [89] presented a decomposition of Moran's I test into three components so that each component represents a global test statistic. The three components tests for the existence of high-value clustering, low-value clustering, and negative autocorrelation. A set of simulations shows that the first test statistic is likely to be significant only for high-value clustering, the second test statistic is likely to be significant only for low-value clustering, and the last test statistic is likely to be significant only for negatively correlated spatial structures. Two real data examples where studied, and in both cases low-value clustering and high-value clustering were shown to exist simultaneously.

Lin and Zhang (2007) [56] combined the permutation test of Moran's I to the residuals of a loglinear model under the asymptotic normality assumption. It provides the versions of Moran's I based on Pearson residuals and deviance residuals so that they can be used to test for spatial clustering while at the same time account for potential covariates and heterogeneous population sizes.

Aggregation is commonly used as a mask to protect health data confidentiality of individuals. Ozonoff et al. (2007) [72] studied the association between spatial resolution and power of detection through thousands of simulations with the spatial scan statistic. Power to detect clusters decreased from nearly 100% when using exact locations to roughly 40% at the coarsest level of spatial resolution. The authors conclude that aggregation has the potential to obliterate existing clusters.

The usefullness of individual-level health data point locations in providing high quality data for epidemiological research must be balanced with the easiness of breaking the confidentiallity of the identities of the individuals. Geographic masking is being employed as a tool for achieving an appropriate balance between data utility and confidentiality. Usually the masks employ perturbation, aggregation of areas, and a combination of both. Zimmerman and Pavlik (2008) [90] discussed whether certain characteristics of the mask (mask metadata) should be disclosed to data users and whether two or more distinct masked versions of the data can be released without breaching confidentiality.

Glaz and Zhang (2006) [37] defined a maximum scan score-type statistic for testing the null hypotheses that the observed data are *iid* according to a specified distribution, against a class of window clustering-type alternatives. The maximum scan score-type statistic detects clustering effectivelly in the situation where the window size is unknown. The extension to multivariate data is discussed by the authors.

In disease surveillance, anomalies may be detected either by computing confidence intervals for region rates or by running a disease cluster detection algorithm. Rosychuk (2006) [75] attempts to determine when those two approaches give the same answers. The study compared (Besag and Newell (1991) [7]) cluster detection method with confidence intervals for crude and directly standardized rates. Simulations suggest that the cluster detection method is preferred when the cluster size exceeds the number of cases in a region or when the expected number of cases exceeds a threshold.

In some situations of disease surveillance, it is prefereable to use disease-related events instead of individuals as the units of analysis.

Rosychuk et al. (2006) [76] proposed a compound Poisson method that detects event clusters by testing individual areas that may be combined with their nearest neighbors. This technique is useful where the population sizes are diverse and the population distribution by important strata may differ by area.

Song and Kulldorff (2003) [78] compared the statistical power of several disease clustering tests: Besag-Newell's R, Cuzick-Edwards' k-Nearest Neighbors (k-NN), the spatial scan statistic, Tango's Maximized Excess Events

Test (MEET), Swartz' entropy test, Whittemore's test, Moran's I and a modification of Moran's I. Except for Moran's I and Whittemore's test, all other tests have good power for detecting some kind of clustering. The spatial scan statistic is good at detecting localized clusters. Tango's MEET is good at detecting global clustering. With appropriate choice of parameter, Besag-Newell's R and Cuzick-Edwards' k-NN also perform well.

Aamodt et al. (2006) [1] conducted a simulation study to compare three methods: SaTScan, generalized additive models (GAM) and Bayesian disease mapping (BYM).

Ozdenerol et al. (2005) [71] compared the results of Kulldorff's Spatial Scan Statistic with the results of Rushton's Spatial filtering technique through increasing sizes of spatial filters.

.

Chapter 4

The spatial scan statistics and regularity functions

In this chapter we review the spatial scan statistic (Kulldorff (1997) [47]), the geometric (Duczmal et al. (2006) [25]) and the non-connectivity (Yiannakoulias et al. (2007) [84]) regularity functions.

4.1 Kulldorff's Spatial Scan Statistic

A study area map A is divided into M regions, with total population N and C total cases. A non-directed graph denoted by G_A is associated to the study area A with M nodes representing the regions and edges linking nodes associated with adjacent regions. A zone is any collection of connected regions. Under the null hypothesis there are no clusters in the map, and the number of cases in each region is Poisson distributed proportionally to its population. For each zone z, the number of observed cases is c_z and the expected number of cases under null hypothesis is μ_z. The relative risk of z is $I(z) = c_z/\mu_z$ and the relative risk of the complement of z is

$O(z) = (C - c_z)/(C - \mu_z)$. Defining $L(z)$ as the likelihood function under the alternative hypothesis and L_0 as the likelihood function under the null hypothesis, it can be shown (see Kulldorff (2007) [47] for details) that the logarithm of the likelihood ratio for the Poisson model is given by:

$$LLR(z) = \log\left(\frac{L(z)}{L_0}\right) = \begin{cases} c_z \log(I(z)) + (C - c_z)\log(O(z)) & \text{if } I(z) > 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

It is maximized over the chosen set Z of potential zones z, identifying the zone that constitutes the *most likely cluster*. For instance, when the set Z contains the zones defined by circular windows of different radii and centers, $\max_{z \in Z} LLR(z)$ is the circular scan statistic (Kulldorff and Nagarwalla (1995) [46]); when Z contains all the zones defined by elliptical windows of different sizes, centers, elongations and orientations, $\max_{z \in Z} LLR(z)$ is the elliptic scan statistic (Kulldorff et al. (2006) [52]). When Z is the set of all connected zones, the evaluation of every zone of Z is not feasible in practice, and many heuristics have appeared recently to compute approximate values for $\max_{z \in Z} LLR(z)$ (Duczmal et al. (2009) [31]). Those heuristics (often called irregularly shaped spatial scan statistics) employ stochastic algorithms to explore the set of configurations Z or alternatively evaluate a restricted subset of Z.

The statistical significance of the most likely cluster of observed cases is computed through a Monte Carlo simulation, according to Dwass (1957) [32]. Under the null hypothesis, simulated cases are distributed over the study area and the scan statistic is computed for the most likely cluster. This procedure is repeated thousands of times, and the distribution of the obtained values is compared with the LLR of the most likely cluster of observed cases, producing an estimate of its p-value.

4.2 The geometric penalty function

Most irregularly shaped spatial cluster detection algorithms frequently end up with a cluster solution that is merely a collection of the high incidence regions, linked together forming a “tree-shaped” zone spread through the map; the associated sub-graph resembles a tree, possibly except for some few additional edges. In general, it is hard to give a geographical meaning for this kind of cluster, because this kind of solution does not add any new information with regard to its special location in the map. One easy way to avoid that problem is simply to set an upper bound to the maximum number of cells within a zone. This approach is only effective when cluster size is rather small (i.e., for detecting clusters occupying roughly up to 10% of the regions of the map). For larger upper bounds in size, the increased geometric freedom favors the occurrence of very irregularly shaped tree-like clusters, thus impacting the power of detection. The geometric compactness penalty for irregularly shaped clusters was presented in Duczmal et al. (2006) [25], penalizing the zones in the map that are highly irregularly shaped. For this purpose the geometric compactness $K(z)$ of a zone z is defined as the area of z divided by the area of the circle with the same perimeter as the convex hull of z.

We will penalize the zones in the map that are highly irregularly shaped. Given a planar geometric object z, define $A(z)$ as the area of z and $H(z)$ as the perimeter of the convex hull of z. Define the compactness of z is as:

$$K(z) = \frac{4\pi A(z)}{H(z)^2} \tag{4.2}$$

Compactness is dependent on the shape of the object, but not on its size. Compactness also penalizes a shape that has small area compared to the area of its convex hull (Duczmal et al. (2006) [25]). The circle is the most compact shape ($K(z) = 1$) and a square has compactness ($K(z) = 0.785$). The compactness penalyzed scan statistic is defined as $max_{z \in Z} LLR(z).K(z)$. A user defined exponent a can be attached to $K(z)$ in order to control its strength; the resulting scan statistic is then $max_{z \in Z} LLR(z).K(z)^a$. Larger values of a increase the effect of the penalty, allowing the presence of more compact clusters only. Similarly, lower values of a allow for more freedom in shape. The idea of using a penalty function for spatial cluster detection, based on shape irregularity, was first used for ellipses (Kulldorff et al. (2006) [52]) although a different formula was employed.

4.3 The non-connectivity penalty function

Yiannakoulias et al. (2007) [84] proposed a greedy algorithm to explore the space Z of all possible zones z. A new non-connectivity penalty function was based on the ratio of the number of edges $e(z)$ to the number of vertices $v(z)$ in the candidate cluster z.

The non-connectivity penalty of z is defined as:

$$Y(z) = \frac{e(z)}{3(v(z) - 2)} \tag{4.3}$$

The non-connectivity penalty was employed as a multiplier to $LLR(z)$, analogously to the geometric compactness penalty. In the same way, a user defined exponent a is attached to the non-connectivity penalty to control its strength. Although the non-connectivity penalty is in many ways similar to the geometric compactness penalty, it has an important difference: it does not rely on the geometric shape of the candidate cluster, which could be an interesting advantage when searching for real clusters which are highly irregularly shaped, but present good connectivity properties.

Some examples can illustrate the proposal of the non-connectivity penalty function, see the figure 4.1.

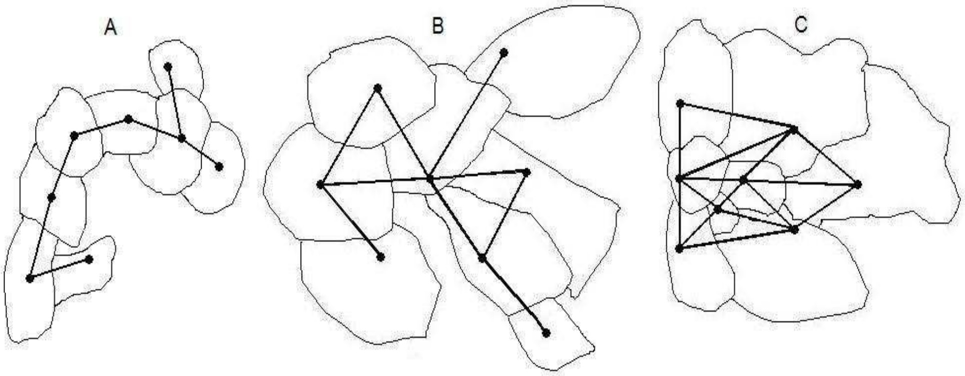


Figure 4.1: Non-connectivity penalty function evaluation for several clusters.

We can observe through the example that the Cluster A is less connected if consider its nodes and edges, already Cluster B is a little more connected and the Cluster C better connected of all. In this case that we have the following measures for each one of clusters

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

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

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

Chapter 5

The disconnection node cohesion function

In this chapter, we present a new penalty proposal for zones in the study map. In this case, we will penalize a zone z according to its topological structure.

5.1 Cohesion function

Consider a study area map A with its associated non-directed graph G_A, and a connected zone z with the corresponding connected sub-graph $G = (V, E)$ of G_A. The nodes in set V correspond to the regions of z and each non-directed edge (i, j) in set E occurs whenever the regions i and j share a common boundary. A node $x \in V$ is called a *disconnection node* of G if the sub-graph obtained from G with the nodes set $V - \{x\}$ is not connected. Let $G_D = \{x_1, \ldots, x_d\} \subset V$ be the set of all the disconnection nodes of G. For each $x_i \in G_D$, let $pop(x_i)$ be the population of the region associated with node x_i. Let μ_{x_i} be the expected number of cases of the region corresponding to node x_i under the null hypothesis, which is proportional to $pop(x_i)$. The sub-graph with the nodes set $V - G_D$, obtained from G, consists of the L *remaining connected subgraphs* $\hat{z}_1, \ldots, \hat{z}_L$, where $2 \leq L \leq |V| - d$. Let $pop(\hat{z}_j)$ be the population of the *remaining connected zone* associated to $\hat{z}_j$. The L connected parts $\hat{z}_1, \ldots, \hat{z}_L$ are ranked in decreasing order according to their populations, as $\hat{z}_{(1)}, \ldots, \hat{z}_{(L)}$.

The cohesion function of the sub-graph G is now defined as:

$$c(G) = \begin{cases} \left(\prod_{i=1}^{d} \left(1 - e^{-\mu_{x_i}}\right) \right) \prod_{i=1}^{L} \dfrac{pop\left(\hat{z}_{(i)}\right)}{\sum\limits_{j=i}^{L} pop\left(\hat{z}_{(j)}\right)} & \text{if } G_D \text{ is not empty} \\ 1 & \text{otherwise} \end{cases} \tag{5.1}$$

If each region has non-zero population, then $0 < c(G) \leq 1$.

If we assume that the number of cases c_{x_i} in each disconnecting node $x_i \in D$ is a Poisson random variable with mean μ_{x_i}, then the factor $1 - e^{-\mu_{x_i}}$ is equal to $P(c_{x_i} > 0)$, the probability of the number of the cases being greater than zero. It is important to note that we are not assuming independence with respect to the product over the disconnecting nodes of the factors $1 - e^{-\mu_{x_i}}$. Thus the first term in the cohesion formula penalizes those zones which have low populated disconnecting nodes, indicated by lower values of μ_{x_i}.

The second term penalizes homogeneous population distribution among the L connected subgraphs $\hat{z}_1, \ldots, \hat{z}_L$: it is understood that the presence of disconnecting nodes which break the cluster apart more evenly (regarding their populations) strongly impacts its cohesion. Otherwise, breaking the cluster more heterogeneously, i.e., leaving large parts of it intact while breaking away only low populated remaining connected parts, is considered less damaging to its cohesion.

Figure 5.1 presents six clusters *A*-*F* where the regions are represented by hexagons. The disconnecting nodes are indicated by dark gray hexagons. Each cluster consists of one or two disconnecting nodes and two or three remaining connected zones(represented by connected sets of light gray hexagons). Each remaining connected zone carries a number representing its population. The value of the cohesion function $c(z)$ is displayed below each cluster.

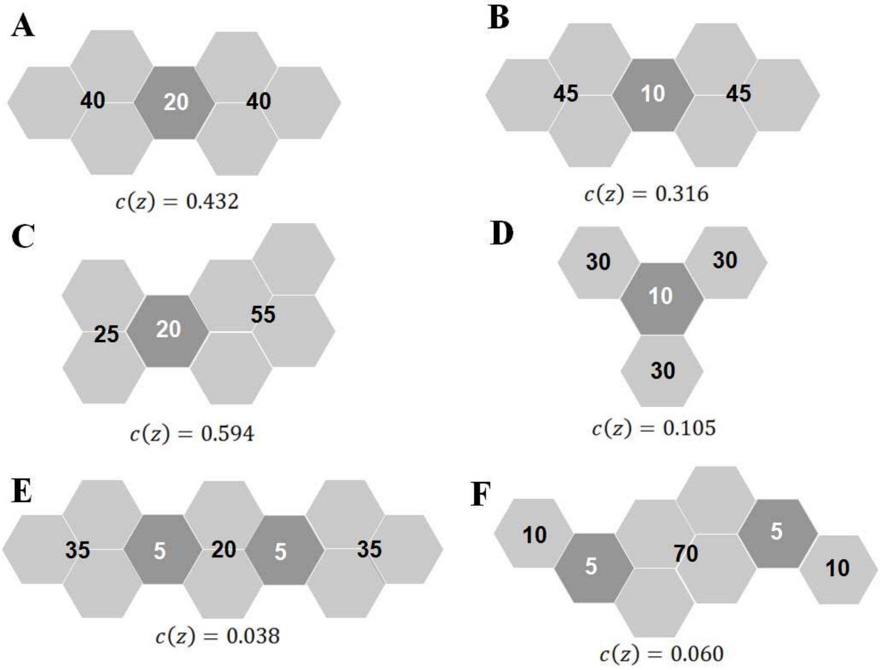


Figure 5.1: Disconnection nodes cohesion function evaluation for several clusters.

Consider that the study area has a total of 100 cases and population 1,000, representing 10% of the total risk population. The cohesion value for cluster E, for instance, is computed as:

$$c(z) = (1 - \exp(-0.1 \times 5))^2 \left(\frac{35}{35+35+20}\right)\left(\frac{35}{35+20}\right)\left(\frac{20}{20}\right) = 0.038$$

Clusters A and B differ in the population size of their disconnection nodes. Cluster A has larger cohesion and is considered more structurally stable because its two remaining zones are linked by a disconnection node with larger population.

Clusters A and C differ in the population heterogeneity of their remaining zones. When removed from cluster C, the disconnection node leaves a relatively large remaining connected zone of population 55 intact. Cluster C has larger cohesion and is considered more structurally stable because its two remaining zones have very different populations, compared with the two evenly distributed remaining zones of cluster A.

Cluster D illustrates the effect of splitting the cluster into more than two remaining connected zones. Compared to cluster B, cluster D has very low cohesion due to the fact that it is split into three equally populated remaining zones after the removal of the disconnection node.

The removal of the two disconnection nodes in clusters E and F produces three remaining connected zones in each cluster. The three remaining connected zones of cluster E are more homogeneously distributed than the corresponding ones of cluster F. Consequently, cohesion for cluster F is higher, due to the fact that the central remaining connected zone of cluster F has relatively higher population 70.

When used as a penalty factor, $c(G)$ is incorporated in the expression (4.1) for the test statistic as a multiplier for the log likelihood ratio, meaning that the penalization is strong when the cohesion function assumes lower values (there is no penalization at all when $c(G) = 1$).

5.2 Multi-objective optimization

As another way to deal with the problem of cluster detection is through multi-objective optimization procedures. Since cluster detection problems can be formulated as multi-objective optimization problem, we will present a brief description of the multi-objective concepts.

A multi-objective optimization problem (MOP) arises when one must optimize simultaneously two or more conflicting objective-functions, subject or not to some constraints. "Conflicting" here refers to the fact that it is not plausible that one choice for the optimization variables will optimize all objectives simultaneously. For that reason, the search for the best solution in a MOP is closely related to the *dominance* concept.

Let a function to be maximized $f(x) = (f_1(x), ..., f_n(x))$ be defined in a space X. A point $x_1 \in X$ dominates another point $x_2 \in X$ if $f_i(x_1) \geq f_i(x_2)$, $i = 1, ..., n$ and $f_k(x_1) > f_k(x_2)$ for at least one $k \in \{1, ..., n\}$. In other words, a point x_1 dominates another point x_2 if the evaluation of x_1 is better than the evaluation of x_2 for at least one objective while not being worse for the other objectives. Then notice given two solutions s_1 and s_2, one, and only one of the three will occur: (i) s_1 dominates s_2, or (ii) s_2 dominates s_1 or (iii) neither s_1 dominates s_2, nor s_2 dominates s_1 (in this case we say that s_1 and s_2 are incomparable). Now, consider a set of solutions. The *Pareto-set* is formed by all solutions that are not dominated by any solution in the search space X. Note that any pair of the solutions in the Pareto-set are incomparable.

From the last paragraph it is clear that the solution of a MOP is a set of non-dominated solutions, called the Pareto-set. This set represents a trade-off between the objectives, meaning that if one tries to improve one objective, at least one of the other(s) objective(s) will fatally suffer a deterioration effect.

5.3 The multi-objective disconnecting node cohesion

We discuss an implementation of the multi-objective treatment employing both the test statistic (4.1) and the disconnecting node cohesion function $c(G)$. Now the cohesion function is not used as penalty correction but instead as the second objective to be maximized. In the example of Figure 5.2, two possible clusters ACB and ADB, are evaluated. The number of cases "X" and population "Y" are represented as "X/Y" for each region. Cluster ACB has weak link C and population 29, and cluster ADB has weak link D and population 25 (both have 2 cases). Cluster ACB has lower LLR than cluster ADB, but cluster ACB has greater cohesion than cluster ADB, due to the the larger population of weak link C, and neither one dominates the other (see the graph in Figure 5.2).

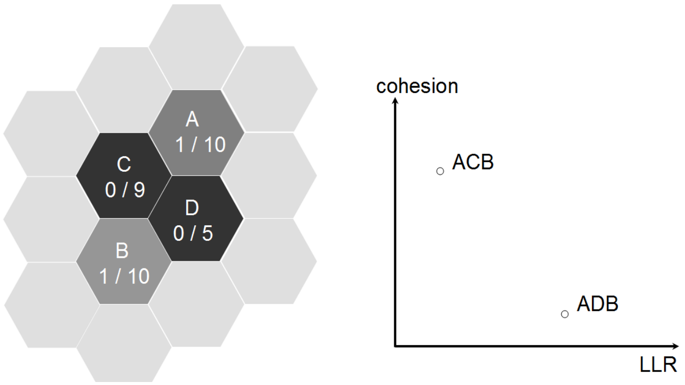


Figure 5.2: LLR and disconnection nodes cohesion for the clusters ACB and ADB.

5.4 The simultaneous compactness and disconnection nodes multi-objective scan

In our second implementation of the multi-objective treatment, one of the objectives is defined as $f_1(z) = k(z).LLR(z)$, while the second objective is defined as $f_2(z) = c(z).LLR(z)$, for each cluster z. This is motivated by the single-objective compactness penalized spatial scan statistic discussed in section 4.1. The motivation for building this algorithm is to evaluate clusters based simultaneously on their regularity of shape and abscence of disconnecting nodes. The algorithm automatically discards those clusters which do not have high cohesion (due to the presence of disconnecting nodes) and at the same time are very irregularly shaped.

Next we discuss the computation of the significance of each cluster in the non-dominated set, and how it is used to determine the best solution. The multi-objective algorithm is first executed to find the non dominated set of the best clusters candidates in the quadrant $(0, \infty) \times (0, \infty)$ of the space $k(.).LLR(.) \times c(.).LLR(.)$.

Considering clockwise sense starting from the topleft cluster we have a cluster with high cohesion and low compactness (1), high cohesion and compactness (2), low cohesion and high compactness (3), low cohesion and compactness (4) (see figure 5.3).

Considering a set of non-dominated solutions clusters like cluster (4) with simultaneous low cohesion and compactness would be automatically discarded.

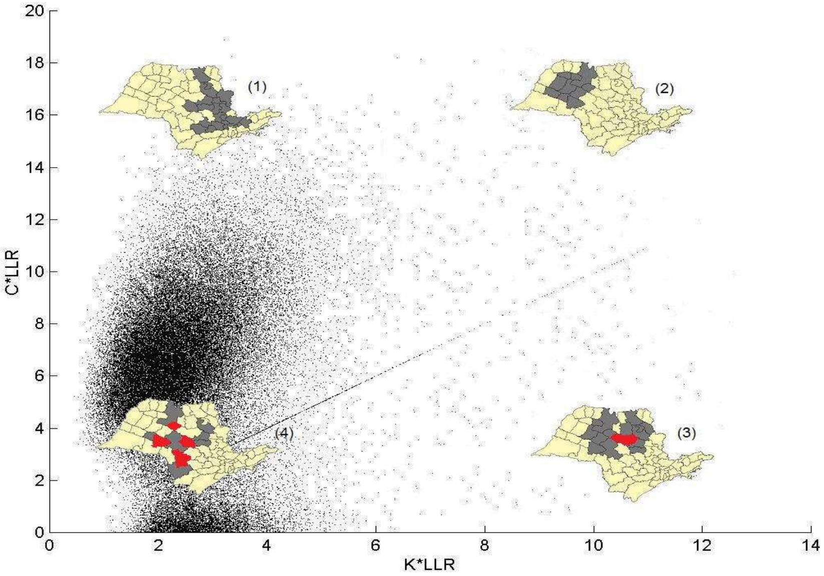


Figure 5.3: LLR and disconnection nodes cohesion.

Chapter 6

Genetic Algorithms for Cluster Detection and Spatial Cluster Inference

In this chapter we describe how we use a genetic algorithm in a cluster detection algorithm. Actually we use a genetic algorithm in two versions. In the first, a single-objective version. We use a genetic algorithm to optimize a objective function given by spatial scan Statistic (Duczmal (2007) [27]). In the second, a multi-objective version, a genetic algorithm is used to optimize two objective functions. In Duczmal (2008) [26] one of the objective functions is the spatial scan statistic and the second is the compactness. In this work we propose three new versions of a multi-objective genetic algorithm. The first version uses the Spatial Scan Statistic as one objective and the non-connectivity regularity function as the second objective. In the second version uses the Spatial Scan Statistic as one objective and the new cohesion function as the second objective. In the third version we use the cohesion function penalized spatial scan statistic as one objective function and the compactness function penalized spatial scan statistic as the second objective. The inference about the detected cluster candidates use a new attainment function introduced in da Fonseca et al. (2001) [18], Fonseca et al. (2005) [35] and Cançado (2009) [12].

A genetic algorithm (GA) uses ideas derived from biological evolution to search for the best solutions of an optimization problem, simulating the mechanisms of random variation and adaptive selection. These mechanisms are called operators and usually include operations of crossover, mutation and selection. It is well known that the design of these operators affects the performance of the algorithm. Particularly, specifically tailored operators work better than generic ones, because they take advantage of the problem structure (Carrano et al. (2006) [13]).

6.1 Single-Objective Genetic Algorithms

Conley et al. (2005) [15] proposed a genetic algorithm to explore a configuration space of multiple aglomerations of ellipses for point data sets. The method employed a strategy to "clean-up" the best configuration found in order to geometrically simplify the cluster. In Sahajpal (2004) [77] a genetic algorithm is used to find clusters in point data sets shaped as intersections of circles of different sizes and centers.

The genetic algorithm employed in this work uses the same crossover and mutation operators described in Duczmal et al. (2007) [27] and Duczmal (2008) [26]. These operators were designed specifically for the spatial clustering problem and they have proved to be very efficient for this problem. The single-objective algorithm described in Duczmal (2007) [27] aimed to maximize Kulldorff's spatial scan statistics (4.1) over the set of potential clusters, starting from a set of zones representing individuals of an initial population. The initial population is successively modified for a number of generations, according to the operators' rules. The *cross-over* operator creates new individuals, mixing the features of two randomly chosen parents A and B, which are themselves zones from the previous generation that have at least of region in common. Offspring are thus produced, consisting of a set of intermediate zones between parents A and B (see Figure 6.1). The *mutation* operator introduces few random perturbations in individual zones, either adding or removing one random region. Both operators increase the variance of the population. The *selection* operator ranks the zones according to some objective function values, and chooses the individuals that would remain at the next generation, maintaining a fixed genetic population size. As the algorithm advances through new generations, it is expected to find individuals with increasingly higher values of the objective function.

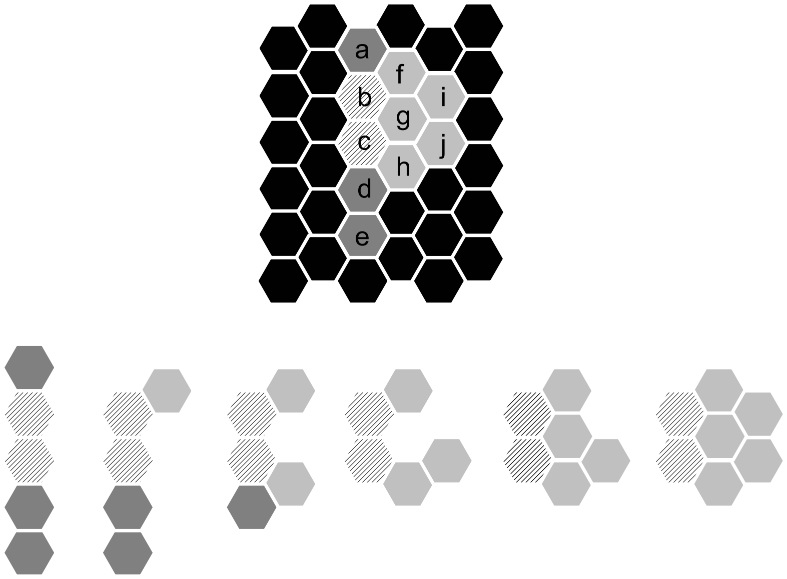


Figure 6.1: Crossover between parents $A = \{a, b, c, d, e\}$ and $B = \{b, c, f, g, h, i, j\}$ in the map (above) generated the offspring formed by the four intermediate zones (below). The offspring constitutes a randomly chosen path in the space of configurations among all the possible paths between the extreme zones, which are parents A and B.

The graph-related operations are minimized by means of a fast offspring generation and evaluation of Kulldorff's spatial scan statistic. A geometric compactness penalty function is employed to avoid excessive irregularity of cluster geometric shape.

6.2 Multi-Objective Genetic Algorithms for Cluster Detection

We now describe the multi-objective optimization approach to the problem of finding spatial clusters. The genetic algorithm described in Section 6.1 will be modified to deal simultaneously with the two quantities: the selected regularity function (either the compactness or the disconnection node cohesion, for instance), and Kulldorff's original spatial scan LLR(.). The selected regularity function will no longer be used as a penalty correction, but instead as a new objective function. That approach simplifies the problem and allows a stronger grasp of the question of finding the "best" cluster solution. The compactness regularity function was used in this context in Duczmal et al. (2008) [26]. The Yiannakoulias regularity function defined in section 4.3 and the novel disconnecting node cohesion function described in chapter 5 will now be used.

For the multi-objective algorithm, the initial population construction, the crossover and mutation operators are identical to those used in the single-objective genetic algorithm, and the reader is referred to Duczmal et al. (2007) [27] and Duczmal et al. (2008) [26] for details. The selection procedure is modified in order to approximate the structure employed by the widely used *Non dominated Sorting Genetic Algorithm* (NSGA-II), described in Deb et al. (2002) [19]. The NSGA-II incorporates the following features:

1. *Non-dominated sorting*: consists of sorting the solutions according to the *non-dominance level*. Individuals belonging to the original set of non-dominated solutions are assigned as level 1. Level 2 is assigned to those individuals belonging to the set of non-dominated solutions obtained after the removal of the individuals of level 1, and so on.

2. *Crowding-distance*: is based on the distance between one individual and its immediate neighbors to the left and to the right (when two objectives are involved).

3. *Binary tournament*: consists of choosing two individuals randomly and comparing them according to some fitness function. The one with best fitness evaluation is selected.

In this context, solutions belonging to lower dominance levels are better than solutions situated at higher levels. An individual belonging to level 1 is not dominated by any of the solutions while an individual belonging to level 2 is dominated by at least one individual (of level 1). The crowding distance is used as a measure of occupation in the neighborhood of a solution in the objectives' space. If one solution has high crowding distance evaluation it means that its neighbors are far away and therefore its neighborhood is sparsely populated. Such individual should have more chances to remain in the population in the evolution process. Alternatively, one individual with low crowding distance evaluation belongs to a well represented area. The use of crowding distance helps us avoid situations where the obtained Pareto-set is too concentrated on one small part of the real Pareto surface, leading the algorithm to the situation where the non-dominated points are more uniformly spread (refer to figure 6.2). For the binary tournaments, the level of the solutions is employed as the fitness function. If both solutions compared belong to the same level, the tie is broken by the crowding distance: the one with higher crowding distance evaluation is selected. The binary tournament is performed with replacement, i.e., all individuals take part in all random draws.

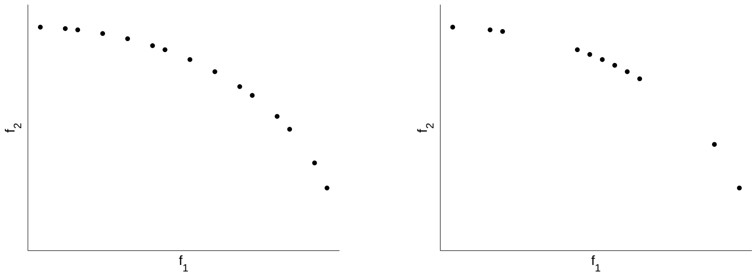


Figure 6.2: Pareto sets - left graph: The Pareto set is well represented by its component points; right graph: the Pareto set is not well represented in some parts where points are absent.

We now describe the general structure of our NSGA-II, including the selection procedure.

1. The initial population P_0 of size N is generated and evaluated regarding the objectives functions. Non-dominance levels and crowding distances are also computed for all the N individuals of P_0.

2. While some stopping criterion is not achieved, the population of the $(i+1)^{th}$ generation is obtained from i^{th} generation following the steps:

    - From P_i we perform N binary tournaments obtaining a list of N selected individuals. These individuals take part in crossover and mutation, generating a list Q_i of M new individuals.
    - A combined population $C_i = P_i \cup Q_i$ of size $N + M$ is formed and the new levels and crowding distances are computed for C_i.
    - Individuals of the lowest levels are inserted into the new population P_{i+1} until we reach N individuals. In general, the last inserted level, say level l, will not totally fit inside the new population, so

the individuals of the last inserted level l are inserted according to the crowding distance criterion, from higher to lower ones. The crowding distance of the selected individuals must be updated and non-dominance levels of P_{i+1} are preserved from C_i.

Note that some individuals may have more than one selected copy after the binary tournament. Particularly, individuals of the lower levels have higher probabilities of being represented with two or more clones, while individuals at higher levels have lower probabilities of being selected. With this structure our GA is very close to the original NSGA-II. The difference is due to our crossover operator nature. The original algorithm performs $N/2$ crossovers at each generation, obtaining a list of N individuals, since most crossover operators generate two new individuals. Since it is not always possible to execute our crossover for a given pair of solutions, we make a maximum of c_{max} crossover attempts at each generation with randomly chosen pairs from the list obtained by the binary tournament, or a maximum of $N/2$ well-succeeded crossovers. So we do not guarantee that $N/2$ crossovers will take place, and even if we attain $N/2$ well-succeeded crossovers we typically do not obtain N new individuals, because each crossover can generate a varying number of new individuals. This explains why the Q_i set has a varying number M of individuals for different generations. We observed a considerable performance gain using the NSGA-II over the simpler Genetic Algorithm employed in Duczmal et al. (2007) [27] and Duczmal et al. (2008) [26].

6.3 Attainment function

Consider a bi-objective maximization problem for the objective functions f_1, f_2. Let $\mathbb{E} = \{x_j, j = 1, ..., Q\}$ be the set of all evaluated solutions, and define its image $\mathbb{I} = \{Y_j = (f_1(x_j), f_2(x_j)), j = 1, ..., Q\}$ contained in the objective space $\mathbb{R}^2$. As mentioned is section 6.2, the solution x_j is called *non-dominated* if x_j is not dominated by any other solution in $\mathbb{E}$. Let $\{x_j^*, j = 1, ..., q\} \subset \mathbb{E}$ be the set of non-dominated solutions of $\mathbb{E}$. The subset $\mathbb{Y} = \{Y_j^* = (f_1(x_j^*), f_2(x_j^*)), j = 1, ..., q\} \subset \mathbb{I}$ is defined as the the *outcome* of a single run of a bi-objective algorithm.

We can associate a boundary to $\mathbb{Y}$ which splits the objective space in two regions R_1 and R_0: R_1 is the region consisting of points dominated by at least one point of $\mathbb{Y}$ plus the points that equal some point of $\mathbb{Y}$ and R_0 consists of the points that are not dominated by any of the points within $\mathbb{Y}$ (see Figure 6.3). When the solution x is dominated by at least one solution of a given outcome $\mathbb{Y}$, we say that x is *attained* by $\mathbb{Y}$. In Figure 6.3, any solution located in the region R_1 is attained by $\mathbb{Y}$. Now consider n runs of the algorithm. As each run produces distinct outcomes we will obtain multiple boundaries, as in Figure 6.4(a).

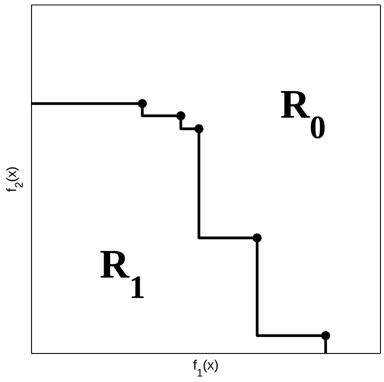
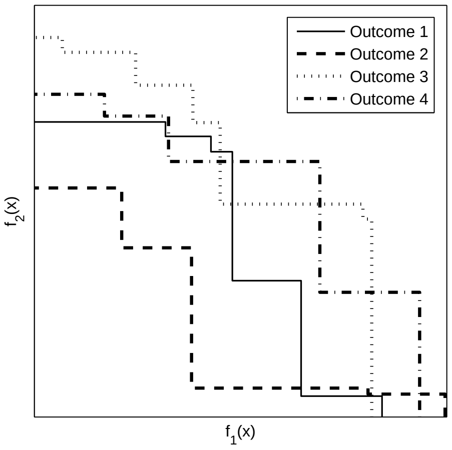
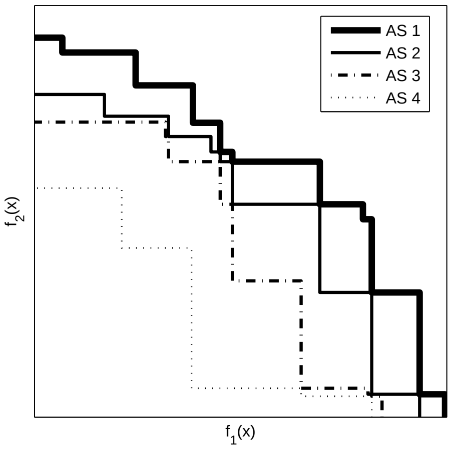


Figure 6.3: The attainment surface splits the objective space in two regions.

Points lying in the upper right of the figure were not attained in any of the runs. Points that lie in the lower left were attained in all the runs. And points lying between different outcomes were attained in some runs but not in others. So we can split the space in $n + 1$ types of regions according to the frequency with which these regions are attained. The boundaries of these regions are called *attainment surfaces* (ref. Figure 6.4(b) and Fonseca et al. (2005) [35]). These frequencies are used to estimate the probability of attaining a point in the objective space, when a large number of runs is executed.



(a)



(b)

Figure 6.4: (a) Outcomes obtained by multiple runs of a biobjective algorithm and (b) the corresponding estimated attainment surfaces.

The attainment function described in da Fonseca et al. (2001) [18] and Fonseca et al. (2005) [35] evaluated at Y can be estimated by the outcome sets $\mathbb{Y}_1, ..., \mathbb{Y}_n$ obtained through n independent runs of the algorithm, as

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

where the symbol "$\trianglerighteq$" means that $\mathbb{Y}_i$ attains Y and $\mathbf{I}$ is the indicator function having value 1 if $\mathbb{Y}_i \trianglerighteq Y$, and value zero otherwise.

In the specific problem of the present work we are interested in estimating the p-value of non-dominated candidate cluster solutions represented by points in the (LLR, Mes) objective space, where Mes is the desired measure, such as compactness, non-connectivity or cohesion, discussed in the previous sections.

Formally, we define $A(Y)$ as the $\lim_{n\to\infty} A_n(Y)$ when it exists. Now, given $0 < p \leq 1$ the *isoline* is defined as the inverse image $A^{-1}(p)$. For sufficiently smooth conditions, $A^{-1}(p)$ is an 1-dimensional surface dividing the objective space into two regions R_0 and R_1, such that if $Y \in R_1$ then $A(Y) > p$, and if $Y \in R_0$ then $A(Y) \leq p$. In practice, given n outcome sets $\mathbb{Y}_1, ..., \mathbb{Y}_n$, we can construct approximations of the p-value isolines for every $p = i/(n+1), i = 1, ..., n$ through the estimated attained function $A_n(Y)$. The example of Figure 6.5 displays some p-value isolines resulting from $n = 1000$ outcome sets under the null hypothesis. The outcome points are displayed in gray.

When a stochastic algorithm is used, only part of the potential set of solutions is evaluated, and there is no guarantee that the optimal non-dominated solutions are found. This of course could lead to a biased estimation of the significance, producing underestimated p-values. Thus the computed p-values are in fact lower bounds for the theoretical p-values.

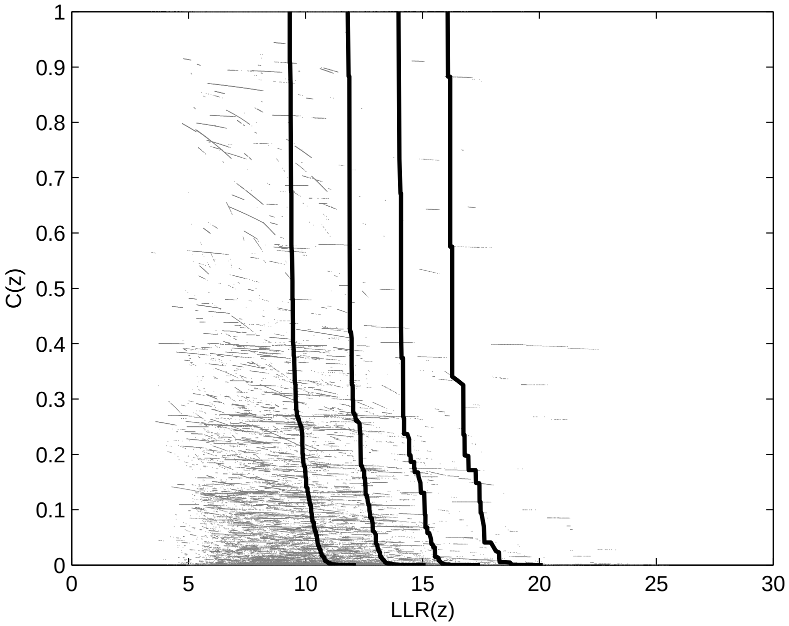


Figure 6.5: The 0.316, 0.1, 0.032 and 0.01 p-value isoline curves for the null hypothesis Monte Carlo simulation, using 1,000 Pareto-sets.

6.4 Spatial Cluster Inference

If we know the probability distribution of the spatial scan statistic under the null hypothesis of cluster non-existence we could determine a critical value such that the significance level (typically 5%) represents the probability of the scan statistic assumes values greater that the critical value. Since, in principle, that probability distribution is unknown, we use Monte Carlo simulations (Dwass(1957) [32]) in order to obtain an empirical distribution of the scan statistic values under the null hypothesis. To make one Monte Carlo simulation, first we distribute a fixed total number of cases throughout the regions of the study area. The cases distribution, conditioned on the total number of cases is made according a multinomial distribution where the probability of an individual to become a case in come region is proportioned to its population. Then the scan statistic is calculated for the most likely cluster given the simulated cases distribution. This procedure is repeated n times and the obtained scan statistic values are ranked (the value corresponding to the 95% quantile is the estimate of the critical value at a 5% significance level). Given the scan statistic value of the observed cases map, the estimate of its p-value is $\frac{n_{obs}}{n+1}$, where n_{obs} is its ranking position among the $n+1$ values (where n is the number of simulated values).

In the multi-objective case when we run the genetic algorithm for a given cases distribution the outcome solution is a set of non-dominated solutions or best clusters candidates. Under the null hypothesis and conditioned on a fixed total number of cases we distribute the cases among the map regions according a multinomial distribution where the probability of an individual to become a case in come region is proportioned to its population. With the outcome of n simulations we use the attainment function to construct p-value isolines in the (LLR, Mes) objective space as explained in section 6.3. To determine the statistical significance of the best cluster candidates obtained from the observed cases map we just plot those candidates in the objective space with the p-value isolines.

Chapter 7

Numerical Evaluations

In this chapter we compare numerically the disconnection nodes cohesion scan (DN), the geometric compactness scan (GC), the non-connectivity scan (NC) and the no-penalty genetic scan (NP). We also compare the corresponding multi-objective scans: the multi-objective disconnection nodes cohesion scan (MDN), the multi-objective geometric compactness scan (MGC), the multi-objective non-connectivity scan (MNC) and the simultaneous multi-objective geometric compactness and disconnection nodes cohesion scan (MGD). We evaluate their power of detection, sensitivity and positive predicted value (PPV).

7.1 New England benchmark tests

A benchmark dataset for real data population for breast cancer of the Northeastern US is used (Duczmal et al. (2006) [25]). This benchmark consists of 245 counties in 10 states and the District of Columbia, with a total population at risk of 29,535,210 women. The map of Figure 7.1 display the counties population quantiles by shades of gray. Nine simulated irregularly shaped clusters, A-F, NY, BOS and $D.C.$, are displayed in the remaining three maps of (Figures 7.2,7.3,7.4). These clusters were chosen with the purpose of testing the limits of the algorithms for some very irregular cluster shapes. Clusters NY, BOS and $D.C.$ are located in highly populated areas, contrasting with the remaining clusters, which are located in rural or mixed areas defined roughly by geographic features such as rivers or shores (see Duczmal et al. (2006) [25]). The lighter shade regions indicate disconnection nodes inside the clusters. All clusters have at least one disconnection node, except B and BOS which have $c(z) = 1$. Clusters F, C and E have the lowest disconnection nodes cohesion.

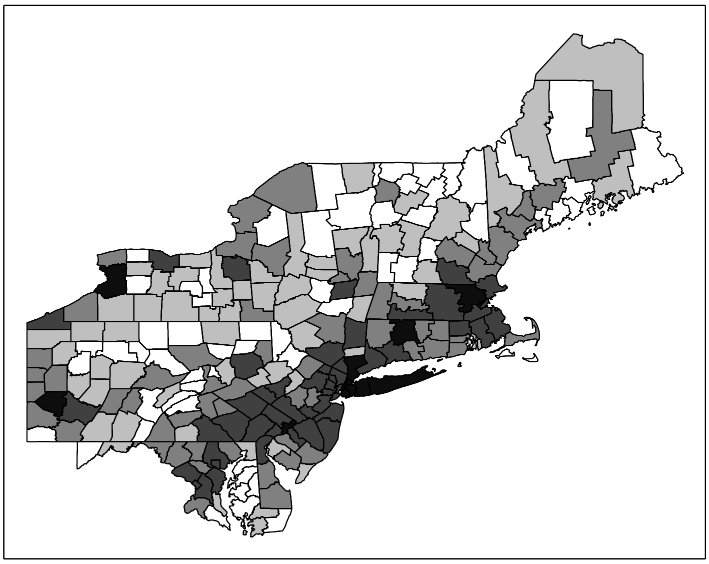


Figure 7.1: Map with counties populations quantiles by shades of gray for 245 counties northeastern U.S. map, shades indicate counties populations.

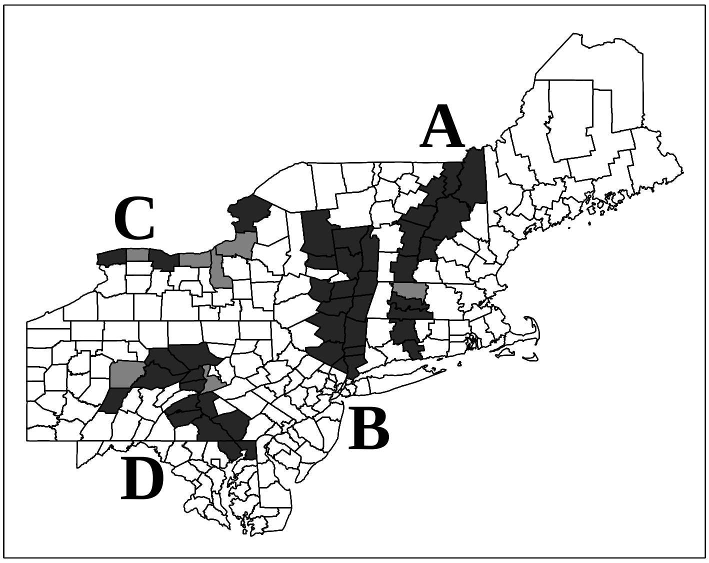


Figure 7.2: Simulated data clusters for the 245 counties northeastern U.S. map, the clusters A, B, C and D were used in the power evaluations. Lighter shades indicate disconnection nodes.

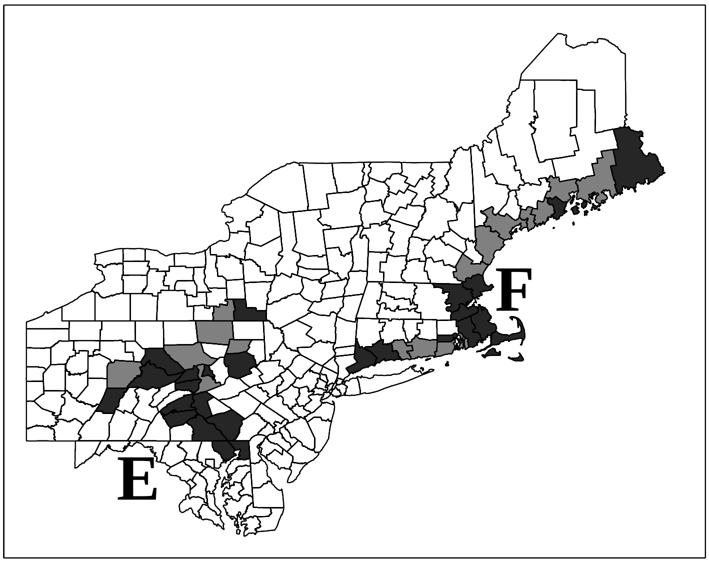


Figure 7.3: Simulated data clusters for the 245 counties northeastern U.S. map, the clusters E and F were used in the power evaluations. Lighter shades indicate disconnection nodes.

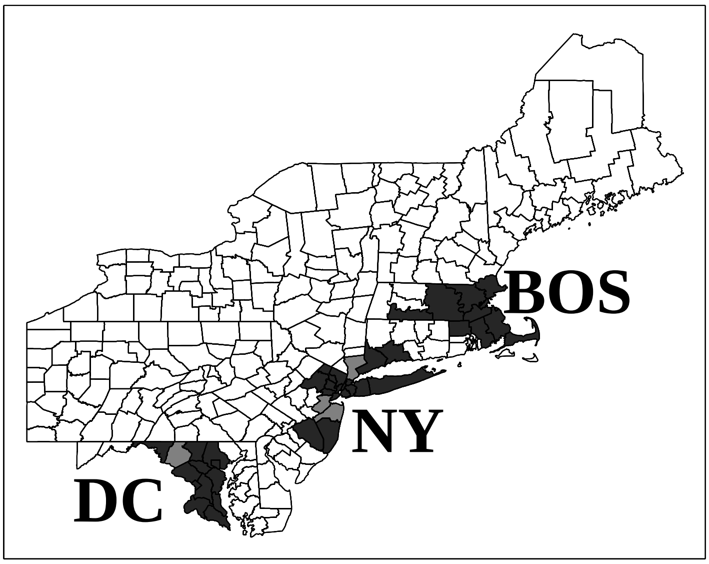


Figure 7.4: Simulated data clusters for the 245 counties northeastern U.S. map, the clusters DC, NY and BOS were used in the power evaluations. Lighter shades indicate disconnection nodes.

From now on, those clusters will be called *real* clusters, in contrast to the *detected* clusters found by the algorithms. For each simulation of data under these nine alternative hypotheses, 600 cases are distributed randomly according to a Poisson model using a single cluster; we set a relative risk equal to one for every cell outside the real cluster, and greater than one and identical in each cell within the cluster. The relative risks for each cluster are defined such that if the exact location of the real cluster was known in advance, the power to detect it should be 0.999 (see Kulldorff et al (2003) [50]).

Given an alternative hypothesis model, the estimate *power* in the single-objective case is the proportion of values of the objective function greates than critical value.

In the multi-objective algorithm, given an alternative hypothesis model, $5,000$ runs produce the corresponding non-dominated sets, which are joined and compared to the 0.05 isoline, obtained under null hypothesis through $10,000$ Monte Carlo replications with the attainment function, as explained in section 6.3. The proportion of non-dominated sets which have at least one point located to the right of the 0.05-value isoline is an estimate of the *power* of the algorithm for that particular alternative hypothesis model.

Additionally, we perform a set of four null hypotheses simulations of $10,000$ runs corresponding to MGC, MNC, MDN and MGD algorithms.

The measures of sensitivity and PPV (Positive Predicted Value) also serve to evaluate the quality of the cluster detection process. These measures are defined in the terms of the population size. We define sensitivity and PPV as:

$$Sensitivity = \frac{Pop(\text{Detected Cluster} \cap \text{Real Cluster})}{Pop(\text{Real Cluster})}$$

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

For the non-penalty single-objective scan, the three measures, namely, detection power, sensitivity an PPV were computed for the most likely cluster in each replication. For the multi-objective scans, they were computed based on the cluster within the Pareto-set which maximized all of those measures.

Tables 7.1 presents the average power, sensitivity and PPV for $5,000$ replications of each of the nine alternative hypotheses obtained with the single-objective algorithms and 7.2 presents the results obtained with the multi-objective algorithms.

When we compare tables 7.1 and 7.2, we can observe a significant gain using the multi-objective strategy. All the multi-objective scans show consistently better performance, regarding power, PPV and sensitivity, in compares on with the single-objective penalized and non-penalized scans.

From Table 7.2 we conclude this section observing that the power of detection of the MGD scan is better in some situations compared to the other three scans. Moreover the MGD scan presented the worse power value among the four multi-objective scans in the specific case of cluster “F”; this particular cluster is highly irregular and presents many disconnection nodes. Both MGC and MNC scans have better PPV performance than the MDN and MGD scans; the MDN scan has better sensitivity in most situations, compared with the other scans. It must be noted that the MNC scan is significantly faster. The MGC scan and the MGD scan have presented consistently better results regarding all three performance measurements.

Table 7.1: Power, positive predicted value and sensitivity comparisons for the mono-objective algorithms.

	Power				PPV				Sensitivity			
cluster	NP	GC	NC	DN	NP	GC	NC	DN	NP	GC	NC	DN
A	0.838	0.822	0.881	0.839	0.624	0.578	0.665	0.619	0.796	0.551	0.792	0.767
B	0.882	0.843	0.926	0.898	0.699	0.691	0.786	0.765	0.707	0.598	0.784	0.743
C	0.827	0.814	0.826	0.667	0.625	0.344	0.659	0.582	0.851	0.360	0.796	0.607
D	0.896	0.840	0.922	0.877	0.696	0.616	0.771	0.734	0.668	0.506	0.713	0.668
E	0.874	0.778	0.885	0.822	0.719	0.633	0.762	0.704	0.534	0.414	0.544	0.508
F	0.629	0.433	0.585	0.510	0.664	0.314	0.650	0.565	0.583	0.170	0.523	0.430
NY	0.759	0.747	0.819	0.868	0.898	0.621	0.929	0.941	0.580	0.364	0.650	0.643
BOS	0.792	0.834	0.864	0.892	0.781	0.389	0.827	0.861	0.747	0.295	0.806	0.841
D.C.	0.803	0.903	0.877	0.901	0.788	0.518	0.865	0.887	0.725	0.426	0.791	0.802

- NP - AG single-objective with $LLR(z)$;
- GC - AG single-objective with $LLR(z).k(z)$;
- NC - AG single-objective with $LLR(z).y(z)$;
- DN - AG single-objective with $LLR(z).c(z)$;

Table 7.2: Power, positive predicted value and sensitivity comparisons for the multi-objective algorithms.

	Power				PPV				Sensitivity			
cluster	MGC	MNC	MDN	MGD	MGC	MNC	MDN	MGD	MGC	MNC	MDN	MGD
A	0.950	0.939	0.946	0.951	0.902	0.806	0.746	0.813	0.827	0.843	0.864	0.808
B	0.954	0.967	0.967	0.963	0.895	0.906	0.832	0.860	0.801	0.857	0.827	0.786
C	0.934	0.910	0.932	0.928	0.813	0.778	0.762	0.744	0.860	0.855	0.881	0.818
D	0.962	0.963	0.972	0.969	0.860	0.891	0.823	0.823	0.740	0.769	0.781	0.796
E	0.947	0.942	0.964	0.953	0.868	0.876	0.822	0.805	0.609	0.596	0.616	0.560
F	0.752	0.733	0.824	0.700	0.796	0.777	0.745	0.686	0.583	0.593	0.645	0.522
NY	0.891	0.906	0.923	0.908	0.961	0.973	0.965	0.977	0.689	0.743	0.715	0.690
BOS	0.918	0.924	0.943	0.956	0.939	0.896	0.888	0.920	0.837	0.873	0.885	0.851
D.C.	0.955	0.933	0.937	0.960	0.977	0.927	0.899	0.956	0.880	0.874	0.849	0.849

- MGC - AG multi-objective with $LLR(z)$ and $k(z)$;

- MNC - AG multi-objective with $LLR(z)$ and $y(z)$;

- MDN - AG multi-objective with $LLR(z)$ and $c(z)$;

- MGD - AG multi-objective with $LLR(z).k(z)$ and $LLR(z).c(z)$.

7.2 Chagas disease clusters

Chagas' disease is caused by the parasite Trypanosoma cruzi. It is transmitted to animals and people by blood-sucking insect vectors (triatomine bugs), which are found only in the Americas. The disease is found chiefly in poor rural areas of Latin America. An individual can be infected if the parasite present in the bug's feces enter the body through mucous membranes, the bite wound itself or others breaks in the skin. Others ways of infection include: consumption of uncooked food contaminated with feces from infected bugs; congenital transmission (from a infected pregnant woman to her baby); blood transfusion and organ transplantation. In the last years, due to better control of the triatomine bugs infestation, the congenital transmission became one of the main transmission mechanism of the Chagas infection. In this work we study the occurrence of Chagas' disease in puerperal women in the state of Minas Gerais, located in Brazil's southeast. The population at risk consists of women that gave birth to babies in the period of July to September, 2006. The new-born babies were blood tested to detect the presence of the Chagas disease antigen, with coverage above 96%. A positive test means that the mother is infected. These tests were conducted through the project PETN-MG (Minas Gerais State Program of New-Born Screening) coordinated by the research group NUPAD-MEDICINA/UFMG from Federal University of Minas Gerais Medical School (`http://www.nupad.medicina.ufmg.br`) in collaboration with Minas Gerais State Health Secretary. The state is divided into 853 municipalities with a total population at risk of $24,969$ women. After a comprehensive screening to eliminate false positives a total number of 113 cases were obtained. In Figure 7.5 the incidence map (cases per 1000 women) for each municipality is shown and in Figure 7.6 and the quantile population map is shown. Most municipalities have zero cases in the period. To detect clusters, we apply the circular scan, the single-penalty mono-objective genetic scan and our four multi-objective scans described in the previous sections.

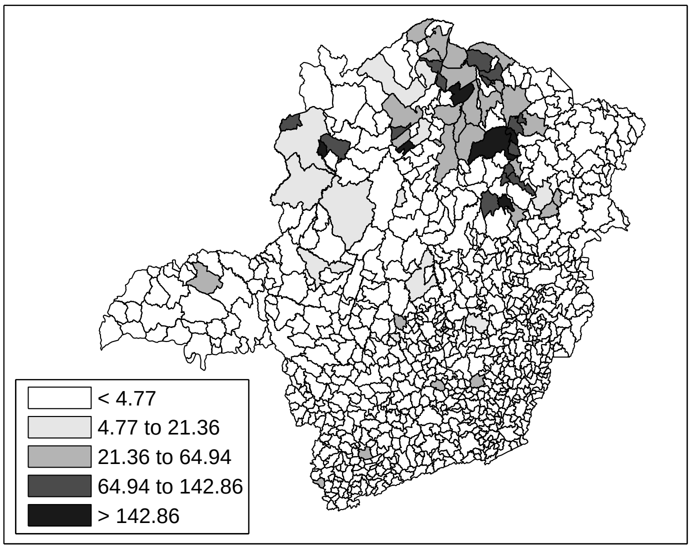


Figure 7.5: Map of rates (per one thousand individuals) of Chagas' disease in the state of Minas Gerais, Brazil.

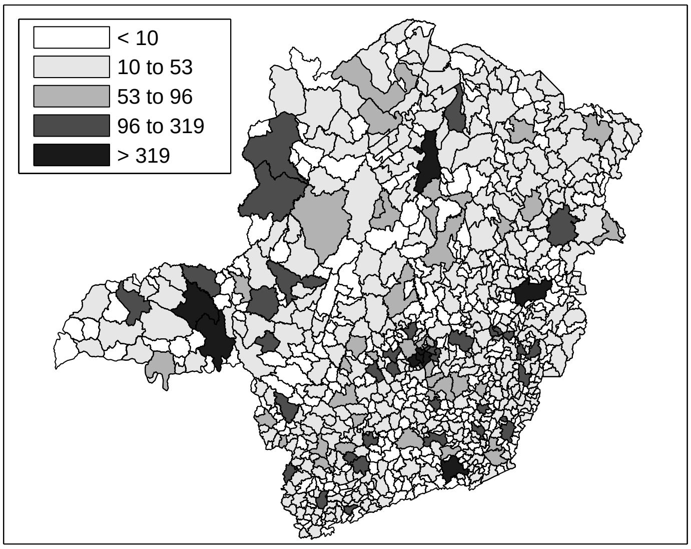


Figure 7.6: Map of populations at risk in the state of Minas Gerais, Brazil.

The primary and secondary cluster detected by the circular scan are shown in Figure 7.7 and described in Table 7.3.

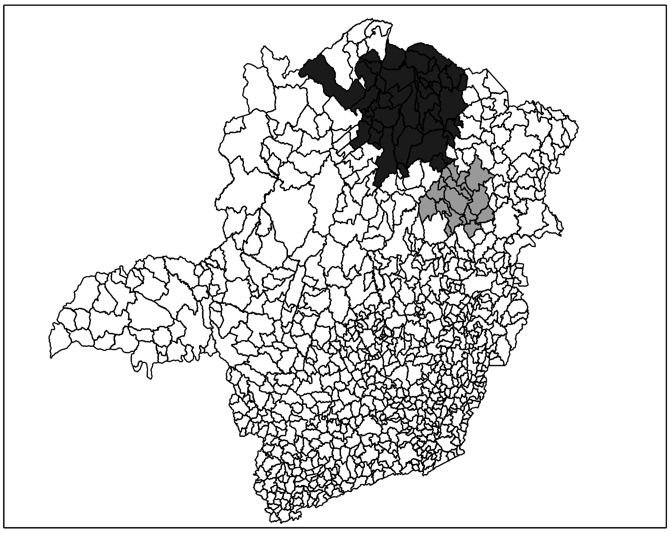


Figure 7.7: Primary (darker shade) and secondary (lighter shade) Chagas' disease clusters detected by the circular scan.

The four graphs of Figures 7.8, 7.9, 7.10, 7.11 display the complete set of non-dominated solutions for the MGC, MNC, MDN and MGD scans respectively. The '×' symbols represent the clusters in the non-dominated solution set for the observed cases map. The MGC, MNC, MDN and MGD scans non-dominated solution sets consist of respectively 150, 63 ,12 and 75 clusters. The gray points at the left part of each graph represents 1000 non-dominated solution sets simulated under the null hypothesis. As explained in section 6.3, for each scan the most likely cluster was selected among the clusters of the non-dominated solution set of the observed cases map according to its smallest estimated p-value. The p-value isolines represents constant p-values for the clusters found under the null hypothesis, ranging from 10^{-3} to 10^{-27}

or less at the rightmost line. Those p-values are estimated through extrapolation from the 1000 null hypothesis non-dominated solutions sets using the attainment function method of section 6.3. Employing the Gumbel semi-parametric model (Abrams (2006) [2] and Duczmal et al. (2008) [26]), we assumed that the p-values decrease according to the logarithm of the LLR. In the specific case of the MGD scan, the extrapolation employs the polar coordinates of the p-value isolines, obtained from the 1000 null hypothesis non-dominated solution sets, using the attainment function method of section 6.3. Of course there is a large amount of uncertainty for the precise location of those very small p-values isolines, but the relevant feature here, namely the overall isolines' slopes, are less prone to extrapolation error. The point representing the most likely cluster is distinguished among the non-dominated solution set according to these slopes.

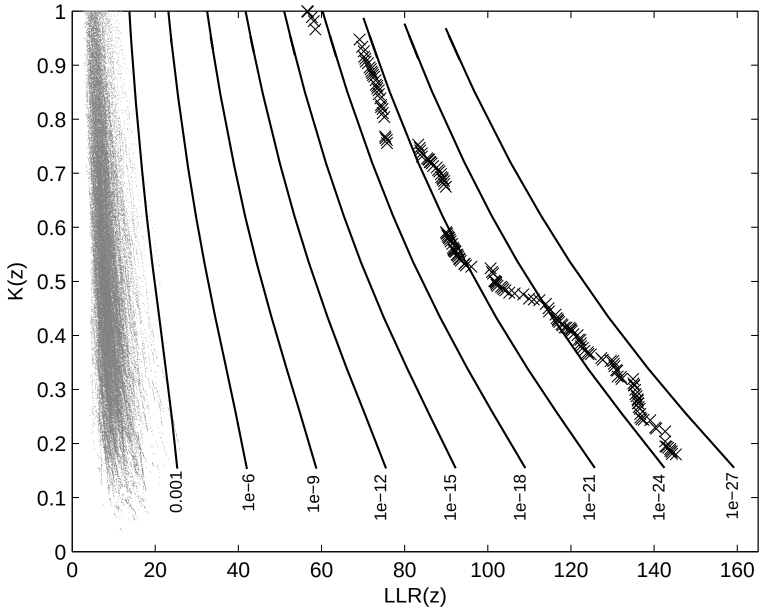


Figure 7.8: Isoline curves and observed clusters (×) found by the MGC scan. Isolines were obtained by extrapolation of the 1,000 Pareto-sets, indicated by gray points.

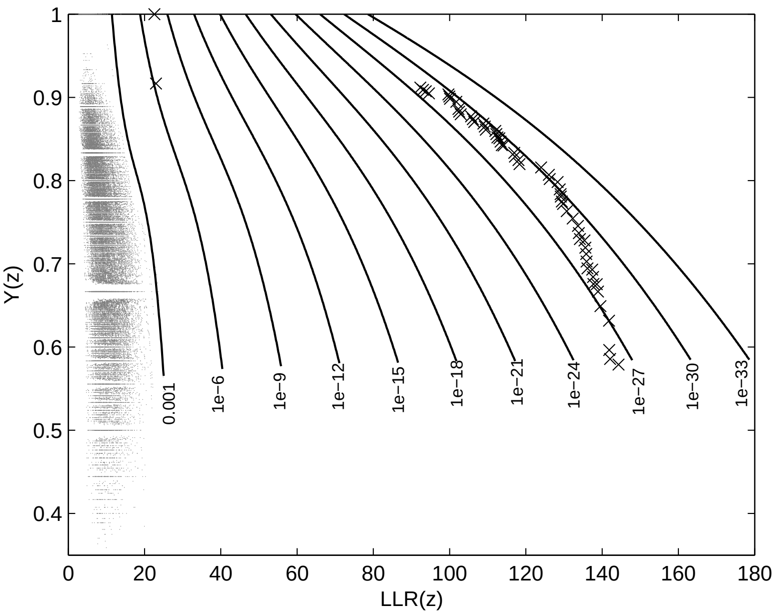


Figure 7.9: Isoline curves and observed clusters (×) found by the MNC scan. Isolines were obtained by extrapolation of the 1,000 Pareto-sets, indicated by gray points.

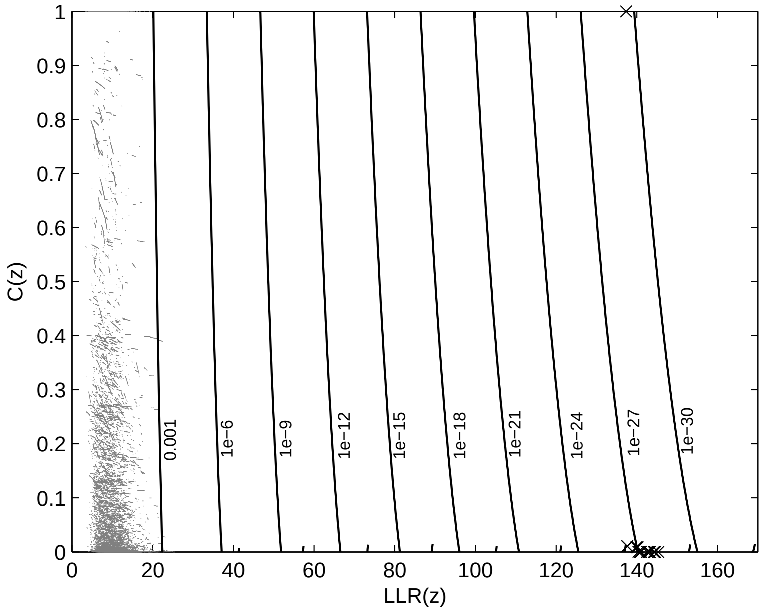


Figure 7.10: Isoline curves and observed clusters (×) found by the MDN scan. Isolines were obtained by extrapolation of the 1,000 Pareto-sets, indicated by gray points.

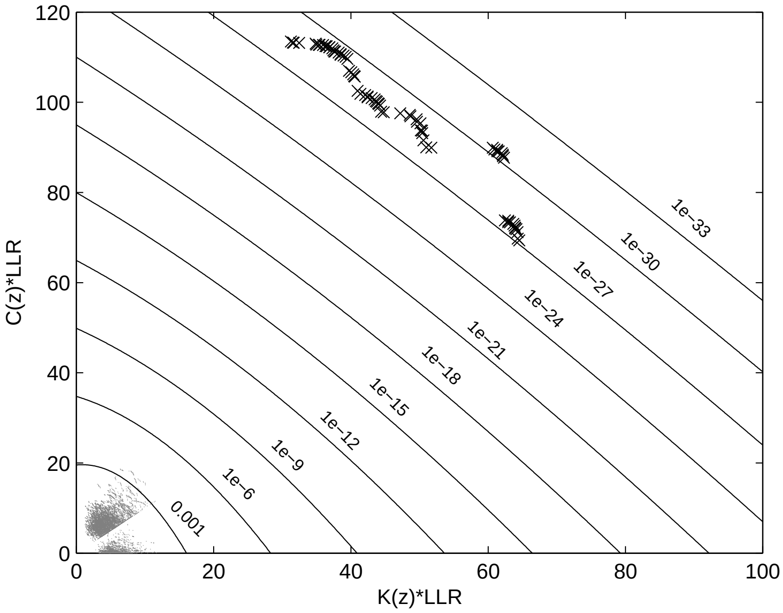


Figure 7.11: Isoline curves and observed clusters (×) found by the MGD scan. Isolines were obtained by extrapolation of the 1,000 Pareto-sets, indicated by gray points.

For the four MDN, MGC, MNC and MGD scans, the most likely clusters found using this procedure are presented in the maps of Figures 7.12, 7.13, 7.14, 7.15. Table 7.3 displays the number of regions, LLR, corresponding values for compactness, non-connectivity and cohesion, population, number of cases, rate and estimated p-value for the most likely clusters found. The p-values shown in the table are conservative estimates based only on counting for the 1000 Monte Carlo simulations, but they are in fact much smaller (less than 10^{-24}), as can be inferred from Figures 7.8, 7.9, 7.10, 7.11.

Table 7.3: Chagas' disease clusters of the Pareto-set of Figures 7.7, 7.12, 7.13, 7.14, 7.15.

	$n(z)$	LLR	measures	pop	cases	rate×1000	p-value
circular prim.	40	87.5	-	1,444	57	39.47	< 0.001
circular sec.	18	13.6	-	453	13	28.70	< 0.001
MGC	40	134.9	$k(z) = 0.319$	1,634	75	45.90	< 0.001
MNC	40	128.3	$y(z) = 0.798$	1,487	71	47.75	< 0.001
MDN	40	137.4	$c(z) = 1.000$	1,732	77	44.46	< 0.001
MGD	25	88.7	$k(z) = 0.698$ $c(z) = 1.000$	720	46	63.89	< 0.001

- MGC - AG multi-objective with $LLR(z)$ and $k(z)$;
- MNC - AG multi-objective with $LLR(z)$ and $y(z)$;
- MDN - AG multi-objective with $LLR(z)$ and $c(z)$;
- MGD - AG multi-objective with $LLR(z).k(z)$ and $LLR(z).c(z)$.

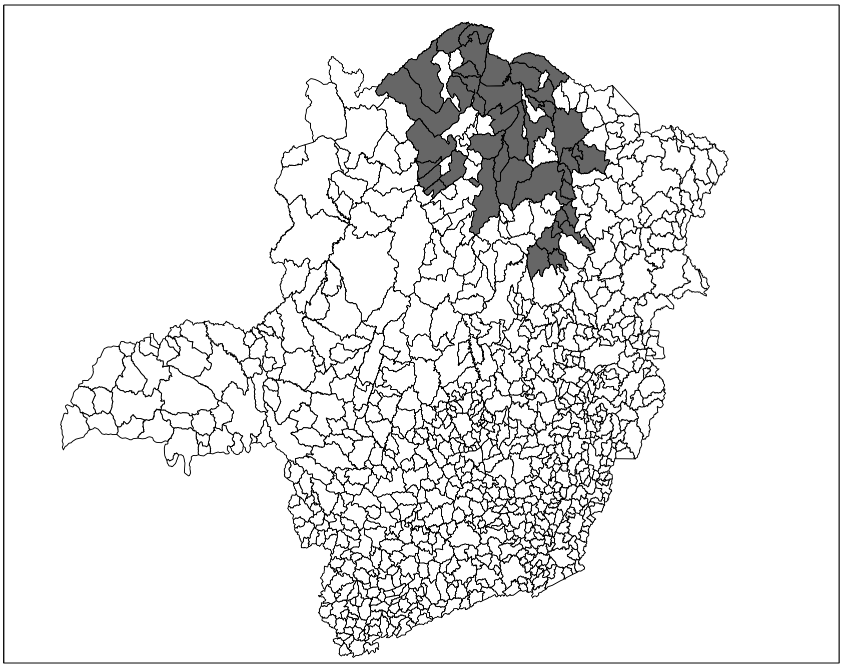


Figure 7.12: Most likely cluster found by the MGC algorithm.

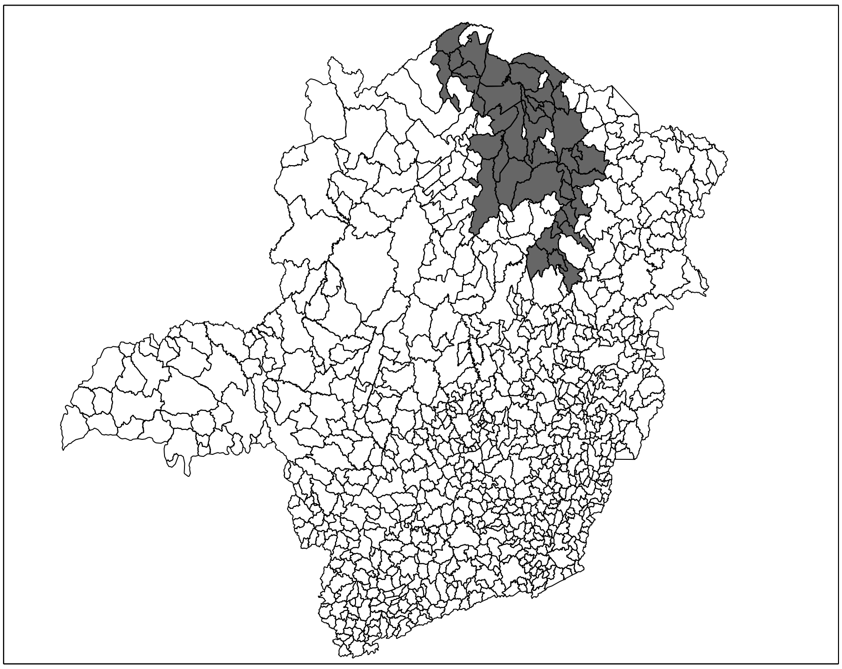


Figure 7.13: Most likely cluster found by the MNC algorithm.

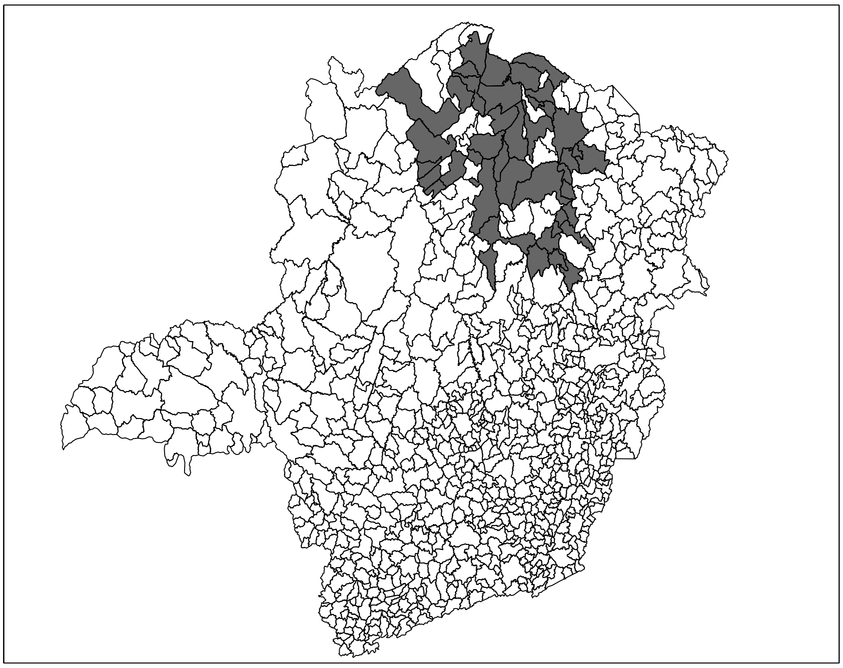


Figure 7.14: Most likely cluster found by the MDN algorithm.

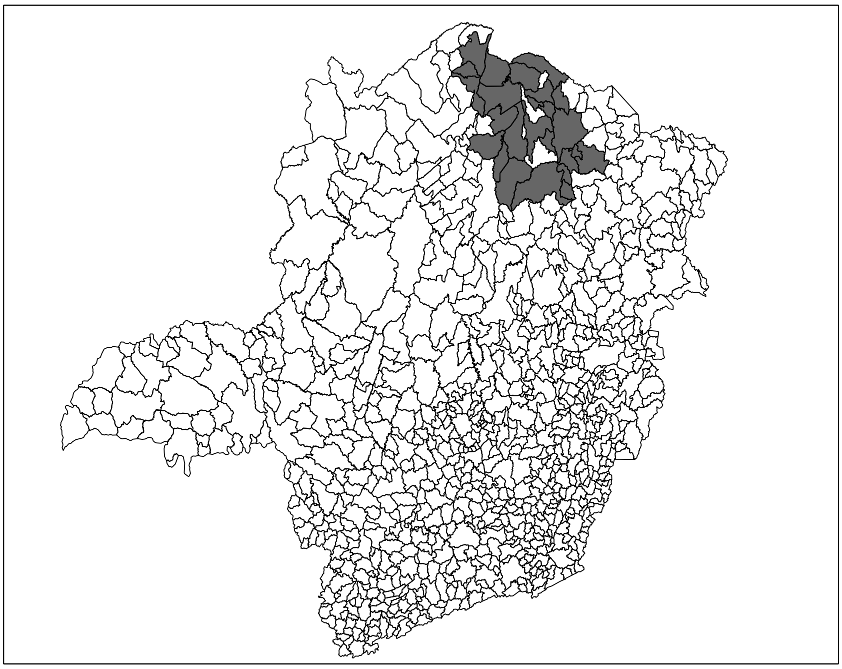


Figure 7.15: Most likely cluster found by the MGD algorithm.

Each of the four maps of Figures 7.16, 7.17, 7.18, 7.19 display simultaneously all clusters in the respective MGC, MNC, MDN and MGD scans' non-dominated solution sets. A gray color coding scheme was used to indicate the proportion of times that each region of the map is present in a non-dominated solution set cluster, from black (the region is present in all clusters) to white (the region is not present in any cluster). This gray scale representation helps the practitioner distinguish those regions which appears in almost all clusters, thus being part of the cluster "core". Note that this core usually does not match exactly the most likely cluster, and constitute an additional tool for identifying the most prevalent regions of the cluster.

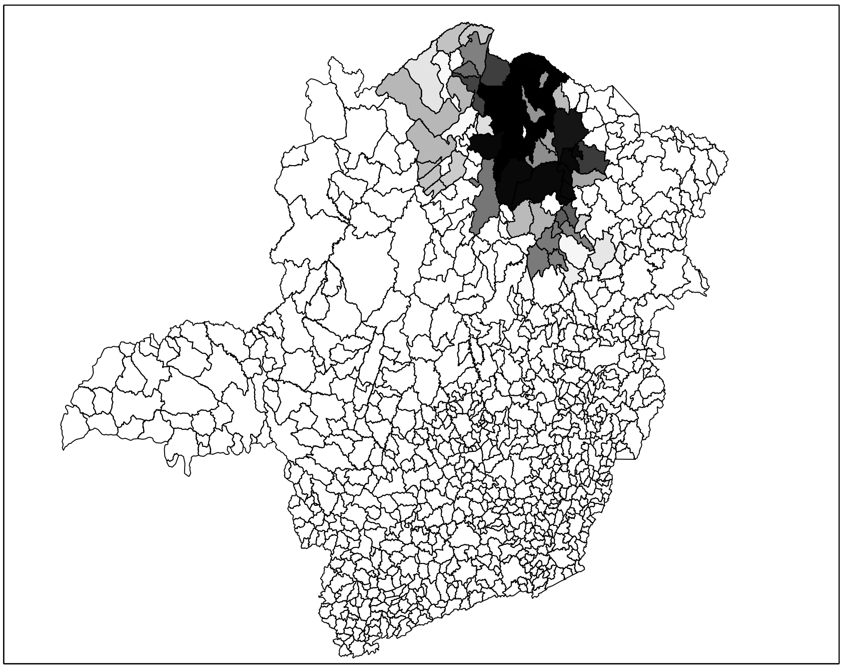


Figure 7.16: Prevalence gray scale map for the observed non dominated solutions obtained by the MGC algorithm.

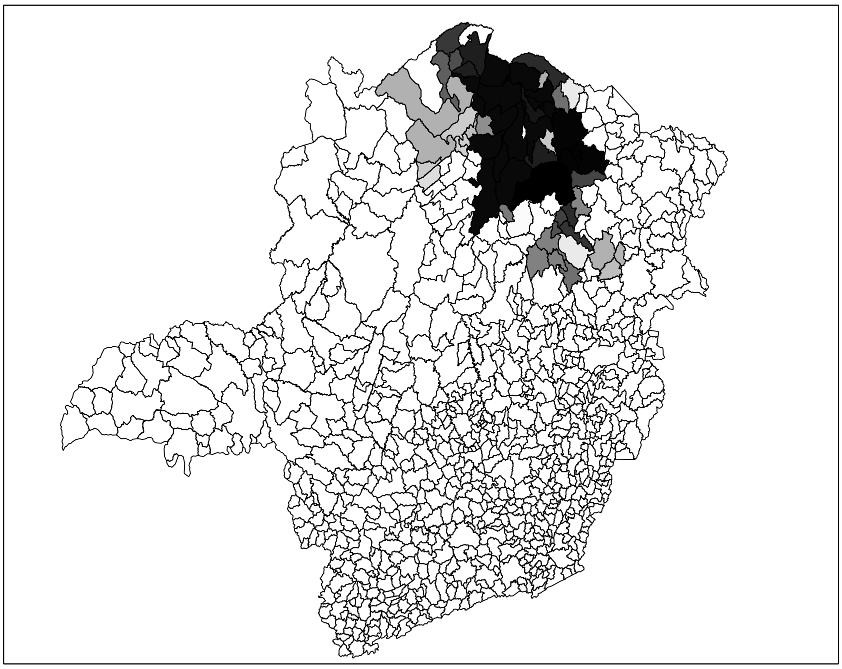


Figure 7.17: Prevalence gray scale map for the observed non dominated solutions obtained by the MNC algorithm.

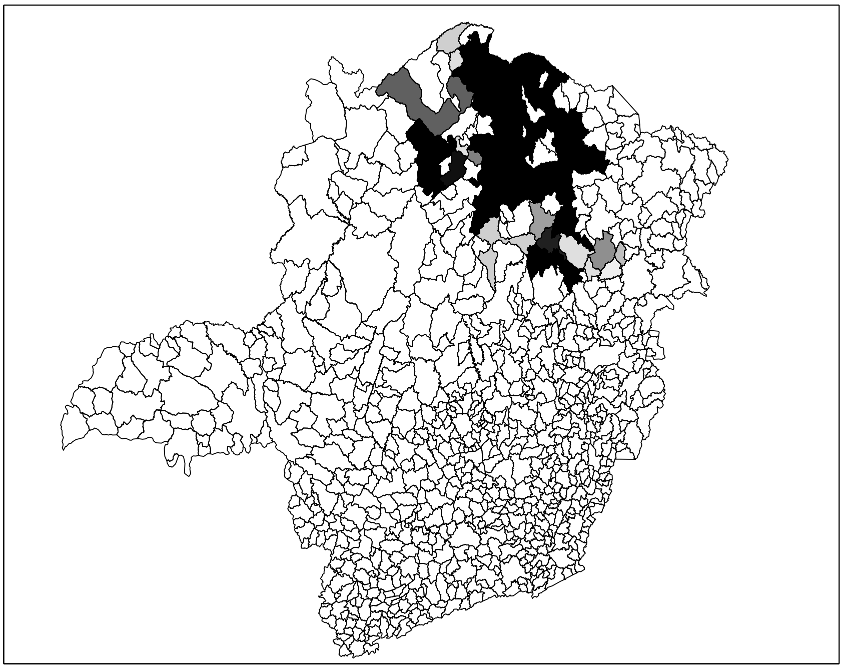


Figure 7.18: Prevalence gray scale map for the observed non dominated solutions obtained by the MDN algorithm.

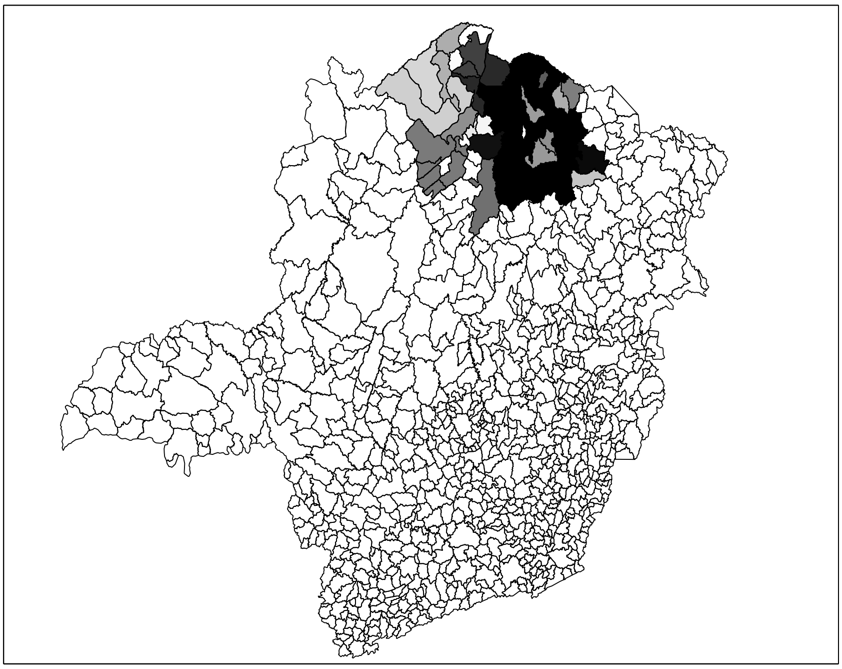


Figure 7.19: Prevalence gray scale map for the observed non dominated solutions obtained by the MDN algorithm.

The MGD most likely cluster is more regularly shaped and has higher rate than the MNC, MGC and MDN most likely clusters. The MDN most likely cluster encompasses more cases and has higher LLR than the other three, but it must be taken in account that the most likely MGD scan solution has only 25 regions. It should be noted that all four MGD, MDN, MGC and MNC scans' most likely clusters are very similar, and in every respect they have better characteristics than the most likely circular cluster found. Even if the primary and secondary circular clusters were added to form a larger cluster, all the four multiobjective scans still present higher rates, number of cases and LLR.

Chapter 8

Conclusions

We compared penalized likelihood and multi-objective methods for the detection and inference of spatial disease clusters employing Kulldorff's spatial scan statistics. Penalized likelihood methods maximize the product of a regularity function by the likelihood ratio scan statistic over the set of potential clusters, employing a genetic algorithm. Regularity functions evaluate a potential cluster, in terms of its geometric shape or topological graph structure, and are used to control the excessive freedom of shape of clusters. The novel disconnection node cohesion function was introduced in this work and compared with two previous regularity functions, the geometric compactness and the non-connectivity functions. The cohesion function is based on the graph topology to penalize the presence of under-populated disconnection nodes in candidate clusters, the *disconnection nodes cohesion function.* A disconnection node is defined as a region within a cluster, such that its removal disconnects the cluster. By applying this function, the most geographically meaningful clusters are sifted through the immense set of possible irregularly shaped candidate cluster solutions. The disconnection nodes cohesion regularity function penalizes inconsistent clusters, but without forbidding the presence of the geographically interesting irregularly shaped ones. It penalizes irregularly shaped clusters selectively: the irregularity is allowed only to the extent that it does not impact the stability of the cluster, or its sensitivity to the removal of disconnection nodes.

We have proposed multi-objective scans to maximize two objectives: the spatial scan statistics and a chosen regularity function. All three regularity functions were used to build the corresponding three multi-objective scans. Additionally the compactness function and the disconnection nodes cohesion function were combined with Kulldorff's spatial scan statistic $LLR(z)$ in a fourth multi-objective optimization algorithm (MGD). Specifically, we maximize simultaneously two objective functions, $LLR(z).k(z)$ and $LLR(z).c(z)$. The advantage of the MGD approach is to provide a criterion for selecting clusters based simultaneously on their shape and presence of disconnection nodes. Those clusters which do not have good internal cohesion, due to the presence of disconnection nodes, and at the same time have very irregular shape, are automatically discarded by the algorithm. Only clusters which fare well in at least one criterion, regularity of shape or disconnection nodes cohesion, are thus selected as potential clusters candidates. A large degree of freedom is available in the choice of the "best" cluster, without compromising interesting and desirable features. Irregularity of shape is allowed, provided that it does not impact the structural stability of the cluster (measured here by the absence of disconnection nodes). Conversely, a few disconnection nodes may occur, if the cluster is not very irregularly shaped. In this fashion, realistic clusters are not discarded by the algorithm, and at the same time a large number of inadequate cluster candidates are eliminated.

The power of detection, sensitivity and positive predicted value of the multi-objective scans were compared with the corresponding evaluations for the three penalized likelihood scans, the non-penalty genetic algorithm scan and the usual circular scan. Our simulations suggest that the four multi-objective scans have better performance than the mono-objective scans. None of the multi-objective scans have shown better performance when compared to the other multi-objective scans: the power of detection of the cohesion scan and the MGD scan were higher for most situations; both geometric and non-connectivity scans shown better PPV performance; the cohesion scan presented generally better sensitivity. The non-connectivity scan is significantly faster.

We applied the statistical methodology of the attainment function to extend way the meaning of the p-value to the bi-objective space in a natural, while preserving the dependence between points within the same Pareto set, for all Pareto-sets obtained by the Monte Carlo simulation. This approach gives a more robust definition clusters's significance for multi-objective scans.

The MGD scan distinguishes clearly those clusters which are worse both from the geometric and the topological viewpoint. The algorithm's power to detect spatial clusters, its sensitivity and positive predicted value were studied through numerical simulations. The power of detection was good for clusters satisfying at least one of the requirements of high geometric compactness or high weak link cohesion. Clusters which are simultaneously highly penalized on both requirements do not fare well using our novel method, as expected. From a geographic perspective, those clusters are undesirable, and it makes sense to penalize them, and at the same time allowing those clusters which are strong in at least one of these two objectives.

An application was presented for Chagas' disease incidence in puerperal women in Brazil. This study case is particularly difficult to analyze, due to the sparsity of cases and the presence of many regions with zero cases, which could potentially produce a large uncertainty in the delineation of clusters. However, the novel algorithms produced very interesting clusters solutions, with the algorithm using zero cases regions as optimal links between the high risk regions. Those solutions compared favorably with the circular clusters or the non-penalty clusters solutions.

We also propose a more robust definition of spatial cluster using these concepts: a *geographically stable spatial cluster* satisfies at least one of the requirements of high geometric compactness or high disconnection nodes cohesion. The *most likely geographically stable spatial cluster* is the most significant cluster of the Pareto-set obtained through the compactness × disconnection nodes cohesion a multi-objective algorithm. We expect that this work should contribute to the investigation of what are the desirable characteristics of geographically sound spatial clusters, and provide adequate quantitative tools to detect them and assess their significance.

Bibliography

[1] Aamodt, G., Samuelsen, S.O. and Skrondal, A. (2006). A simulation study of three methods for detecting disease clusters, *International Journal of Health Geographics*, **5**, 15.

[2] Abrams, A.M., Kulldorff, M. and Kleinman, K. (2006). Empirical / Asymptotic P-Values for Monte Carlo-Based Hypothesis Testing: an Application to Cluster Detection Using the Scan Statistic, *Advances in Disease Surveillance*, **1**, 1.

[3] 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.*

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

[5] 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.

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

[7] Besag , J. and Newell, J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society*, **A154**, 143–155.

[8] 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.

[9] Beeker, E., Bauer, D.W. and Mohtashemi, M. (2007). Benchmark Data Generation from Discrete Event Contact Network Models, *Advances in Disease Surveillance*, **4**, 235.

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

[11] Cami, A., Wallstrom, G.L. and Hogan, W.R. (2007). Effect of Work-related Mobility in the Simulation of Aerosol Anthrax Releases with BARD, *Advances in Disease Surveillance*, **4**, 239.

[12] Cançado, A.L.F. (2009). Doctor thesis: *Detecção de Clusters Espaciais Através de Otimização Multiobjetivo*, *Department of Electric Engineering - UFMG, Brasil.*

[13] Carrano, E.G., Soares, L.A.E., Takahashi, R.H.C., Saldanha, R.R. and Neto, O.M. (2006). Electric distribution network multiobjective design using a problem-specific genetic algorithm, *IEEE Transactions on Power Delivery*, **21(2)**: 995–1005.

[14] Chankong, V. and Haimes, Y.Y. (1983). Multiobjective Decision Making:Theory and Methodology. *North-Holland.*

[15] Conley, J., Gahegan, M. and Macgill, J. (2005). A Genetic Approach

to Detecting Clusters in Point Data Sets, *Geographical Analysis*, **37**, 286–314.

[16] Cook, A.J., Gold, D.R. and Li, Y. (2007). Spatial Cluster Detection for Censored Outcome Data, *Biometrics*, **63**, 540–549.

[17] Costa, M.A., Kulldorff, M. and Assunção, R.M. (2007). A Space Time Permutation Scan Statistic with Irregular Shape for Disease Outbreak Detection, *Advances in Disease Surveillance*, **4**, 243.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Duarte, A.R., Duczmal, L., Ferreira, S.J. and Cançado, A.L.F. (2009). Internal cohesion and geometric shape of spatial clusters, 1–19 (to appear in *Environmental and Ecological Statistics*).

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Duczmal, L., Ferreira, S.J., Duarte, A.R., Soares, M.V., Gontijo, E.D, Cançado, A.L.F., and Takahashi, R.H.C. (2008). Geographically Meaningful Cluster Scanning Through Weak Link Correction, *Advances in Disease Surveillance*, **5**, 28.

[30] Duczmal, L., Cançado, A.L.F., Ferreira, S.J., Duarte, A.R., Fonseca, C.M., and Gontijo, E.D. (2009). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters, 1–22 (submitted).

[31] 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** (to appear).

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

[33] Elliott, P., Martuzzi, M. and Shaddick, G. (1995). Spatial statistical methods in environmental epidemiology: a critique, *Statistical Methods in Medical Research*, **4**, 137–159.

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

[35] 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.

[36] 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.

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

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

[39] Haiman, G. and Preda, C. (2002). A New Method for Estimating the Distribution of Scan Statistics for a Two-Dimensional Poisson Process, *Methodology And Computing In Applied Probability*, **4(4)**, 393–407.

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

[41] 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.

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

[43] Jacquez, G.M., Kaufmann, A., Meliker, J., Goovaerts, P., AvRuskin, G. and Nriagu, J. (2005). Global, local and focused geographic clustering for case-control data with residential histories, *Environmental Health: A Global Access Science Source*, **4**, 4.

[44] Jacquez, G.M., Meliker, J., AvRuskin, G., Goovaerts, P., Kaufmann, A., Wilson, M. and Nriagu, J. (2006). Case-control geographic clustering for residential histories accounting for risk factors and covariates, *International Journal of Health Geographics*, **5**, 32.

[45] 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).

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

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

[48] 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.

[49] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society*, **164(1)**, 61–72.

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

[51] Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R.M. and Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection, *PLoS Medicine*, **2(3)**, 216–224.

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

[53] 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.

[54] 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.

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

[56] Lin, G. and Zhang, T. (2007). Loglinear Residual Tests of Moran's *I* Autocorrelation and their Applications to Kentucky Breast Cancer Data, *Geographical Analysis*, **39**, 293–310.

[57] Meliker, J.R. and Jacquez, G.M. (2007). Space-time clustering of case-control data with residential histories: insights into empirical induction periods, *Journal of Stochastic Environmental Research & Risk Assessment*, **21**, 625–634.

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

[59] 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.

[60] Moran, P.A.P. (1948). The interpretation of statistical maps, *Journal of the Royal Statistical Society*, **10**, 243–251.

[61] Moran, P.A.P. (1950). A test for the serial independence of residuals, *Biometrika*, **37**, 178–181.

[62] 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.

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

[64] Neill, D.B. and Moore, A.W. (2003). A Fast Multi-Resolution Method for Detection of Significant Spatial Overdensities, *Carnegie Mellon CSD Technical Report*.

[65] Neill, D.B. and Moore, A.W. (2004). A Fast Multi-Resolution Method for Detection of Significant Spatial Disease Clusters, *Advances in Neural Information Processing Systems*, **16**, 651–658.

[66] 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.

[67] Neill, D.B. and Moore, A.W. (2006). Methods for detecting spatial and spatio-temporal clusters, *Handbook of Biosurveillance*, 243–254.

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

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

[70] Neill, D.B. and Makatchev, M. (2007). Incorporating Learning into Disease Surveillance Systems, *Advances in Disease Surveillance*, **4**, 107.

[71] Ozdenerol, E., Williams, B.L., Kang, S.Y. and Magsumbol, M.S. (2005). Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters, *International Journal of Health Geographics*, **4**, 19.

[72] Ozonoff, A., Jeffery, C., Manjourides, J., White, L.F. and Pagano, M.

(2007). Effect of spatial resolution on cluster detection: a simulation study, *International Journal of Health Geographics*, **6**, 52.

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

[74] 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.

[75] Rosychuk, R.J. (2006). Identifying geographic areas with high disease rates: when do confidence intervals for rates, *International Journal of Health Geographics*, **5**, 46.

[76] Rosychuk, R.J., Huston, C. and Prasad, N.G.N. (2006). Spatial Event Cluster Detection Using a Compound Poisson Distribution, *Biometrics*, **62**, 465–470.

[77] 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.*

[78] Song, C. and Kulldorff, M. (2003). Power evaluation of disease clustering tests, *International Journal of Health Geographics*, **2**, 9.

[79] Takahashi, K. and Tango, T. (2007). A Scan Statistic based on Anscombe's Variance Stabilization Transformation, *Advances in Disease Surveillance*, **4**, 116.

[80] Takahashi, K., Kulldorff, M., Tango, T. and Yie, K. (2007). A Flexible Space-Time Scan Statistic for Disease Outbreak Detection and Monitoring, *Advances in Disease Surveillance* , **2**, 70.

[81] 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.

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

[83] Tango, T. (2007). A Spatial Scan Statistic Scanning Only the Regions with Elevated Risk, *Advances in Disease Surveillance*, **4**, 117.

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

[85] 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**.

[86] Waller, L.A. and Jacquez, G.M. (2000). Disease models implicit in statistical tests of disease clustering, *Epidemiology*, **6**, 584–590.

[87] 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.

[88] Womble, W.H. (1951). Differential systematics, *Science*, **114**, 315–322.

[89] Zhang, T. and Lin, G. (2007). A decomposition of Moran's I for clustering detection, *Computational Statistics & Data Analysis*, **51**, 6123–6137.

[90] Zimmerman, D.L. and Pavlik, C. (2008). Quantifying the Effects of Mask Metadata Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data, *Geographical Analysis*, **40**, 52–76.

List of Figures

List of Tables