TESE SE DOUTORADO Nº 126

THE DETECTION OF SPATIAL CLUSTERS: GRAPH AND DYNAMIC PROGRAMMING METHODS

Gladston Juliano Prates Moreira

DATA DA DEFESA: 01/07/2011

Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica

The Detection of Spatial Clusters: Graph and Dynamic Programming Based Methods

Gladston Juliano Prates Moreira

Thesis presented to the Graduate Program in Electrical Engineer of the Federal University of Minas Gerais in final fulfillment of the requirements for the degree of Doctor in Electrical Engineer.

Advisor: Prof. Luiz Henrique Duczmal

Co-advisor: Prof. Ricardo H. C. Takahashi

During the development of this work the author received financial support from CAPES / FAPEMIG

Belo Horizonte, July 01, 2011

Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica

A Detecção de *Clusters* Espaciais: Métodos Baseados em Grafos e Programação Dinâmica.

Gladston Juliano Prates Moreira

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito final para obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Prof. Luiz Henrique Duczmal

Co-orientador: Prof. Ricardo H. C. Takahashi

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES/FAPEMIG

Belo Horizonte, 01 de Julho de 2011

"The Detection Of Spatial Clusters: Graph And Dynamic Programming Based Methods"

Gladston Juliano Prates Moreira

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Engenharia Elétrica.

Aprovada em 01 de julho de 2011.

Por:

Prof. Dr. Luiz Henrique Duczmal
Estatística (UFMG) - Orientador

Prof. Dr. Ricardo Hiroshi Caldeira Takahashi
DMAT (UFMG) - Co-Orientador

Prof. Dr. Luís Paquete
Universidade de Coimbra

Prof. Dr. Rodney Rezende Saldanha
DEE (UFMG)

Prof. Dr. André Luiz Fernandes Cançado
Estatística (UnB)

Profa. Dra. Elizabeth Fialho Wanner
Computação (CEFET/MG)

Profa. Dra. Denise Burgarelli Duczmal
DMAT (UFMG)

Dedico esse trabalho à *Raquel*, *Rossana* e ao meu filho *Gael*.

I dedicate this work to *Raquel*, *Rossana* and my son *Gael*.

Acknowledgments

I would like to thank first my whole family, specially my parents for their support, comprehension and for everything they have done for me.

I am also very grateful to Jussara, David, Eloy, Anderson and Emerson for the great friendship and companionship.

Thanks to Professor Luiz Henrique Duczmal and Professor Ricardo Takahashi which were essential in this work. Thanks for the shared knowledge, the excellent orientation, the credit and the opportunity given to me.

I am grateful to Flávia Magalhães, the community of health agents of the Family Health Program, and the Secretary of Health and Epidemiological Surveillance in Lassance City.

Thanks also to Professor Luís Paquete the great opportunity that he offered me, the excellent orientation and assistance during the Phd internship I worked at the University of Coimbra. I also thank Professor Carlos Fonseca by the doors that were opened.

Many thanks to the members of my thesis committee, professors Luís Paquete, Rodney Rezende Saldanha, André Luiz Fernandes Cançado, Elizabeth Fialho Wanner and Denise Burgarelli Duczmal for their valuable suggestions.

I am grateful to colleagues in the group's statistics department and engineering school for help and companionship.

I would like to thank CAPES and FAPEMIG for the financial support and everyone that contributed to this work.

Agradecimentos

Gostaria de agradecer primeiramente toda minha família, especialmente ao meus pais, pelo apoio e por tudo que fizeram por mim.

Sou muito grato também à Jussara, David, Eloy, Anderson e Emerson pela grande amizade e companherismo.

Agradeço ao Professor Luiz Henrique Duczmal e Professor Ricardo Takahashi os quais foram essenciais na realização deste trabalho. Agradeço pelo conhecimento compartilhado, pela excelente orientação, pela confiança e pela oportundade que me foi dada.

Agradeço a Flávia Magalhães, a comunidade dos agentes de saúde do Programa Saúde da Família e a secretaria de Saúde e Vigilância Epidemiológica da cidade de Lassance.

Agradeço também ao Professor Luís Paquete, pela grande oportunidade que ele me proporcionou, pela excelente orientação e acolhimento durante o doutorado sanduíche realizado na Universidade de Coimbra. Agradeço também ao Professor Carlos Fonseca pelas portas que foram abertas.

Gostaria tambéem de agradecer aos membros da minha banca, os professores Luís Paquete, Rodney Rezende Saldanha, André Luiz Fernandes Cançado, Elizabeth Fialho Wanner e Denise Burgarelli Duczmal pelas valiosas sugestões

Sou grato aos colegas do grupo do departamento da estatística e da escola de engenharia pelo grande companheirismo.

Meus agradecimentos à CAPES e FAPEMIG pelo suporte financeiro e a todos aqueles que de alguma forma contribuíram para a realizaçãao deste trabalho.

Abstract

This thesis addresses the spatial and space-time cluster detection problem. Two algorithms to solve the typical problem for spatial data sets are proposed.

A fast method for the detection and inference of point data set spatial and space-time disease clusters is presented, the Voronoi Based Scan (VBScan). A Voronoi diagram is built for points representing population individuals (cases and controls). The number of Voronoi cells boundaries intercepted by the line segment joining two cases points defines the Voronoi distance between those points. This distance is used to approximate the density of the heterogeneous population and build the Voronoi distance Minimum Spanning Tree (VMST) linking the cases. The successive removal of edges from the VMST generates sub-trees which are the potential clusters. Finally, those clusters are evaluated through the scan statistic. Monte Carlo replications of the original data are used to evaluate the significance of the clusters. The ability to promptly detect space-time clusters of disease outbreaks, when the number of individuals is large, was shown to be feasible, due to the reduced computational load of VBScan. Numerical simulations showed that VBScan has higher power of detection, sensitivity and positive predicted value than the Elliptic PST. Furthermore, an application for dengue fever in a small Brazilian city is presented.

In a second approach, the typical spatial cluster detection problem is reformulated as a bi-objective combinatorial optimization problem. We propose an exact algorithm based on dynamic programming, Geographical Dynamic Scan, which empirically was able to solve instances up to large size within a reasonable computational time. We show that the set of non-

dominated solutions of the problem, computed efficiently, contains the solution that maximizes the Kulldorff's Spatial Scan Statistic. The method allows arbitrary shaped clusters, which can be a collection of disconnected or connected areas, taking into account a geometric constraint. Note that this is not a serious disadvantage, provided that there is not a huge gap between its component areas. We present an empirical comparison of detection and spatial accuracy between our algorithm and the classical Kulldorff's Circular Scan, using the data set of Chagas disease cases in puerperal women in Minas Gerais state, Brazil.

Keywords: spatial scan statistic, spatial cluster, space-time cluster, voronoi diagram, minimum spanning tree, combinatorial optimization problem, dynamic programming.

Resumo

Esta tese aborda o problema de detecção de clusters espaciais e espaços-temporais. Dois algoritmos para resolver o típico problema de conjuntos de dados com processos espaciais são propostos.

Um método eficiente para a detecção e inferência de clusters de doenças espaciais e espaços-temporais de dados pontuais é apresentado, o Voronoi Based Scan (VBScan). Um diagrama de Voronoi é construído para os pontos que representam indivíduos da população (casos e controles). O número de células de Voronoi interceptadas pelo segmento de linha que une de dois pontos que representam dois casos define a distância de Voronoi entre esses pontos. Esta distância é usada para aproximar a densidade da população heterogênea e construir a árvore geradora mínima baseada na distância de Voronoi (VMST) ligando os casos. A remoção sucessiva de arestas da VMST gera sub-árvores que são os clusters candidatos potenciais. Finalmente, os clusters são avaliados através da estatística scan de Kulldorff. Simulações de Monte Carlo dos dados originais são usados para avaliar a significância dos clusters. A capacidade de detectar rapidamente clusters de surtos da doença, quando o número de indivíduos é grande, mostrou-se viável, devido à redução da carga computacional obtida com o VBScan. As simulações numéricas mostraram que o VBScan tem maior poder de detecção, sensibilidade e valor preditivo positivo do que o scan elíptico. Além disso, uma aplicação de casos e controles georeferenciados de dengue em uma cidade do Brasil é apresentado.

Numa segunda abordagem, o problema típico de detecção de clusters espaciais é reformulado como um problema bi-objetivo de otimização combinatória. Nós propomos um algoritmo exato baseado em programação dinâmica, Geographical Dynamic Scan, que empiricamente foi capaz de re-

solver os casos até de grande porte dentro de tempo computacional aceitável. Nós mostramos que o conjunto de soluções não dominadas do problema, encontradas eficientemente, contém a solução que maximiza a estatística scan de Kulldorf. O método permite clusters de formatos arbitrários, que podem ser uma coleção de regiões desconectadas ou conectadas, tendo em conta uma restrição geográfica. Note-se que esta não é uma séria desvantagem, desde que não haja um grande espaçamento entre as suas áreas. Apresentamos uma comparação empírica de detecção e precisão espacial entre o nosso algoritmo e o clássico Scan circular, utilizando dados de casos de doença de Chagas em mulheres parturientes no estado de Minas Gerais, Brasil.

Palavras-chave: estatística espacial scan, cluster espacial, cluster espaço-temporal, diagrama de voronoi, árvore geradora mínima, problema de otimização combinatória, programação dinâmica.

Resumo Estendido

Esta seção consiste em um resumo estendido sobre o trabalho desenvolvido nesta tese. Primeiramente, este texto introduz o problema abordado, a principal motivação para solucioná-lo e alguns dos principais métodos relacionados. Em seguida, uma breve descrição dos objetivos principais das metodologias desenvolvidas para resolver o problema. Finalmente, as conclusões são apresentadas.

Introdução

Testes estatísticos de vigilância de doença no espaço e no espaço-tempo, geralmente, procuram determinar se a incidência da doença em um subconjunto definido espacial e/ou temporalmente é incomum em relação à incidência na região de estudo como um todo. Assim, essa classe de métodos é projetada para detectar clusters de doença no espaço e no tempo, e adaptar sistemas de vigilância concebidos para a detecção de surtos. O desenvolvimento de métodos de detecção de clusters espaços-temporais, naturalmente, evoluiu a partir de métodos puramente espaciais. Podemos estratificar os métodos em três tipos de classe de testes estatísticos: testes para interação espaço-tempo, os métodos de soma cumulativa, e a estatística scan.

A estatística espacial scan de Kulldorff (Kulldorff, 1997) é atualmente o método mais usual para encontrar clusters (aglomerados) espaciais, espaço-temporais e temporais. Estudada em detalhe pela primeira vez por (Naus, 1965) é um método estatístico com muitas aplicações potenciais, com o objetivo de detectar um excesso de eventos locais. A estatística espacial scan supera o problema de testes múltiplos (comuns a muitos métodos locais de

análise espacial), tomando o cluster mais provável definido pela maximização da razão de verossimilhança. (Kulldorff & Nagarwalla, 1995) apresentam o clássico 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.

Em (Kulldorff, 2001), a estatística espacial scan é estendida para o espaço-tempo, de modo que cilindros são utilizadas para o formato dos potenciais candidatos a cluster. A base circular representa a área espacial e a altura do cilindro representa o período de tempo. Na análise prospectiva, cilindros candidatos são limitados àqueles que começam a qualquer momento durante o período de estudo e termina no período de tempo atual (ou seja, clusters vivos).

O problema típico de detecção de clusters espaciais

No enfoque desta tese, propomos dois métodos de detecção de clusters espaciais quando dois tipos de dados de processo espacial para um determinado fenômeno de interesse estão disponíveis. Suponha que tenhamos, por exemplo, um mapa dividido em regiões, cada uma delas com uma população conhecida e um número de casos observados. Assim, cada caso pode ser, por exemplo, um indivíduo infectado por uma certa doença. Neste mapa um *cluster* é um aglomerado de regiões geograficamente limitadas 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 geograficamente limitado de regiões do mapa. Denotaremos por Z o conjunto de todas as zonas. Por exemplo, uma janela circular sobre a área em estudo define uma zona formada pelas regiões cujos centróides estão dentro da janela, veja a Figura 1. Suponha agora que tenhamos um conjunto de dados pontuais

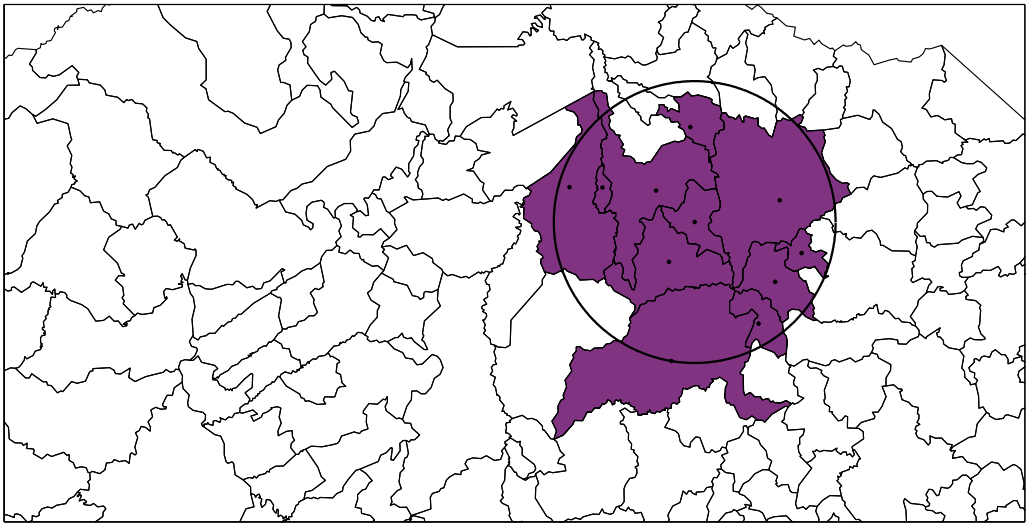

Figure 1: Uma possível zona obtida para uma dada janela circular.

de casos e controles, onde cada ponto no mapa representa indivíduos infectados (casos) e indivíduos com similares características mas não infectados (controles). Uma *zona* é qualquer subconjunto geograficamente limitado de indivíduos do mapa. Dentre os N indivíduos, n são casos e $N - n$ são controles. Para cada círculo de raio $r > 0$ centrado em cada indivíduo, uma zona é o conjunto de indivíduos dentro do círculo. Um exemplo é visto na Figura 2.

Seja $z \in Z$ uma zona, definindo $L(z)$ como a função de verossimilhança sob a hipótese alternativa de que exista uma zona z^* que é um cluster, e L_0 como a verossimilhança sob a hipótese nula de que não exista um cluster, foi mostrado em (Kulldorff, 1997) que o logaritmo da razão de verossimilhança, $K(\mathbf{z}) = \log\left(L(\mathbf{z})/L_0\right)$, é dado por

$$K(\mathbf{z}) = \begin{cases} C\log\left(\dfrac{N}{C}\right) + \mathbf{z_C}\log\left(\dfrac{\mathbf{z_C}}{\mathbf{z_N}}\right) + (C - \mathbf{z_C})\log\left(\dfrac{C - \mathbf{z_C}}{N - \mathbf{z_N}}\right) & \text{if } \dfrac{\mathbf{z_C}}{\mathbf{z_N}} > \dfrac{C - \mathbf{z_C}}{N - \mathbf{z_N}} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

assumindo que o número de casos na zona $\mathbf{z}$, $\mathbf{z_C}$, segue uma distribuição de Poisson com média proporcional à sua população $\mathbf{z_N}$. A função K é maximizada sobre todas as zonas em Z, identificando a zona que constitui

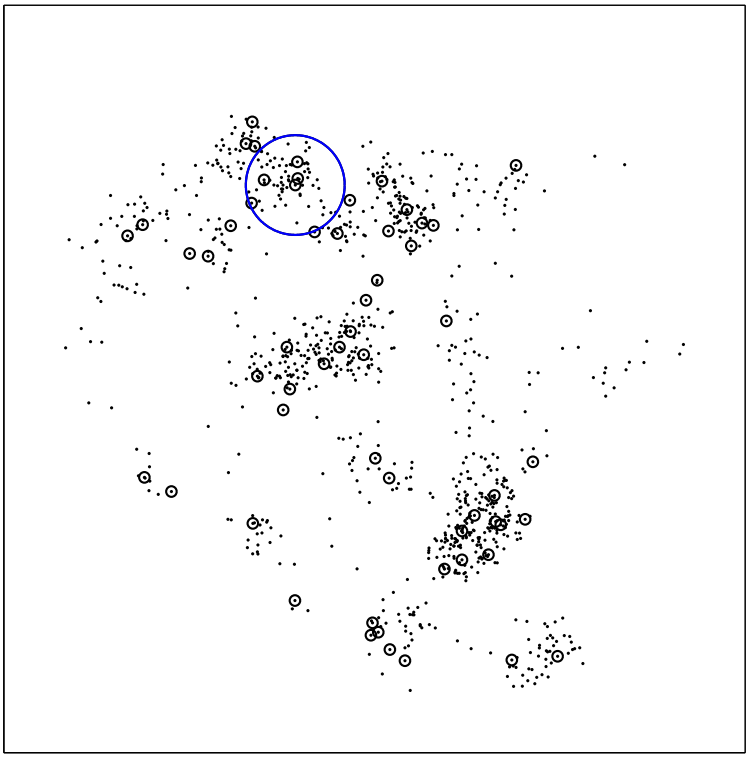

Figure 2: Uma possível zona obtida para uma dada janela circular.

o cluster mais verossímil. Então, temos a estatística de teste dada por $T = \max_z K(z)$.

A busca por soluções eficientes seria feita então dentro do conjunto Z. O fato é que seria computacionalmente invíavel testar todas as zonas possíveis. Para contornar esse problema, os algoritmos para detecção de clusters espaciais fazem uso de duas técnicas:

- Redução do conjunto das soluções canditadas Z para outro conjunto Z' das zonas promissoras ou que permita uma busca exaustiva.

- Utilização de métodos estocásticos de otimização.

Em ambas as técnicas, geralmente, os métodos só garantem uma boa aproximação para a solução ótima global do problema. Outra restrição dos métodos de detecção de clusters está relacionada com o formato dos clusters encontrados. Muitos algoritmos 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. Outros apresentam clusters detectados com formatos fixos, tipicamente circulares.

Neste sentido, propomos nesta tese dois algoritmos distintos de detecção de clusters. Um que aborda conceitos de teoria de grafos, com o propósito de obter um conjunto Z' de potenciais candidatos a cluster. O outro implementa conceito de programação dinâmica com o objetivo final de reduzir o conjunto das soluções candidatas ao cluster, encontrando uma solução ótima global para o problema.

Conclusões

Este trabalho apresenta dois métodos de detecção de clusters espaciais. Um primeiro método direcionado a detectar clusters espaciais e espaço-temporais para dados de processos pontuais de casos e controles. O segundo direcionado a ambos os tipos de dados. Os métodos discutidos são eficientes na melhoria das medidas de avaliação utilizadas em comparação com métodos clássicos.

Principais contribuições

- Proposição de um algoritmo de detecção de cluster espaciais e espaço-temporais, *Voronoi Based Scan*, quando disponíveis dados de processos pontuais de casos e controle;

- Proposição de um algoritmo de cluster espaciais, *Geographical Dynamic Scan*, tanto para dados de área quanto para dados pontuais de casos e controle;

- Elaboração, disponibilização e análise de casos e controles geo-referenciados de dengue na cidade de Lassance em Minas Gerais, em colaboração com o Programa de Saúde da Família.

Esboço da tese

Esta tese está organizada em 6 capítulos. O capítulo 1 apresenta os conceitos gerais do método de detecção de clusters espaciais, a estatística scan espacial, tipos de dados aplicáveis, inferência. Traz ainda uma breve descrição do método clássico para análise prospectiva de clusters espaço-temporais e uma breve revisão da literatura.

O capítulo 2 apresenta nosso método proposto para detecção de clusters espaciais e espaço-temporais quando um conjunto de dados de processos pontuais do tipo caso-controle é avaliado. Ele utiliza conceitos de teoria de grafos e otimização geométrica para caracterizar as soluções candidatas a cluster. Já no capítulo 3 implementamos um outro método de detecção de clusters espaciais usando um algoritmo de programação dinâmica que encontra eficientemente as soluções candidatas. Neste último, tanto um conjunto de dados agregadoa ou pontuais podem ser avaliado.

No capítulo 4 apresentamos análises numéricas que mostram o desempenho dos métodos aqui propostos na detecção de clusters. Estudos com dados artificiais e dados reais encontrados na literatura foram realizados. No capítulo 5 uma aplicação de casos e controles georeferenciados de dengue na cidade de Lassance em Minas Gerais é apresentada.

Finalmente, no capítulo 6 apresentamos nossas conclusões e perspecitvas de trabalhos futuros.

Contents

List of Acronyms

VBScan - Voronoi Based Scan

VMST - Voronoi Minimum Spanning Tree

Elliptic PST - Elliptic Prospective Space Time

MST - Minimum Spanning Tree

GDScan - Geographical Dynamic Scan

ND - Non-dominated operator

PPV - Positive Predicted Value

FHP - Family Health Program

List of Figures

List of Tables

Chapter 1

The spatial scan statistics

Spatial cluster detection methods are statistical tests which generally seek to determine whether a phenomenon of interest in a spatially defined subset is unusual compared to the incidence in the study region as a whole.

In this chapter we review the spatial scan statistics (Kulldorff, 1997). The first part provides a list of the types of data often used in typical spatial cluster detection problem.

1.1 Types of Data

1.1.1 Point Data

Locations of spatial entities (e.g., disease cases) are often represented as a point in two-dimensional map space, see Figure 1.1, such data are called *point data, point process data, event data.* Each record in this data must have its positional information represented by that x- and y-coordinates, and may also contain additional attributes (for example, age, gender, ...).

1.1.2 Case-Control Data

We will be interested in cluster detection and geographic surveillance when data are in the form of point locations for cases and controls. Here,

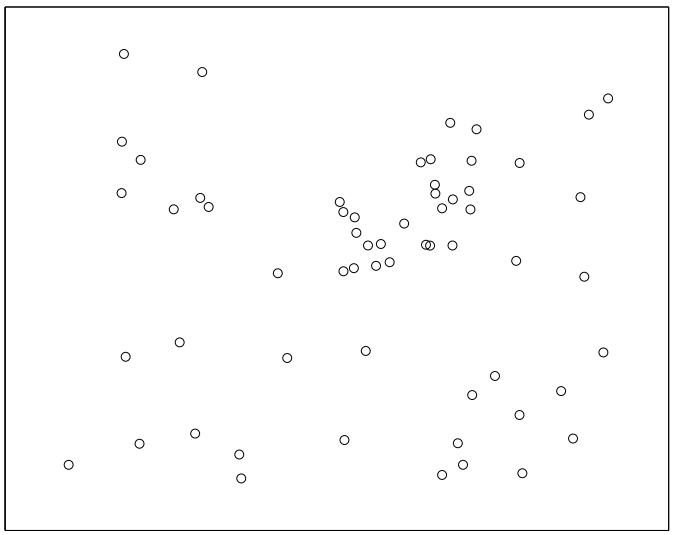

Figure 1.1: Spatial distribution of the observed cases in arbitrary data.

cases refer to individuals with a particular disease of interest, and controls refer to individuals with similar characteristics as cases but do not have the disease. When controls can be seen as a representative subset of population without the disease, comparison of the spatial distribution of the cases that of controls helps us identify spatial patterns in the cases distribution that are beyond what is merely reflective of the spatial distribution of the population. An example of case-control artificial data is shown in the Figure 1.2.

1.1.3 Aggregated Data

With aggregated data (or areal data), a study region is divided into a set of non-overlapping zones (such as counties, ZIP code zones ...), and each zone has associated attribute values such as the number of disease cases and population. This type of data may also contain other physical and socioeconomic attributes associated with each zone. The dataset shown in Figure 1.3. is a typical example of aggregated data. This is perhaps the most commonly available form of spatial data because exact locations of disease cases are often not publicly releasable.

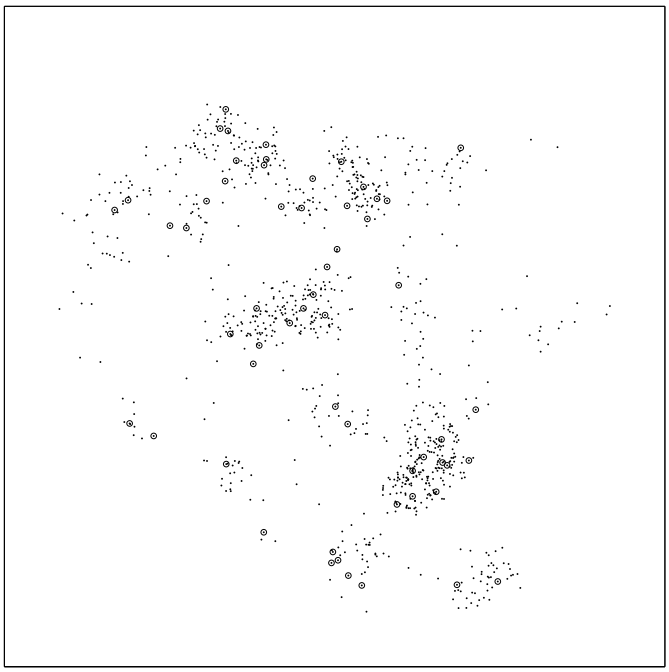

Figure 1.2: Spatial distribution of the observed cases (circles) and controls (dots) in Lancashire-UK data.

1.1.4 Space Time Data

Although less commonly available, we will also be interested in data which have time subscripts in addition to positional information. For example, a point dataset that represents each patient as a point on his or her residential address with associated date of diagnosis or an aggregated dataset that contains the number of cases and population for each zone over multiple years belong to this category of data. Such data offer us opportunities to examine not only spatial patterns in the data distribution but also spatial-temporal patterns and temporal changes in spatial patterns. When investigating a disease outbreak, for example, one's objective is not merely to detect spatial clusters of disease cases but also to identify how the size, shape, or location of the clusters is changing over time.

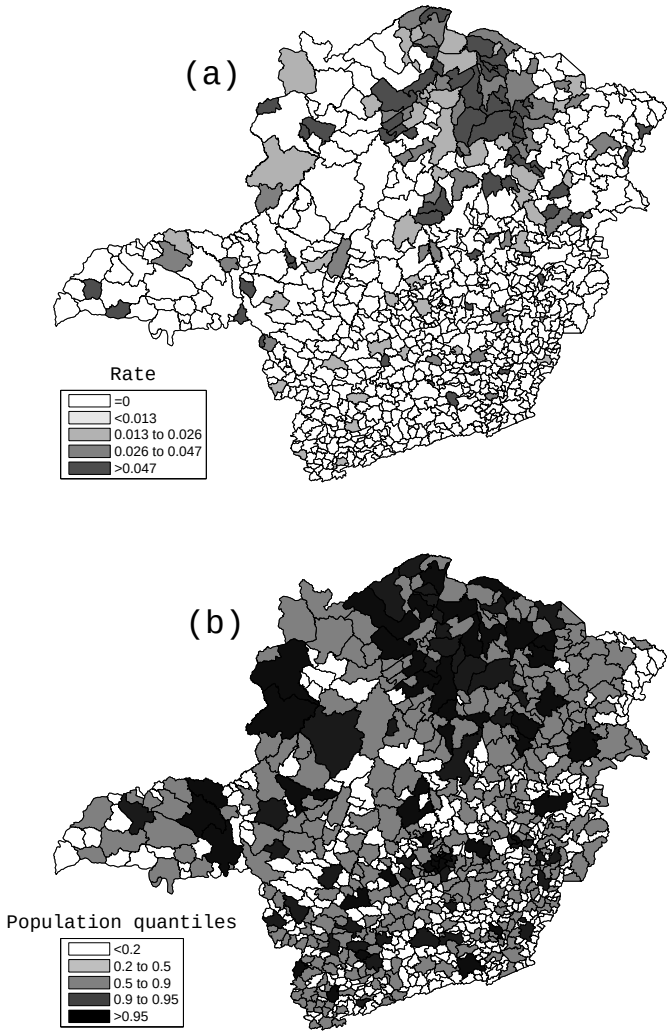

Figure 1.3: Mapping spatial variations of Chagas disease in the State of Minas Gerais - Brazil by county during 2006: (a) disease rates map; (b) Population at risk map.

1.2 Kulldorff's Spatial Scan Statistic

Consider a spatial dataset with M locations, for example, a disease map divided into M regions, with total population N and total number of cases C. A zone $\mathbf{z}$ is any subset of regions of the map. The null hypothesis states that 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 $\mathbf{z}$, the number of observed cases is $\mathbf{z_C}$ and the expected number of cases under null hypothesis is $\mathbf{z}_\mu = \mathbf{z_N}(C/N)$, where $\mathbf{z_N}$ is the population in the zone $\mathbf{z}$. Defining $L(\mathbf{z})$ as the likelihood function under the alternative hypothesis and L_0 as the likelihood function under the null hypothesis, it can be shown (Kulldorff, 1997) that the logarithm of the likelihood ratio, $K(\mathbf{z}) = \log\left(L(\mathbf{z})/L_0\right)$, for the Poisson model is given by:

$$K(\mathbf{z}) = \begin{cases} C\log\left(\frac{N}{C}\right) + \mathbf{z_C}\log\left(\frac{\mathbf{z_C}}{\mathbf{z_N}}\right) + (C-\mathbf{z_C})\log\left(\frac{C-\mathbf{z_C}}{N-\mathbf{z_N}}\right) & \text{if } \frac{\mathbf{z_C}}{\mathbf{z_N}} > \frac{C-\mathbf{z_C}}{N-\mathbf{z_N}} \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

The function K is maximized over the chosen set Z of potential zones $\mathbf{z}$, identifying the zone which constitutes the most likely cluster. Hence we have the test statistic, given by $T = \max_z K(z)$.

Given the definition above, it follows that $K(\mathbf{z}) = K(\mathbf{z_C}, \mathbf{z_N})$ is a function of variables $\mathbf{z_C}$ and $\mathbf{z_N}$. Assuming that $\mathbf{z_C}$ and $\mathbf{z_N}$ take positive values, it is trivial to note that $K(\mathbf{z_C}, \mathbf{z_N})$ satisfies the following property:

Property 1 *Let* $\mathbf{z_C}/\mathbf{z_N} > (C-\mathbf{z_C})/(N-\mathbf{z_N})$. *The function* $K(\mathbf{z_C}, \mathbf{z_N})$ *is strictly increasing in the variable* $\mathbf{z_C}$ *and strictly decreasing in the variable* $\mathbf{z_N}$.

1.3 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 than the critical value. Since, in principle, that probability distribution is unknown, we use Monte Carlo simulations (Dwass, 1957) 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 the fixed total number of cases C 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 become a case in any region is proportional 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, T, 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).

1.4 Prospective Space-Time Scan

The Prospective Space-Time Scan (Kulldorff, 2001) considers all cylindrical clusters in the space-time domain. All the possible circular windows in the space domain are taken as the bases of the cylinders to be considered. The study period is given by the time interval $[Y_1, Y_2]$. The likelihood for the observed data set is obtained as the maximum over all cylinders in the time interval $[s, t]$ reaching the end of the study period, with $Y_1 \leq s \leq t = Y_2$. For the random data sets generated under null hypothesis, the likelihood is maximized over all cylinders for which $Y_1 \leq s \leq t \leq Y_2$ and $Y_m \leq t$, where Y_m is the time instant in which the time periodic surveillance began, in order to adjust for the multiple analysis. See (Kulldorff, 2001) for details. SaTScan software implements the Prospective Space-Time Scan for both area and point data sets. In order to establish some comparisons for the evaluation of the proposed method, in this paper we have implemented a version of the Prospective Space-Time Scan for point data sets using elliptic cylinders instead of circular zones (Kulldorff *et al.*, 2006).

1.5 State-of-the-art

The spatial scan statistic (Kulldorff, 1997) constitutes the main statistic used for cluster detection, being employed, for instance, by the software packages SaTScan (Kulldorff, 1999) to detect static circularly shaped disease clusters (Kulldorff & Nagarwalla, 1995). Recently, several attempts have been developed in order to relax the assumption of cluster circular shape. (Sahajpal *et al.*, 2004) used a genetic algorithm to find clusters shaped as intersections of circles of different sizes and centers. The SaTScan approach has been extended to the case of elliptic shaped clusters (Kulldorff *et al.*, 2006), in this way allowing the detection of elongated clusters. Other methods have also been proposed to detect connected clusters of irregular shape (Duczmal & Assunção, 2004; Patil & Taillie, 2004; Tango & Takahashi, 2005; Duczmal *et al.*, 2006, 2008; Neill, 2008, 2010). The Static Minimum Spanning Tree (SMST) proposed by (Assunção *et al.*, 2006) used a greedy algorithm to aggregate regions. The Flexibly Shaped (FS) spatial scan statistic (Tango & Takahashi, 2005) made an exhaustive search of all possible first-order connected clusters contained within a set encompassing the nearest k neighbors of a given region.

A key point for the construction of such methods for detection of irregularly shaped clusters is that, as the geometrical shape receives more degrees of freedom, some correction should be employed in order to compensate the increased flexibility, so avoiding the increase of false-positive errors (Duczmal *et al.*, 2006, 2007). This fact has been recognized since the early study of elliptically shaped clusters (Kulldorff *et al.*, 2006). These corrections were also treated in a multi-objective framework (Duczmal *et al.*, 2008; Duarte *et al.*, 2010; Cançado *et al.*, 2010). (Yiannakoulias *et al.*, 2007) proposed a topological penalty.

Neill's Fast Subset Scan (Neill, 2008) presented a significant advance in spatial methods for aggregated area maps, finding exactly the optimal irregularly spatial clusters in linear computing time. The clusters found may sometimes be disconnected, but this is not a serious disadvantage, provided that there is not a huge gap between its areas. A way to control the pres-

ence of those potential gaps is to limit the number of component areas of the cluster, e.g. allowing only clusters which are subsets of a circular zone of moderate maximum size.

Recently, methods have been proposed when point process data are available. (Conley *et al.*, 2005) proposed a genetic algorithm to explore a configuration space of multiple agglomerations of ellipses in point data set maps, implemented in the software PROCLUDE. (Wieland *et al.*, 2007) introduced a graph theoretical method for detecting arbitrarily shaped clusters based on the Euclidean minimum spanning tree of cartogram transformed case locations, which is quite effective, but the cartogram construction step of this algorithm is computationally expensive and complicated. (Demattei *et al.*, 2007) proposed a method based on the construction of a trajectory for multiple cluster detection in point data sets. (Cucala, 2009) proposed a method for identifying clusters in spatial point processes. It relies on a specific ordering of events and the definition of area spacings which have the same distribution as one-dimensional spacings. (Demattei & Cucala, 2011) introduce a spatio-temporal distance which allows the extension of the spatial cluster detection methods (Demattei *et al.*, 2007; Cucala, 2009) used for detecting spatio-temporal clusters.

Chapter 2

Voronoi based scans for point data sets

2.1 Motivation

Algorithms for the detection and inference of clusters are useful tools in etiological studies (Lawson *et al.*, 1999) and in the early warning of infectious disease outbreaks (Duczmal & Buckeridge, 2006; Kulldorff *et al.*, 2005, 2006, 2007; Neill, 2009). A *spatial cluster* is defined as a localized portion of the domain containing a higher than average proportion of cases over controls, whose appearance is unlikely under the assumption that cases are randomly distributed in the population. *Space-time clusters* are defined as unexpected concentrations of disease cases in a time series sequence of geographical maps, and could potentially indicate an outbreak or epidemic, due to environmental or biological causes.

The mechanism behind the enhancement of the power of cluster detection methods when arbitrary shapes are considered can be described as:

- If the shape of the possible cluster was known a priori, the most powerful method of detection would be to assume such a shape, and search for empirical clusters of that format. In this way, objects of other shapes

would be disregarded, and the statistics would be evaluated only over the legitimate cluster candidates.

- If the shape of possible clusters was not known a priori, considering a fixed shape would lead to two kinds of errors: either the true cluster would be included inside a greater cluster estimate of the assumed shape, or only a portion of the true cluster which coincides with the assumed shape would be considered. In both cases, the power of the method would be decreased due to such errors.

- An entropy-like argument is employed at this point: the relatively rare regular shapes that would represent a homogeneous formation of the cluster are considered better than the more numerous flexible irregular shapes, that represent a rather non-homogeneous cluster propagation.

- Within the multi-objective framework, there is no need to state precisely the relative weight of the different shapes. The suitable balancing of the minimization of shape flexibility and the maximization of the likelihood ratio of the existence of the cluster can be attained by a hypothesis test, which reveals a cluster estimate with minimal p-value.

These developments related to flexible cluster shapes have been mostly performed for the static case only. The first motivation of this work is the concepts from the graph theory, applied to evaluate the set of potential clusters, for detecting arbitrarily shaped clusters based on the minimum spanning tree representation.

For the space-time case, the Prospective Space-Time Scan (Kulldorff, 2001) considers all cylindrical clusters in the space-time domain as cluster candidates. A version of Space-Time Scan has been developed too for the case of the elliptical scan, also considering cylindrical clusters stated as projections of the ellipses along the time dimension (Kulldorff *et al.*, 2006). The second main motivation of this work is the observation that, although the elliptical spatial shape endows some flexibility to the scan procedure, allowing a high detection power in space coordinates, the cylinder shape assumed in order to extend such a spatial shape to time coordinates is too restrictive, leading

to inaccuracies in space-time cluster detection. This issue has been dealt in some references (Iyengar, 2005; Takahashi *et al.*, 2008; Demattei & Cucala, 2011). See (Robertson & Nelson, 2010) for a review of space-time cluster detection software.

Our proposed methodology builds different graphs for each considered time interval. In this way, the flexibility that is necessary for dealing with the variation of the disease spread along the time dimension is obtained in a direct way.

2.2 Definitions and Methods

The idea of employing a Minimum Spanning Tree (MST) in order to characterize clusters has been already studied by (Assunção *et al.*, 2006), in the context of area data sets. For dealing with point data sets, the application of the scan statistics requires a proper definition of disease case density related to each data point. As, clearly, a single sphere radius was not suitable for estimating the population density in all regions, due to the heterogeneity in the geographical distribution of population, a correction procedure was necessary. The procedure proposed by (Wieland *et al.*, 2007) performed a non-linear cartogram transformation of the map, leading to a new map with an approximately homogeneous control population distribution. It should be noticed that this procedure is highly computing intensive.

A much simpler procedure for the estimation of disease density is proposed in this work. The general idea is: a Voronoi diagram is depicted, defining regions associated to each individual point in the map (both for disease and non-disease cases). A new distance, called Voronoi distance, between two points, is defined as the number of Voronoi cell boundaries that must be crossed in order to establish a path between those points. A ball of radius R in this distance, centered in the point A, would consist of the set of points which can be reached from A with up to R Voronoi cells crossings. Therefore, the Voronoi distance can be used in order to define a variable metric of the original coordinates that exactly performs the correction that transforms a non-homogeneous population density map into a homogeneous

one. The computation of the Voronoi distance and all associated entities can be performed with efficient polynomial algorithms. Using the VMST, the computation of disease clusters in a fixed time coordinate can be performed very fast. In order to deal with space-time clusters, a simple procedure that connects the graphs of different time instants by the common nodes is employed. The program was written in Dev C language.

2.2.1 Minimum spanning tree representation

We will use a Minimum Spanning Tree to represent a set of event data point and determine the subsets which potentially constitute a cluster.

In order to characterize point data set clusters, the Voronoi distance is defined. The population at risk consists of N individuals in the space domain, divided into n disease *cases* and $N - n$ *controls*. Consider the set $P = \{(x_i, y_i) : i = 1, ..., N\} \subset \mathbb{R}^2$, indicating the geographic location of the cases and controls. For $i = 1, ..., N$ the *Voronoi cell* $v(i)$ consists of those points in $\mathbb{R}^2$ which are closer to (x_i, y_i) than to any other point in P. The *Voronoi diagram* is formed by the collection of cells $v(i), i = 1, ..., N$.

Definition 1 (Voronoi distance) *Let v_{ij} be the number of Voronoi cells intercepted by the line segment joining the points (x_i, y_i) and (x_j, y_j) (including the cells containing the points i and j). In this work we define the Voronoi distance between points i and j as $\delta(i, j) = v_{ij} - 1$. When the points i and j occupy neighboring Voronoi cells, $\delta(i, j) = 1$.*

A geometric routine is used to compute the number of intersections of the segment linking two cases i and j with the edges of the Voronoi cells. If that segment intercepts tangentially a Voronoi cell, a potential problem may occur in the computation of $\delta(i, j)$. However, this problem occurs only rarely, supposing that the point coordinates follow a random pattern.

As an attempt to identify subsets of such a set that are likely to constitute a cluster, the following heuristic is employed here, (Xu *et al.*, 2002; Wieland *et al.*, 2007): A nonempty subset S of D forms a candidate cluster if the smallest distance separating the sets S and $D - S$ is greater than the

maximum internal distance of S, where $D - S$ is the subset of D removing all points of S. Formally, this can written as:
Let $S = S_1 \cup S_2$ be any partition and ρ represent the distance between two points of D. If $S \subseteq D$ forms a cluster then

$$\arg \min_{d \in D - S_1} \{\min\{\rho(d, s) : s \in S_1\}\} \in S_2.$$

Hence, the potential cluster is a connected graph with tree structure, linking the disease cases in the space domain. Our algorithm builds a set of subtrees of the minimum spanning tree of the complete graph of cases, defining a small set of potential space clusters. For example, considering the data points of the Figure 2.1 the points of the same cluster are connected with each other by short edges while long edges link cluster together.

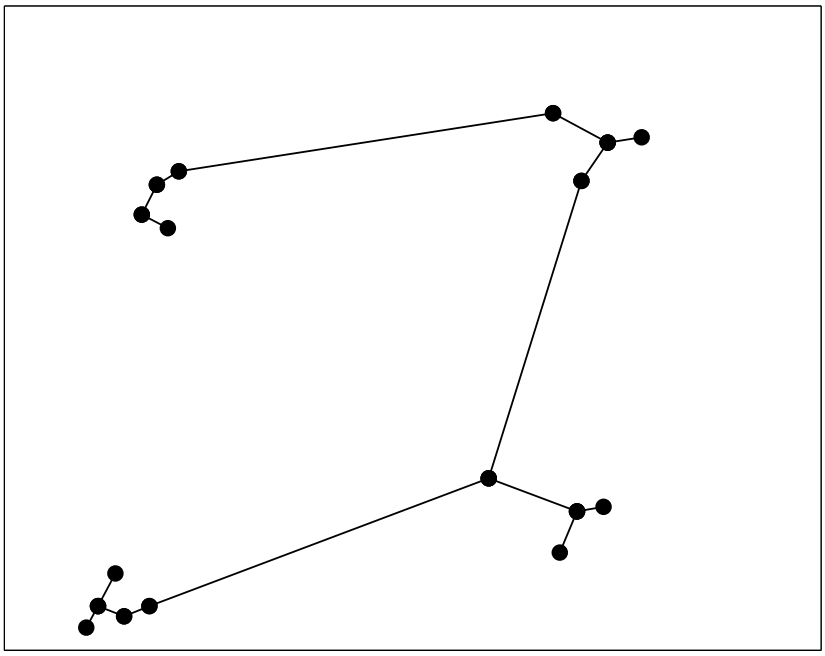

Figure 2.1: An minimum spanning tree connecting all the data points, using Euclidean distance.

2.3 Voronoi based spatial scan

Formally, let $D = \{c_i\}$ be the subset representing the disease cases where each $c_i = (x_i, y_i)$ indicates its geographic location. We define a weighted complete graph $G(D) = (V, E)$ with vertex set $V = \{c_i : c_i \in D\}$ and edge set $E = \{(c_i, c_j) : c_i, c_j \in D, i \neq j\}$. Each edge $(c_i, c_j) \in E$ has weight defined by the *Voronoi distance* $\delta(i, j)$.

A minimum spanning tree (MST) of a weighted complete graph $G(D)$ can be defined as a minimal set of edges of $G(D)$ that connect all vertices with minimum total distance. The Voronoi Minimum Spanning Tree (VMST) of the weighted graph $G(D)$ defined above is a spanning tree with the minimum total Voronoi distance. A set of discrete values characterizes the Voronoi distance. This would cause the emergence of multiple solutions very often. This effect is eliminated by ordering the edges with identical Voronoi distances according to the Euclidean distance. This procedure ensures the following lemma, which is an extension of the result proposed by (Wieland *et al.*, 2007):

Lemma 1 *Assume that the Euclidean distance between any two points belonging to the set P is different from any other distance between two points of the same set. Then the set of potential clusters are in one-to-one correspondence with connected components among all graphs T_w, with T_w defined as the graph derived from VMST by deleting all edges having weight greater than w.*

Proof: Define the order of descending weights w to the edges of VMST untied by Euclidean distance as discussed above. Hence, the proof follows the same way as performed in (Wieland *et al.*, 2007), replacing the Euclidean distance by Voronoi distance. □

The set of potential clusters may be quickly found from a VMST by using a greedy edge deletion procedure, improving and simplifying the strategy employed by the Density-Equalizing Euclidean MST method (Wieland *et al.*, 2007). The procedure is: After constructing the VMST of the set of case locations D, we iteratively remove the largest remaining edge, giving rise to two additional cluster candidates in each iteration. For a map with n cases, we obtain $2n - 1$ cluster candidates, including n unitary clusters.

Figure 2.2 shows the spatial distribution of 70 coordinates, with 10 observed cases (circles) and 60 non-cases (dots) in an artificial data set and the associated Voronoi minimum spanning tree. Figure 2.3 shows a simple visualization of the greedy edge deletion procedure for the example above. The successive steps of edge deletion are represented, with the new cluster candidates shown in each iteration.

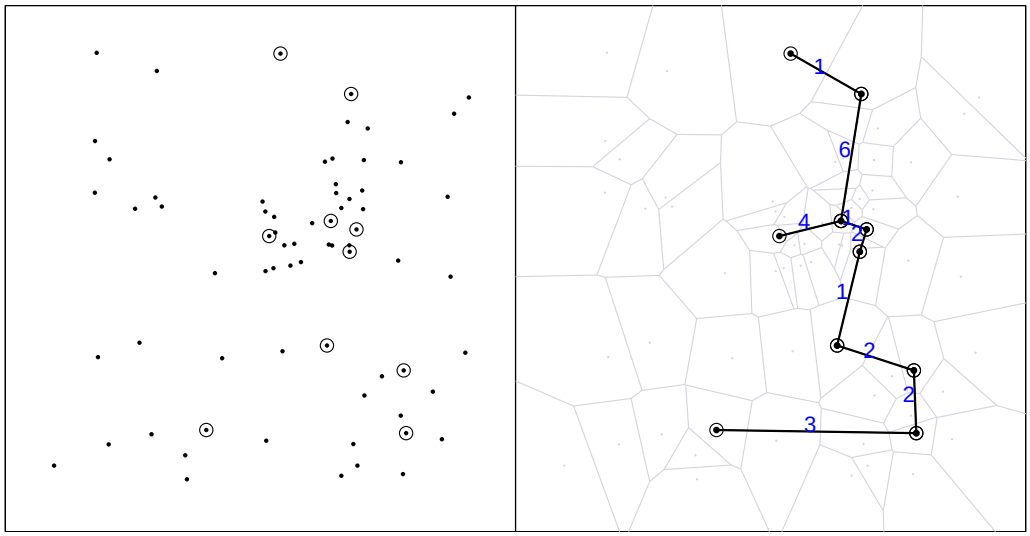

Figure 2.2: Left: spatial distribution of the 10 observed cases (circles) and 60 non-cases (dots). Right: corresponding Voronoi minimum spanning tree.

Given a case with geographic location $c_i = (x_i, y_i)$, consider the circle $\mathcal{C}(c_i, r)$ centered in the point (x_i, y_i), with radius r. If the local density around the point (x_i, y_i) is given by s individuals per unit area, then the expected number of individuals inside the circle $\mathcal{C}(c_i, r)$ is computed as $s\pi r^2$. When the radius r is expressed locally in units of the Voronoi distance as R, then the expected number of individuals inside $\mathcal{C}(c_i, r)$ is simply πR^2. Thus the Voronoi distance definition contains the necessary information to compute approximately the local density function of the heterogeneous population, for a suitable choice of neighbors of each individual case.

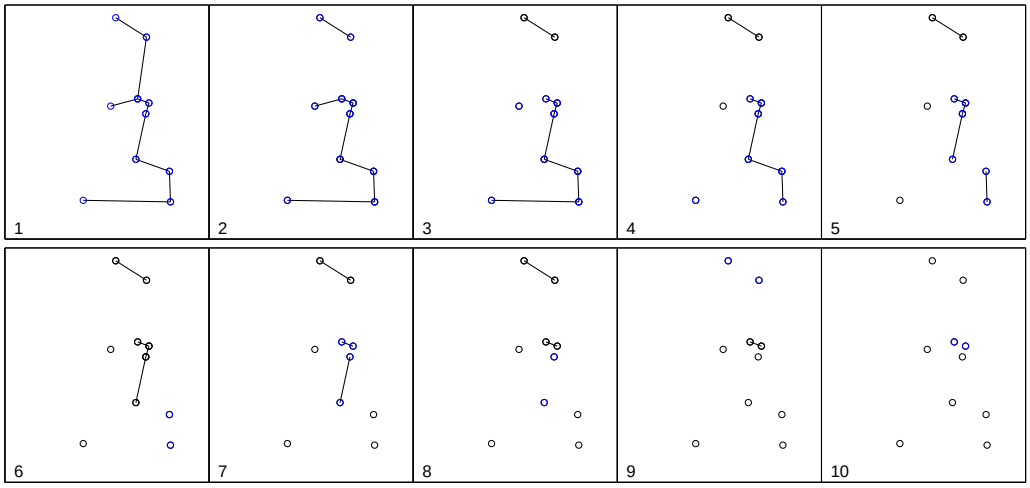

Figure 2.3: Visualization of the greedy edge deletion procedure, in successive steps numbered from 1 to 10. Sub-graphs linking blue circles represent the new cluster candidates that appear in each iteration, and sub-graphs linking black circles represent cluster candidates that have already appeared in former steps.

Proposition 1 *Consider a case dataset D and its corresponding VMST, denoted by $\mathcal{V}$. Let T_S be a connected subgraph of $\mathcal{V}$ whose nodes constitute the set S, and denote by $f(x)$ the local population density in x. For each case $c_i \in S$ let ω_i be equal to the minimum weight of the edges that are incident to c_i in $\mathcal{V}$ and $\mathcal{B} = \bigcup \mathcal{C}(c_i, \omega_i/2)$. The local population of S can be approximated by $\int_{\mathcal{B}} f(x)dx = \frac{1}{4}\sum_{c_i \in S} \pi\omega_i^2$.*

This defines a "region of influence" of the cluster S through the composition of the regions of influence of each case, which are defined as circular regions, with radii $\omega_i/2$ chosen as large as possible, such that there is no interference between neighboring circles in the VMST. An example is shown in Figure 2.4.

We further note that this definition is robust, in the following sense. Consider two situations: first, a case dataset D spread evenly in a map of control points, and second, a case dataset D' with the same number of points and overall shape as D but geographically smaller, inserted in the same map of control points. It is easy to see that the regions of influence

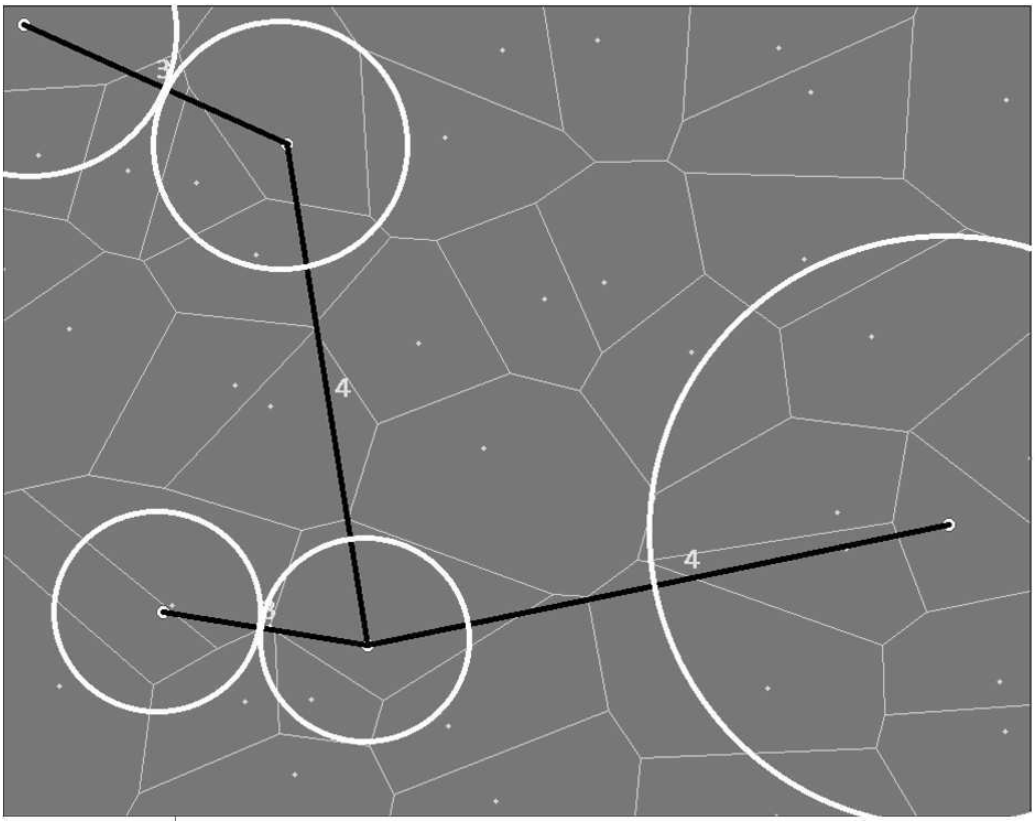

Figure 2.4: The "region of influence" of each individual case in an arbitrary map.

of the clusters associated to D is larger than the corresponding regions of influence associated with D', as we could expect.

We shall use this information to estimate the number of control individuals under the "region of influence" of each case individual, which in turn will allow the use of the scan statistic and also define a corresponding cluster finding algorithm employing a minimum spanning tree.

2.4 Voronoi based space-time scan

In order to deal with space-time clusters, a simple procedure that connects the cases of different time instants for each time interval is employed. On what follows, we specify a parameter τ to indicate the maximum allowed

temporal gap within the candidate cluster.

Let P_T be the set of the geographic coordinates of the $N-n$ controls and the n_T disease cases present in the interval time window given by $T = [s, t]$, where s is the initial time and t the final time of the interval T. The Voronoi diagram of P_T and the corresponding Voronoi distance is defined similarly to the former procedure, in space coordinates only. For the space-time domain, let t_i be the onset time of the disease for the i-th case, $i = 1, ..., n_T$. Then, establish connections linking only cases whose temporal distance is limited by τ.

Formally, let $D^T = \{c_i^{t_i} : i = 1, ..., n_T\}$ be the set of cases observed in the interval $T = [s, t]$, where $s \leq t_i \leq t$ and (x_i, y_i) indicates the geographic location for the $c_i^{t_i}$ case, $i = 1, ..., n_T$. In this way, two observed cases $c_i^{t_i}, c_j^{t_j} \in D^T$ will be connected if the temporal distance is such that $|t_i - t_j| \leq \tau$. We define a weighted complete graph $G^\tau(D^T) = (V^T, E^\tau)$ with vertex set $V^T = \{c_i^{t_i} : c_i^{t_i} \in D^T\}$ and edge set $E^\tau = \{(c_i^{t_i}, c_j^{t_j}) : c_i^{t_i}, c_j^{t_j} \in D^T, i \neq j, |t_i - t_j| \leq \tau\}$. The weights are the usual Voronoi distances between points (x_i, y_i) and (x_j, y_j).

The procedure is repeated for every time interval $T = [s, t]$ such that $Y_1 \leq s \leq t = Y_2$, as seen in the Prospective Space-Time Scan section, building a different Voronoi based MST for each time interval T.

When using the parameter value $\tau = 1$, the produced clusters of cases have no time gaps. Larger values of the parameter τ, otherwise, may produce clusters with cases separated by more than one unit of time, which could be undesirable in some circumstances. In the applications of the chapter 5, we consider several possible values for τ.

Chapter 3

Dynamic Programming based Scan

3.1 Motivation

In general, the greatest difficulty of methods for detection of spatial clusters is to identify over all subsets of the data the subset that corresponds to the pattern of discrepancy. The evaluation of all subsets is computationally infeasible for large dataset. Recently, several attempts have been developed in order to outline this problem. Many heuristics have appeared recently to compute approximate values that maximizes the logarithm of the likelihood ratio (Duczmal *et al.*, 2009), other methods have made to reduce the search space (Duczmal *et al.*, 2011; Wieland *et al.*, 2007; Demattei *et al.*, 2007). Neill's Fast Subset Scan (Neill, 2008) presented a significant advance in spatial methods for aggregated area maps, finding exactly the optimal irregularly spatial clusters.

In (Cancado, 2009), the spatial cluster detection problem is formulated as the classic knapsack problem. The problem can be modeled as a *bi-objective combinatorial optimization problem*. The set of non-dominated solutions of the problem contains the solution that maximizes the logarithm of the likeli-

hood ratio, $K(\mathbf{z})$. In this work, a knapsack problem (unconstrained versions) is proposed.

We propose an exact algorithm based on dynamic programming, Geographical Dynamic Scan, that empirically was able to solve instances up to large size within a reasonable computational time. The set of non-dominated solutions of the problem, computed efficiently, contains the solution that maximizes the logarithm of the likelihood ratio, $K(\mathbf{z})$. The method allows arbitrary shaped clusters, which can be a collection of disconnected or connected regions, taking into account a geometric constraint. Note that this is not a serious disadvantage, provided that there is not a huge gap between its areas. Finding exactly the optimal irregularly spatial clusters, the method allows multiple clusters.

We present an empirical comparison of detection and spatial accuracy between our algorithm and the classical Kulldorff's Circular Scan, using the data set of Chagas disease cases in puerperal women in Minas Gerais state, Brazil.

3.2 Multi-objective optimization problem

Multi-objective optimization deals with the problem of finding optimal solutions due to more than one objective function. A multi-objective optimization problem is formally defined as:

$$\begin{aligned} &\min \mathbf{f}(\mathbf{x}), \quad \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_m(\mathbf{x})) \\ &\text{subject to: } \mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathbf{X} \end{aligned} \tag{3.1}$$

in which $\mathbf{x} \in \mathbf{X} \subseteq \mathbb{R}^n$ is the *decision variable vector*, $\mathbf{X}$ is the *optimization parameter domain*, $\mathbf{Y} \subseteq \mathbb{R}^m$ is the *objective space*, i.e. $\mathbf{Y} = \mathbf{f}(\mathbf{X})$.

The goal of multi-objective optimization methods is to obtain a set of points belonging to the optimization parameter domain of the problem, such that they minimize, in a sense, a vector function. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$. We define

the following relational operators:

$$\begin{array}{lcll} & & \mathbf{u}, \mathbf{v} \in \mathbb{R}^m & \\ \mathbf{u} \leqq \mathbf{v} & \Longleftrightarrow & u_i \leq v_i,\ i = 1, \ldots, m & \\ \mathbf{u} \neq \mathbf{v} & \Longleftrightarrow & \exists i \in \{1, \ldots, m\} : u_i \neq v_i & \\ \mathbf{u} \leq \mathbf{v} & \Longleftrightarrow & \mathbf{u} \leqq \mathbf{v} \text{ and } \mathbf{u} \neq \mathbf{v} & \end{array}$$

These operators enable a well-defined definition of optimality in multi-objective optimization.

Definition 2 *A feasible solution* $\mathbf{x}^* \in \mathbf{X}$ *is called optimal solution of a multi-objective optimization problem if there is no* $\mathbf{x} \in \mathbf{X}$ *such that* $\mathbf{f}(\mathbf{x}) \leq \mathbf{f}(\mathbf{x}^*)$.

In that case $\mathbf{f}(\mathbf{x}^*)$ is called optimal value of the multi-objective optimization problem.

Given the multi-objective optimization problem (3.1), a decision vector $\mathbf{x}$ dominates another decision vector $\mathbf{x}'$ and $\mathbf{f}(\mathbf{x})$ dominates $\mathbf{f}(\mathbf{x}')$ if and only if $\mathbf{f}(\mathbf{x}) \leq \mathbf{f}(\mathbf{x}')$. An optimal solution is called non-dominated vector. In this way, the non-dominated set of solutions, or the Pareto-optimal set, $\mathcal{P}$, is defined as:

$$\mathcal{P} = \{\mathbf{x}^* \mid \nexists \mathbf{x} \in \mathbf{X} : \mathbf{f}(\mathbf{x}) \leq \mathbf{f}(\mathbf{x}^*),\ \mathbf{x} \in \mathbf{X}\}\,. \tag{3.2}$$

A Pareto-optimal solution is a non-dominated vector $\mathbf{x} \in \mathbf{X}$. The Pareto-optimal set of the multi-objective optimization problem is the set of all Pareto-optimal solutions. The image of this set in the objective space is called the *Pareto front* and denoted as $\mathbf{f}(\mathcal{P})$.

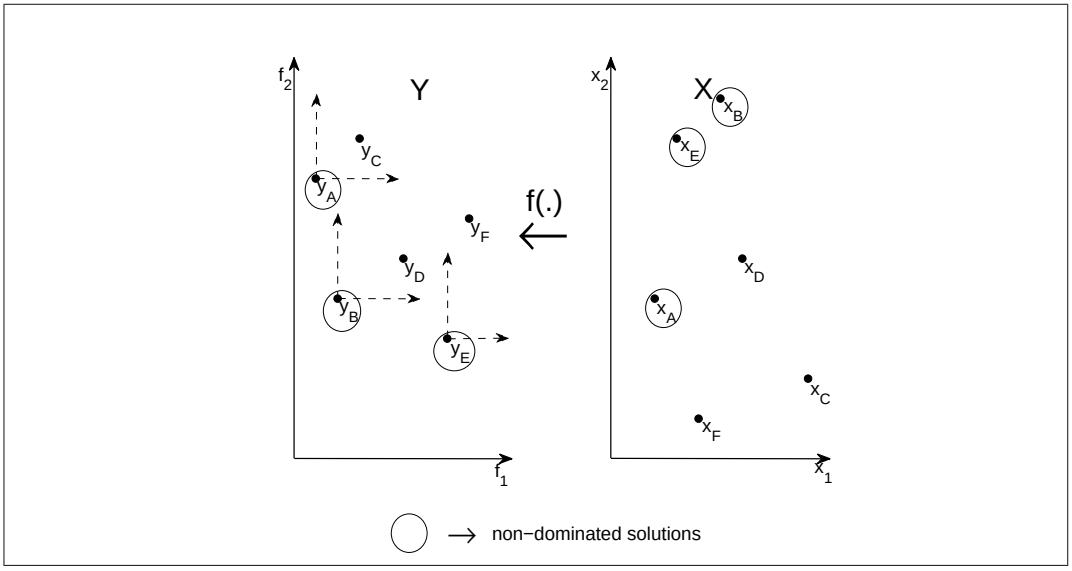

Figure 3.1: Non-dominated solutions. Right: the Pareto-optimal set. Left: the *Pareto front set*.

3.3 System and Method

In this section we propose an algorithm to solve the problem of detecting clusters using dynamic programming.

Dynamic programming is a stage-wise search method suitable for optimization problems whose solutions may be viewed as the result of a sequence of decisions (Gupta *et al.*, 2008). The most attractive property of this strategy is that during the search for a solution it avoids full enumeration by pruning early partial decision solutions that cannot possibly lead to optimal solution. The dynamic programming relies on a principle of optimality. This principle states that in an optimal sequence of decisions or choices, each subsequence must also be optimal.

3.3.1 Mathematical Formulation

The typical detection of spatial clusters problem (1.1) across the set Z of all possible zones, is reformulated, in the same way as done in (Cancado, 2009). The proposed method this work enumerates a subset of feasible solu-

tions, and discards those that will not lead to optimal solutions.

Given a map with m regions, consider the binary variables $x_1, ..., x_m$, where $x_i = 1$ if the i-th region is present in the cluster and 0 otherwise. Let $\mathbf{c_i}$ and $\mathbf{n_i}$ denote the number of cases and population of i-th region. Consider the following unconstrained bi-objective combinatorial optimization problem:

$$\begin{aligned} \min \quad & \mathbf{f}(\mathbf{x}) = \left(\mathbf{C}(\mathbf{x}) = -\sum_{i=1}^{m} \mathbf{c_i} x_i, \quad \mathbf{N}(\mathbf{x}) = \sum_{i=1}^{m} \mathbf{n_i} x_i \right) \\ \text{s.t.} \quad & \mathbf{x} \in \{0,1\}^m \end{aligned} \tag{3.3}$$

A non-dominated solution $\mathbf{x} = (x_i, ..., x_m)$ of the Problem (3.3) above, represents a subset of regions (zone) of the map with number of cases $|\mathbf{C}(\mathbf{x})|$ and population $\mathbf{N}(\mathbf{x})$.

The following proposition shows that is possible to find the maximum of the function K by solving Problem (3.3), see (Cancado, 2009).

Proposition 2 *The set of non-dominated solutions of Problem (3.3) contains the solution that maximizes K.*

Proof: Let $\mathcal{P}$ be the set of non-dominated solutions of Problem (3.3)

$$\mathcal{P} = \{\mathbf{x} \mid \nexists \mathbf{x}^* \in \{0,1\}^m : \mathbf{f}(\mathbf{x}^*) \leq \mathbf{f}(\mathbf{x}),\ \mathbf{x} \in \{0,1\}^m\},$$

and $\mathbf{x}^{**}$ the subset of regions of the map that maximizes the function K, with number of cases $|\mathbf{C}(\mathbf{x}^{**})|$ and population $\mathbf{N}(\mathbf{x}^{**})$. We show that if $\mathbf{x}^{**} \notin \mathcal{P}$ then it leads to a contradiction. Indeed, if $\mathbf{x}^{**} \notin \mathcal{P}$ then there exists a pair $(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x}))$ such that $\mathbf{C}(\mathbf{x}) \leq \mathbf{C}(\mathbf{x}^{**})$ and $\mathbf{N}(\mathbf{x}) \leq \mathbf{N}(\mathbf{x}^{**})$, with at least one inequality being strict. Hence, as K satisfies the property 1, it follows that:

1. If $\mathbf{C}(\mathbf{x}) < \mathbf{C}(\mathbf{x}^{**})$ and $\mathbf{N}(\mathbf{x}) = \mathbf{N}(\mathbf{x}^{**})$ then
   $K(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x})) = K(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x}^{**})) > K(\mathbf{C}(\mathbf{x}^{**}), \mathbf{N}(\mathbf{x}^{**}))$;

2. If $\mathbf{C}(\mathbf{x}) = \mathbf{C}(\mathbf{x}^{**})$ and $\mathbf{N}(\mathbf{x}) < \mathbf{N}(\mathbf{x}^{**})$ then
   $K(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x})) = K(\mathbf{C}(\mathbf{x}^{**}), \mathbf{N}(\mathbf{x})) > K(\mathbf{C}(\mathbf{x}^{**}), \mathbf{N}(\mathbf{x}^{**}))$;

3. If $\mathbf{C}(\mathbf{x}) < \mathbf{C}(\mathbf{x}^{**})$ and $\mathbf{N}(\mathbf{x}) < \mathbf{N}(\mathbf{x}^{**})$ then
   $K(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x})) > K(\mathbf{C}(\mathbf{x}^{**}), \mathbf{N}(\mathbf{x})) > K(\mathbf{C}(\mathbf{x}^{**}), \mathbf{N}(\mathbf{x}^{**}))$;

All cases above implies that $K(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x})) > K(\mathbf{C}(\mathbf{x}^{**}), \mathbf{N}(\mathbf{x}^{**}))$. This shows that if $\mathbf{x}^{**}$ is the solution that maximizes the function K, then $\mathbf{x}^{**} \in \mathcal{P}$. □

3.3.2 Dynamic programming algorithm

In this section, we introduce the dynamic programming approach, which is an adaptation of the Nemhauser-Ullman algorithm for the $\{0, 1\}$ knapsack problem (Nemhauser & Ullmann, 1969).

The algorithm that we use to solve Problem (3.3) can be regarded as a dynamic programming algorithm. In what follows, we present the theoretical background of the algorithm. Given a map with m regions and a random list enumerated of the regions, a zone of the map can be represented by a vector $\mathbf{x} \in \{0, 1\}^m$. Let sets $\mathcal{Z}^i = \{(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x})) \mid x_k = 0, \forall k > i, \ \mathbf{x} \in \{0, 1\}^m\}$ that represent all zones with at most i-th first regions (given a random list enumerated of the regions), $i = 0, \ldots, m$. We call $\mathbf{z}^i = (\mathbf{z}^{\mathbf{i}}_{\mathbf{C}}, \mathbf{z}^{\mathbf{i}}_{\mathbf{N}}) \in \mathcal{Z}^i$ a state. The inclusion chain

$$\{(0, 0)\} = \mathcal{Z}^0 \subseteq \mathcal{Z}^1 \subseteq \ldots \subseteq \mathcal{Z}^m = \mathcal{Z}$$

holds by definition of the sets $\mathcal{Z}^i$. Note that the states in $\mathcal{Z} = \mathcal{Z}^m$ represent the image of the feasible solutions of Problem (3.3), i.e. $\{\mathbf{z} = (\mathbf{z_C}, \mathbf{z_N}) \in \mathcal{Z}\} = \{(\mathbf{C}(\mathbf{x}), \mathbf{N}(\mathbf{x})) \mid \mathbf{x} \in \{0, 1\}^m\}$. Hence, we can naturally define the concept of dominance between two states. We say that a state $\mathbf{z} = (\mathbf{z_C}, \mathbf{z_N}) \in \mathcal{Z}$ dominates a state $\mathbf{z}' = (\mathbf{z}'_{\mathbf{C}}, \mathbf{z}'_{\mathbf{N}}) \in \mathcal{Z}$ if $(\mathbf{z_C}, \mathbf{z_N})$ dominates $(\mathbf{z}'_{\mathbf{C}}, \mathbf{z}'_{\mathbf{N}})$.

Definition 3 *A state $\mathbf{z} \in \mathcal{Z}$ is called extension of a state $\mathbf{z}^i = (\mathbf{z}^{\mathbf{i}}_{\mathbf{C}}, \mathbf{z}^{\mathbf{i}}_{\mathbf{N}}) \in \mathcal{Z}^i$, $i < m$, if $\mathbf{z} = (\mathbf{z}^{\mathbf{i}}_{\mathbf{C}} - \mathbf{c_j}, \mathbf{z}^{\mathbf{i}}_{\mathbf{N}} + \mathbf{n_j})$ for some $j \in \{i + 1, \ldots, m\}$. If $j = i + 1$, the state $\mathbf{z}$ is called successor of $\mathbf{z}^i$, denoted by $s(\mathbf{z}^i)$.*

This notion of states matches the notion of zones.

Example. Consider an arbitrary disease map with 20 locations and its distribution of population at risk and cases according to the Figure 3.2. The solution $\mathbf{x1} = (\mathbf{0}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, ..., \mathbf{0})$, see the Figure 3.3, represent

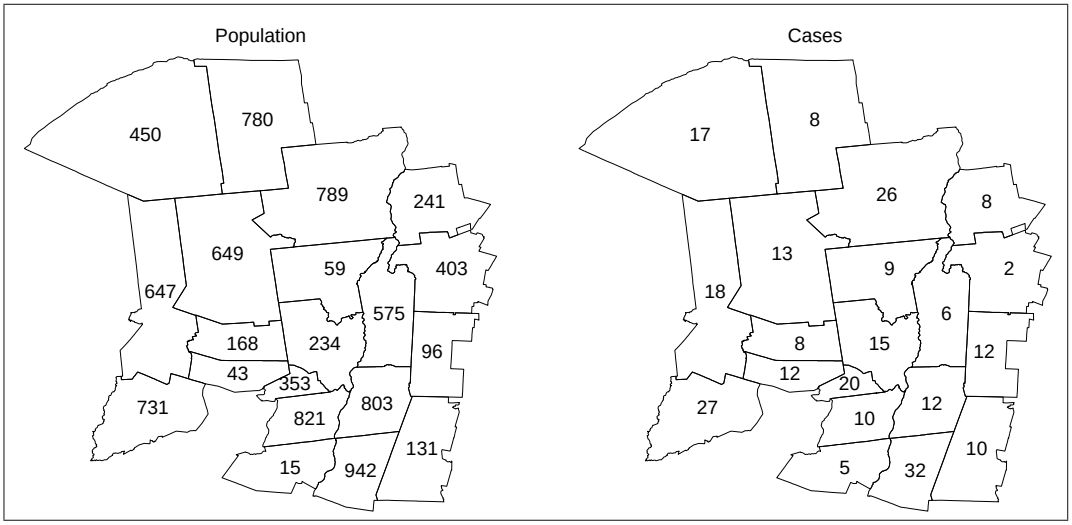

Figure 3.2: An arbitrary map with 20 locations and its distribution of population at risk and cases per location.

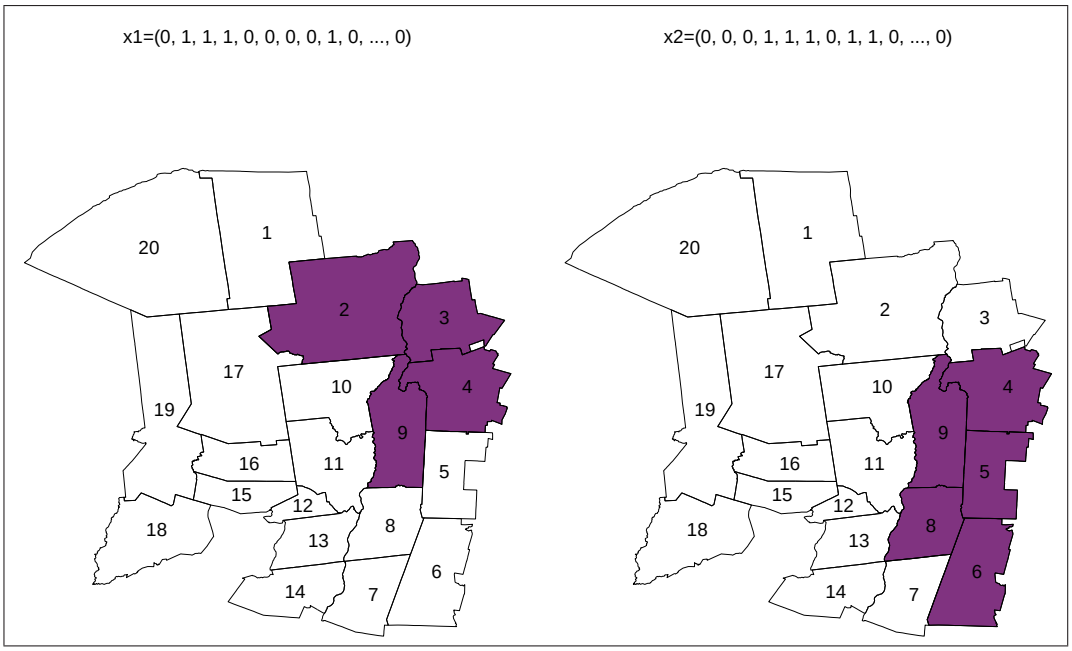

Figure 3.3: Different solutions mapped to same state $\mathbf{z} = (-\mathbf{42}; \mathbf{2008}) \in \mathcal{Z}^9$.

the state $\mathbf{z} = (-\mathbf{42}; \mathbf{2008}) \in \mathcal{Z}^9$. Note that the different solution given by binary vector $\mathbf{x2} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1}, \mathbf{0}, ..., \mathbf{0})$ is mapped by the same

state $\mathbf{z} = (-\mathbf{42}; \mathbf{2008}) \in \mathcal{Z}^9$

By the definition of a state successor and by construction we have the following recursion formula

$$\mathcal{Z}^{i+1} = \mathcal{Z}^i \cup \left\{ s(\mathbf{z}^i) \mid \mathbf{z}^i \in \mathcal{Z}^i \right\}, \tag{3.4}$$

for $i = 0, \ldots, m-1$.

We now state a theorem that justifies the use of dynamic programming in order to solve the problem of maximizing K.

Theorem 1 *Let $i < m$. If $\mathbf{z} = (\mathbf{z_C}, \mathbf{z_N}) \in \mathcal{Z}^i$ is dominated by $\mathbf{z}' = (\mathbf{z}'_\mathbf{C}, \mathbf{z}'_\mathbf{N}) \in \mathcal{Z}^i$, then there is an extension of $\mathbf{z}'$ that will dominate any extension of $\mathbf{z}$.*

Proof: Let $\mathbf{z} = (\mathbf{z_C}, \mathbf{z_N}) \in \mathcal{Z}^i$ be dominated by $\mathbf{z}' = (\mathbf{z}'_\mathbf{C}, \mathbf{z}'_\mathbf{N}) \in \mathcal{Z}^i$ and let the state $\mathbf{ext}(\mathbf{z}) = (\mathbf{z_C} - \mathbf{c_{j_z}}, \mathbf{z_N} + \mathbf{n_{j_z}}) \in \mathcal{Z}$ denote an extension of $\mathbf{z}$ with index $j_z \in \{i+1, \ldots, m\}$. Make the extension $\mathbf{ext}(\mathbf{z}')$ of $\mathbf{z}'$ defined by setting $j_{z'} = j_z$. It has to be shown that $\mathbf{ext}(\mathbf{z}')$ dominates $\mathbf{ext}(\mathbf{z})$. Because $\mathbf{z}$ is dominated by $\mathbf{z}'$, we get $\mathbf{z}'_\mathbf{C} \leq \mathbf{z_C}$ and $\mathbf{z}'_\mathbf{N} \leq \mathbf{z_N}$ where at least one inequality is strict. By definition of $\mathbf{ext}(\mathbf{z}')$ and $\mathbf{ext}(\mathbf{z})$ the same values are added to $\mathbf{z}'_\mathbf{C}, \mathbf{z}'_\mathbf{N}$ and $\mathbf{z_C}, \mathbf{z_N}$. Therefore, $\mathbf{z}'$ dominates $\mathbf{z}$ implies that $\mathbf{ext}(\mathbf{z}')$ dominates $\mathbf{ext}(\mathbf{z})$. □

The basic idea of the dynamic programming algorithm we are considering in the following is based on Theorem 1. The algorithm generates a sequence of sets of states $\mathcal{Z}^i$, $i = 0, \ldots, m$. The set $\mathcal{Z}^{i+1}$ contains successors of the states in $\mathcal{Z}^i$ but not those states that are dominated since they do not lead to non-dominated states. We rewrite recursive formula Eq. (3.4) as follows:

$$\mathcal{Z}^{i+1} = \max \left\{ \mathcal{Z}^i \cup \left\{ s(\mathbf{z}^i) \mid \mathbf{z}^i \in \mathcal{Z}^i \right\} \right\}, \tag{3.5}$$

for $i = 0, \ldots, m-1$, where "max" denotes component-wise maxima.

While generating sets of states, a non-dominated operator ND perform the deletion of dominated states. Therefore, the algorithm in its final phase generates a set of states $\mathcal{Z}^m = \mathcal{Z}$, the set of non-dominated solutions of

Problem (3.3). Since $\mathcal{Z}$ represents the image of the feasible solutions that maximizes the logarithm of the likelihood ratio, the algorithm test which one maximizes K function.

We showed that the dynamic programming algorithm allows to solve the unconstrained maximization of the function K for a spatial dataset. However, note that the unconstrained maximization over subsets of the Problem (3.3) is typically not sufficient to solve practical spatial detection problem. The method allows arbitrary shaped cluster, which can be a collection of regions with high likelihood that spreads randomly across the map. In the following section, we modify the dynamic programming algorithm in order to take into account a geometric constraint.

3.4 Geographical dynamic scan

The dynamic programming algorithm explained in the previous section is modified in order to consider a *geographical proximity constraint*. Consider a map with m regions and a fixed index k, $1 < k < m$. For each region i, we define a centroid c_i, an arbitrary point in its interior, $i = 1, ..., m$. Let $d(c_i, c_j)$ be the Euclidean distance between any two centroids c_i and c_j of the map. Then, for each region i, we define its geographical proximity G_i to be the region i and its $k-1$ nearest neighbors regarding the distance to the centroid c_i. We use the dynamic programming approach to find the non-dominated solutions of G_i for each region i. From the set of all non-dominated solutions found for every region, we choose the one that is maximal with respect to function K. Algorithm 1 introduces our approach to solve the classical spatial cluster detection problem.

Note that the geographical proximity constraint adopted is the same defined in the classical Kulldorff method, circular scan, see the Figure 3.4. However, assuming that the geographical proximity of a region i contains k regions, while the circular scan only evaluates k of the 2^k subsets, the geographical dynamic scan guarantees the optimal solution "evaluating" efficiently the 2^k subsets. Furthermore, another difference between the two methods is that classical Kulldorff requires the resulting region to be con-

Algorithm 1 Geographical dynamic scan algorithm

1. Let $S = \emptyset$;
2. Define neighborhood size k and centroids c_i for each region $i = 1, ..., m$ of the map;
3. For each region $i = 1, \ldots, m$
   (a) Let G_i be the geographical proximity for each region $i = 1, ..., m$;
   (b) Let S_{nd_i} be the non-dominated set of Problem (3.3) using the dynamic programming algorithm with input data G_i;
   (c) $S := S \cup S_{nd_i}$;
   (d) $S := ND(S)$, where ND define a non-dominated operator.
4. $s := \max\{K(S)\}$;
5. Return s.

nected, while our algorithm can return a disconnected region if it satisfies the geographical proximity constraint.

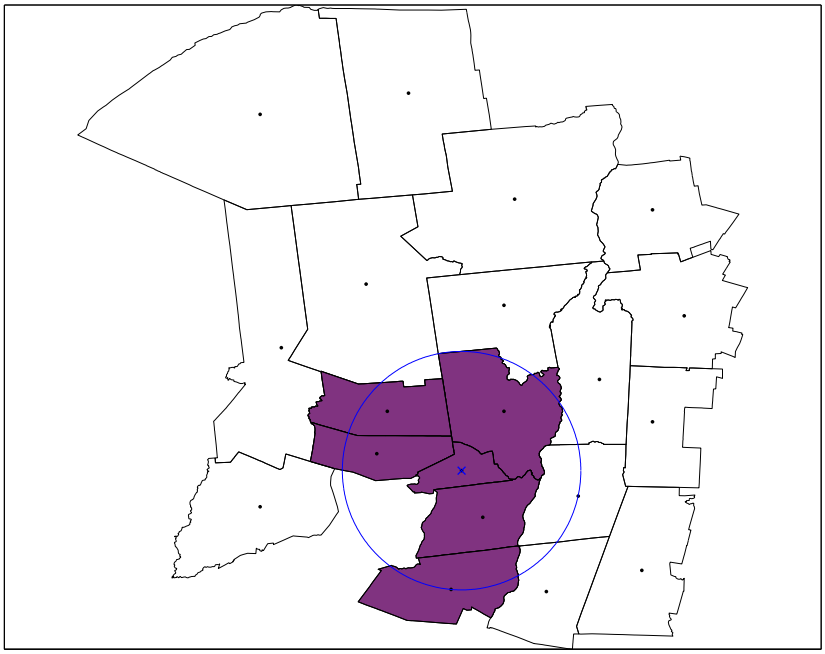

Figure 3.4: The geographical proximity for the region with cross centroid when neighborhood size $k = 5$.

Chapter 4

Results and Discussion

In this chapter we evaluate the numerical performance of Voronoi Based Scan and Geographical Dynamic Scan algorithms proposed in this work.

4.1 Evaluated Measures

A good detection method is that it is sensitive enough to detect a cluster when it actually exists. We will evaluate the efficiency of the algorithms in this thesis calculating their power.

Definition 4 (Power) *The power of a statistical test measures the test's ability to reject the null hypothesis when it is actually false.*

In other words, the power of a hypothesis test is the probability of not committing a type II error. We can estimate the power by Monte Carlo simulations, running the algorithm a large number of times in artificial settings, constructed so that there is the presence of a cluster. The maximum power a test can have is 1, the minimum is 0. Ideally we want a test to have high power, close to 1.

We also use the measures of sensitivity and positive predicted value (ppv) that serve to evaluate the quality of the cluster detection process. The measures were defined differently according to the form of spatial data evaluated.

For the aggregated spatial data, the sensitivity and positive predicted value (ppv) are defined in the terms of the population size 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 case-control data spatial set, let $\{X_1, X_2, \ldots, X_n\}$ be random variables that denote the spatial coordinates of n cases observed in the data set. The sensitivity and positive predicted value are defined as

$$Sensitivity = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \in \text{Detected Cluster} \cap \text{Real Cluster})}{\sum_{i=1}^{n} \mathbb{1}(X_i \in \text{Real Cluster})}$$

$$PPV = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \in \text{Detected Cluster} \cap \text{Real Cluster})}{\sum_{i=1}^{n} \mathbb{1}(X_i \in \text{Detected Cluster})}$$

where $\mathbb{1}(.)$ is the indicator function.

Using artificial clusters, the measures of power, sensitivity and positive predicted value of the algorithms are estimated. In each scenario a relative risk equal to 1.0 was set for every region (considering aggregated spatial data) and every control (considering case-control spatial data) outside the real cluster, and greater than 1.0 and identical otherwise. 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 would be 0.999 (Kulldorff *et al.*, 2003).

4.2 Numerical Tests

The Voronoi Based Scan and the Geographical Dynamic Scan are compared through numerical simulations to the elliptic scan statistic.

4.2.1 Voronoi Based Scan

In this section we present a set of numerical results. The Voronoi Based Scan (VBScan) was compared numerically with the elliptic version of the spatial scan and prospective space-time scan (Kulldorff *et al.*, 2006; Kulldorff, 2001), according to power of detection, sensitivity and positive predictive value.

In the first set of simulations, we evaluated only the spatial structure of the proposed algorithm.

A verification for purely spatial clusters

The Voronoi based method, in its purely spatial setting, is applied for the well known data set of residential locations of larynx and lung cancer cases of the Chorley-Ribble area in Lancashire-UK, from 1973 to 1984. The 917 lung cancer cases are used as controls for the 57 larynx cancer cases (see http://cran.r-project.org/web/packages/splancs/splancs.pdf - pag. 55). In Figure 4.1 the spatial distribution of the observed cases (circles) and controls (dots) is shown on the left, and the Voronoi minimum spanning tree is shown on the right, with the Voronoi cells in the background. The elliptic spatial scan is also run as comparison. The p-values associated to the two scans are computed based on $9,999$ Monte-Carlo simulations under the null hypothesis. The most likely clusters found in both runs are identical, consisting of the five cases (triangles) of Figure 4.1. Table 4.1 shows the likelihood values, number of cases, p-values and running times for both scans. The set of possible elliptic clusters forms a more restrictive space of configurations than the set of of irregularly shaped clusters; not surprisingly, the elliptic scan p-value is smaller than the VBScan p-value, because the five cases in the most likely cluster fit very well inside an elongated ellipse.

An additional artificial dataset with total population at risk of $1,000$ individuals, 100 cases, was also used. The instance was simulated in the map constructed with the spatial locations of population at risk following an uniform point process within the square $[0,1] \times [0,1]$. Different spatial cluster geometries were evaluated. The three spatial cluster zones, as shown

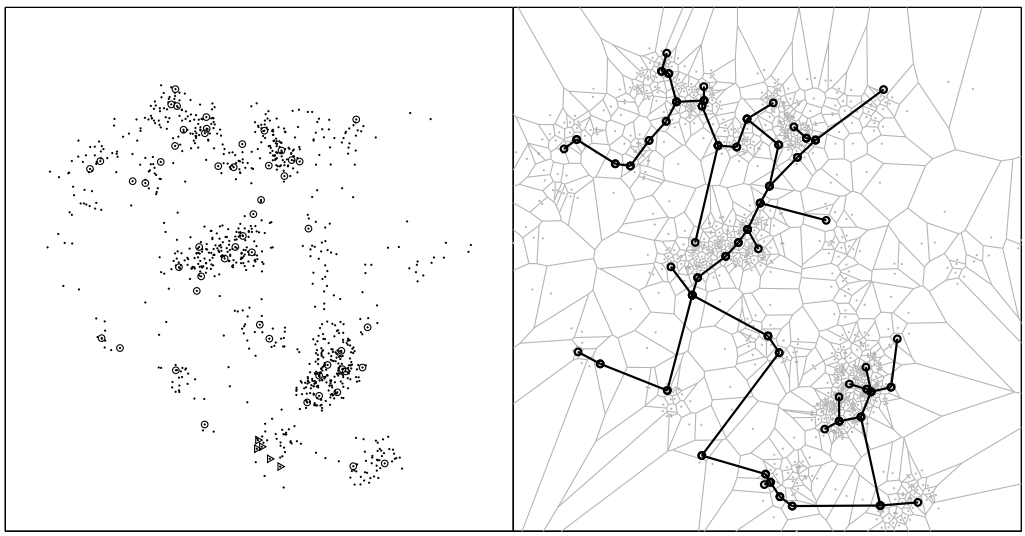

Figure 4.1: Left: Spatial distribution of the observed cases (circles) and controls (dots) in Lancashire-UK and the most likely cluster (triangles). Right: associated Voronoi minimum spanning tree.

Table 4.1: Comparisons spatial clusters detection of the cancer in Lancashire, match values to elliptic scan and VBScan methods.

Method	LLR	cases	p-value	CPU-Time(sec.)
Elliptic Scan	14.4049	5	0.0089	896
VBScan	10.8357	5	0.0470	449.5

in Figure 4.2, aggregate spatial areas:

1. A circular shaped cluster was simulated with radius equal to 0.195.

2. A "T-2D"-shaped cluster was simulated with zone $T = T_1 \cup T_2$ where $T_1 = [0.2, 0.4] \times [0.5, 0.8]$, $T_2 = [0.0, 0.6] \times [0.8, 0.9]$.

3. An "L-2D"-shaped cluster was simulated with zone $L = L_1 \cup L_2$ where $L_1 = [0.2, 0.4] \times [0.5, 0.8]$, $L_2 = [0.2, 0.8] \times [0.8, 0.9]$.

Given a cluster model, exactly the same sets of data were used for all algorithms. $10,000$ Monte Carlo simulations of the null hypothesis were performed, and also $10,000$ Monte Carlo replications for each one of the three alternative hypothesis models. The three measures above, namely, detection

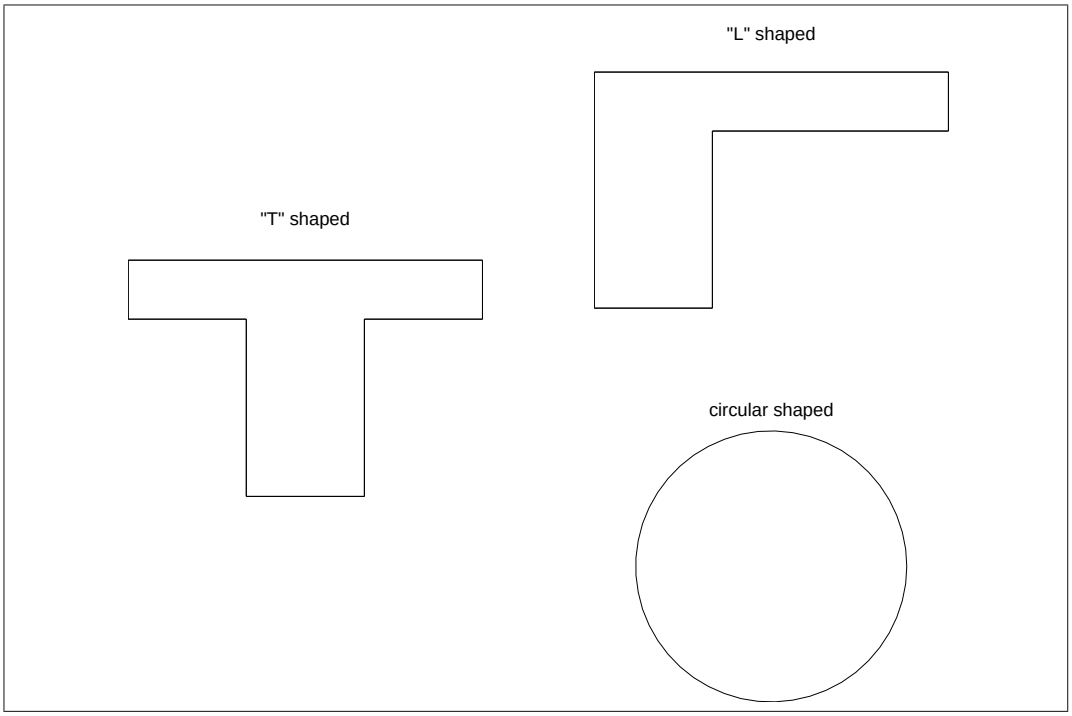

Figure 4.2: Three alternative artificial spatial clusters.

power, sensitivity an PPV were computed for the most likely cluster in each replication.

Table 4.2 shows the results. The power and PPV values are slightly higher for the elliptic spatial scan than for the Voronoi based method but the sensitivity is lower for the elliptic scan. In addition, the Voronoi based method requires less computational time for point data set compared to the elliptic scan statistic.

By Proposition 1, we attached a ball of radius $\omega_i/2$ to each case c_i belonging to the cluster S. The value ω_i was chosen as the minimum weight of the edges that are incident to c_i in the VMST. An alternative definition may use the average (or even the median) of the weights of the edges that are incident to c_i, instead of the minimum value of the weights. We have conducted numerical simulations suggesting that there are negligible differences of performance using these alternative definitions, compared with the

Table 4.2: Power, positive predicted value and sensitivity comparisons for shaped spatial clusters

shape of cluster	**Power**		**Sensitivity**		**PPV**	
	Elliptic	VBScan	Elliptic	VBScan	Elliptic	VBScan
Circle	0.8400	0.7963	0.7257	0.8199	0.8347	0.7871
"T-2D"	0.7320	0.7067	0.5508	0.7270	0.7837	0.7398
"L-2D"	0.7206	0.6696	0.5501	0.7144	0.7740	0.6932

original definition using the minimum value of the weights, see Table 4.3. This is a good indication that proposed definition of local population of the cluster is stable.

Table 4.3: Power, positive predicted value and sensitivity comparisons for alternatives values of ω_i.

ω_i	shaped cluster	**Power**	**Sensitivity**	**PPV**
minimum edge weight	Circle	0.7004	0.8260	0.7771
	"T-2D"	0.5910	0.7326	0.7263
	"L-2D"	0.5703	0.7248	0.6801
average edge weight	Circle	0.7963	0.8196	0.7873
	"T-2D"	0.7075	0.7278	0.7404
	"L-2D"	0.6716	0.7152	0.6932
median edge weight	Circle	0.7921	0.8149	0.7888
	"T-2D"	0.6982	0.7225	0.7424
	"L-2D"	0.6705	0.7112	0.6970

Analysis of the Voronoi based space-time scan

We used artificial datasets with total population at risk of $1,000$ individuals, including 100 cases and 900 controls. The instances were simulated with a square space region $[0,1] \times [0,1]$ and a time interval $[1,10]$. Space-time clusters with different shapes were considered. Numerical simulations were conducted using an artificial map constructed with the spatial locations of the individuals of the population at risk following an uniform point pro-

cess, and the time of occurrence of the events following a discrete uniform distribution.

The Voronoi based method was compared to the prospective elliptic space-time scan statistic. Three alternative models of space-time clusters with different shapes were simulated. The three space-time cluster zones, as shown in Figure 4.3, aggregate spatial areas in consecutive time coordinates:

1. A cylinder shaped cluster was simulated with radius of the circular base and height equal to 0.198 and $[3, 6]$, respectively.

2. A cone shaped cluster was simulated as a frustum of a cone. The radius of lower and upper circular base were equal to 0.115 and 0.265, respectively. The time window was equal to $[3, 6]$.

3. An "L-3D"-shaped cluster was simulated with zone $L = L_1 \cup L_2$ where $L_1 = [0.3, 0.7] \times [0.3, 0.7] \times [3, 4]$, $L_2 = [0.484, 0.7] \times [0.3, 0.7] \times [5, 6]$.

Table 4.4 presents the resulting average power, sensitivity and PPV for $10,000$ replications of each one of the three cluster models obtained with the VBScan and Elliptic PST algorithms. For all three space-time clusters, the power of detection of the VBScan was higher than the power of the Elliptic PST. This also occurs for PPV and Sensitivity. The results found in the three measures evaluated for "L-3D"-shaped cluster show the greater flexibility of VBScan, compared with Elliptic PST method.

Table 4.4: Power, sensitivity and positive predicted value comparisons for the three alternatives space-time clusters.

	Power		**Sensitivity**		**PPV**	
shaped cluster	Elliptic PST	VBScan	Elliptic PST	VBScan	Elliptic PST	VBScan
Cylinder	0.4789	0.6510	0.5447	0.6532	0.6415	0.6738
Cone	0.3863	0.5093	0.4683	0.5947	0.5822	0.6157
"L-3D"	0.3316	0.5768	0.4530	0.6141	0.5323	0.5943

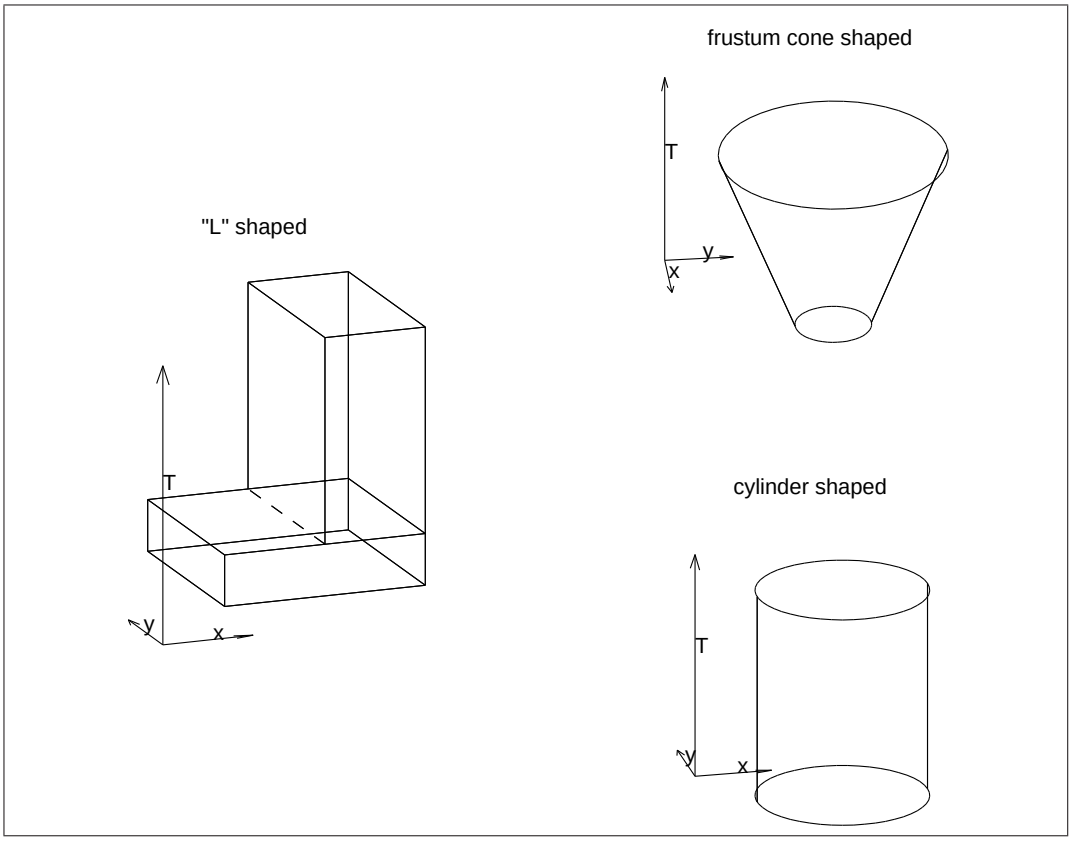

Figure 4.3: Three alternative artificial space-time clusters.

4.2.2 Geographical dynamic scan

We now present an empirical comparison of detection and spatial accuracy for the Geographical Dynamic Scan algorithm proposed and the classical Kulldorff method, circular scan, using the data set of Chagas' disease cases in puerperal women in Minas Gerais state, Brazil employed in Oliveira *et al.* (2011). 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.nupadmedicina.ufmg.br in collaboration with Minas Gerais State Health Secretary. The state is divided into 853 municipalities with a total population at risk of 63,519 women. After a comprehensive screening to eliminate false positives a total number of 803 cases were obtained. The raw rates map is presented in Figure 4.4(a) and the population at risk map in Figure 4.4(b).

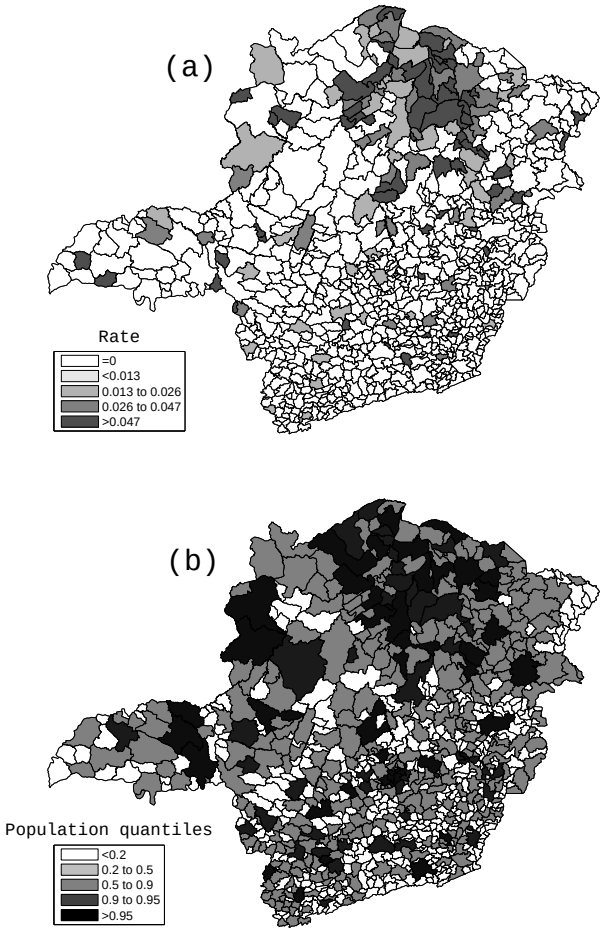

Figure 4.4: Mapping spatial variations of Chagas disease in the State of Minas Gerais - Brazil by county during 2006: (a) disease rates map; (b) Population at risk map.

Verifications

In the first set of simulations, we use the data set of Chagas to show the performance for the geographical dynamic scan. In the Figure 4.5, we report the total run time required versus fixed neighborhood size k to analyze of Chagas' data set. The geographical dynamic scan method is able to solve within a reasonable computational time with increasing neighborhood size, while we note that the run time increases exponentially with neighborhood size without the use of the dynamic programming method (naive search).

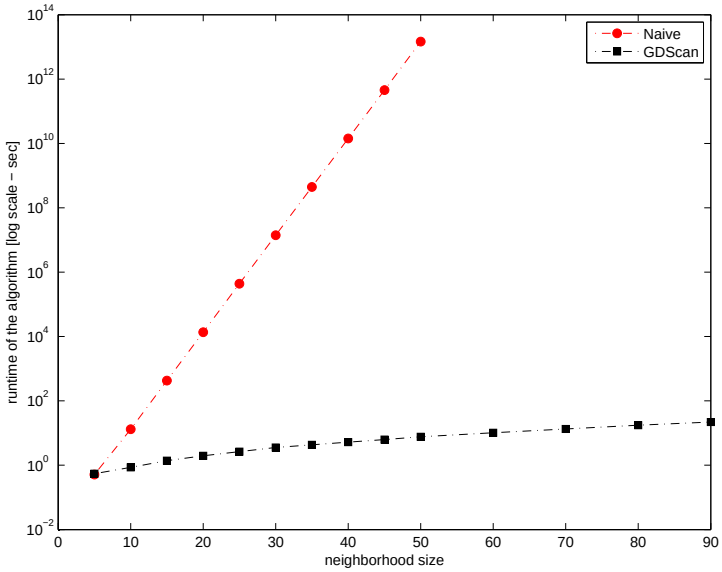

Figure 4.5: Run time versus neighborhood size k for the data set of Chagas, with and without the dynamic programming method.

Figure 4.6 shows the trade-off between quality solution and runtime of the algorithms. The geographical dynamic scan algorithm finds the spatial region with the global maximum value of logarithm of the likelihood ratio K. In Figure 4.6 bottom, we compared the different gaps between the values of logarithm of the likelihood ratio K found running the geographical dynamic scan and the classical Kulldorff's methods. Occasionally, the classical Kulldorff's method can find the global maximum value of logarithm of the likelihood ratio K, in this case occurs only for the geometric constraint size $k = 10$. On the other hand, Figure 4.6 top shows an increase in the runtime

of the geographical dynamic scan method, but a computational time quite plausible.

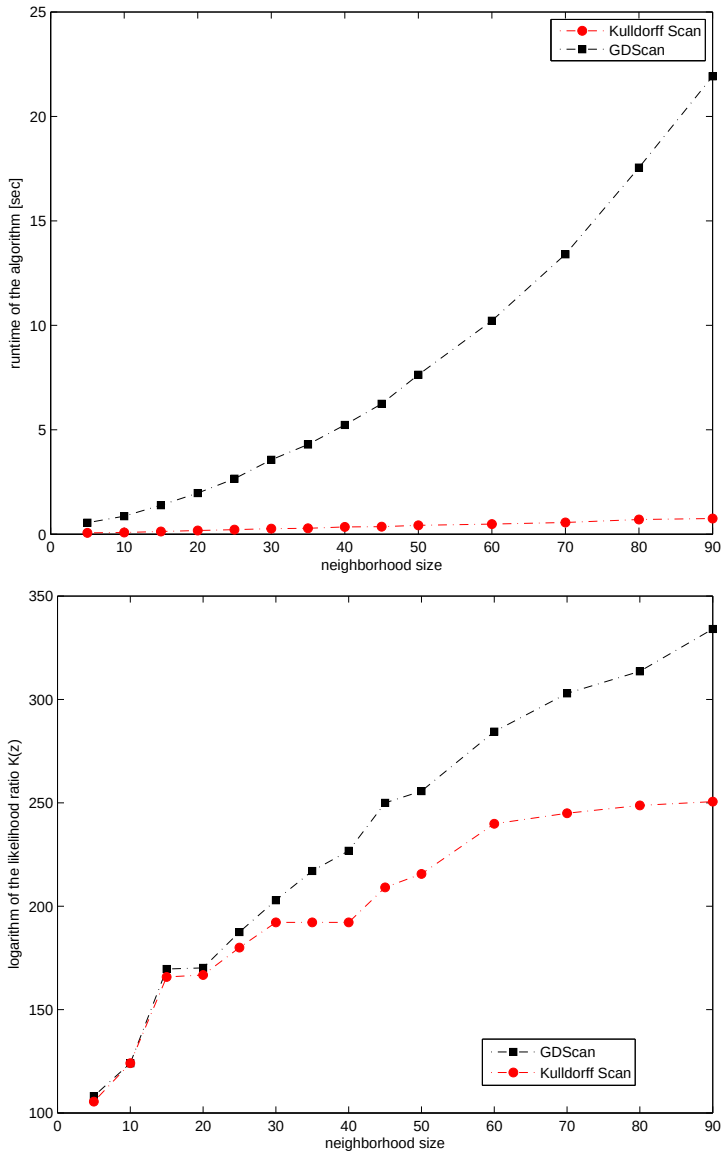

Figure 4.6: Comparisons between the dynamic programming and the classical Kulldorff methods for the data set of Chagas. Bottom: logarithm of the likelihood ratio versus geometric constraints size k. Top: runtime versus geometric constraints size k.

Figures 4.7 and 4.8 show the different clusters found by two methods,

respectively for the neighborhood size $5, 20, 50, 90$. Of all the clusters found by dynamic programming scan method, shown in Figure 4.7, only one cluster (with neighborhood size 50) appears disjointed. Figure 4.8 shows that even for small windows size, the clusters found by the classical Kulldorff scan overestimates the "optimal clusters" (Figure 4.7).

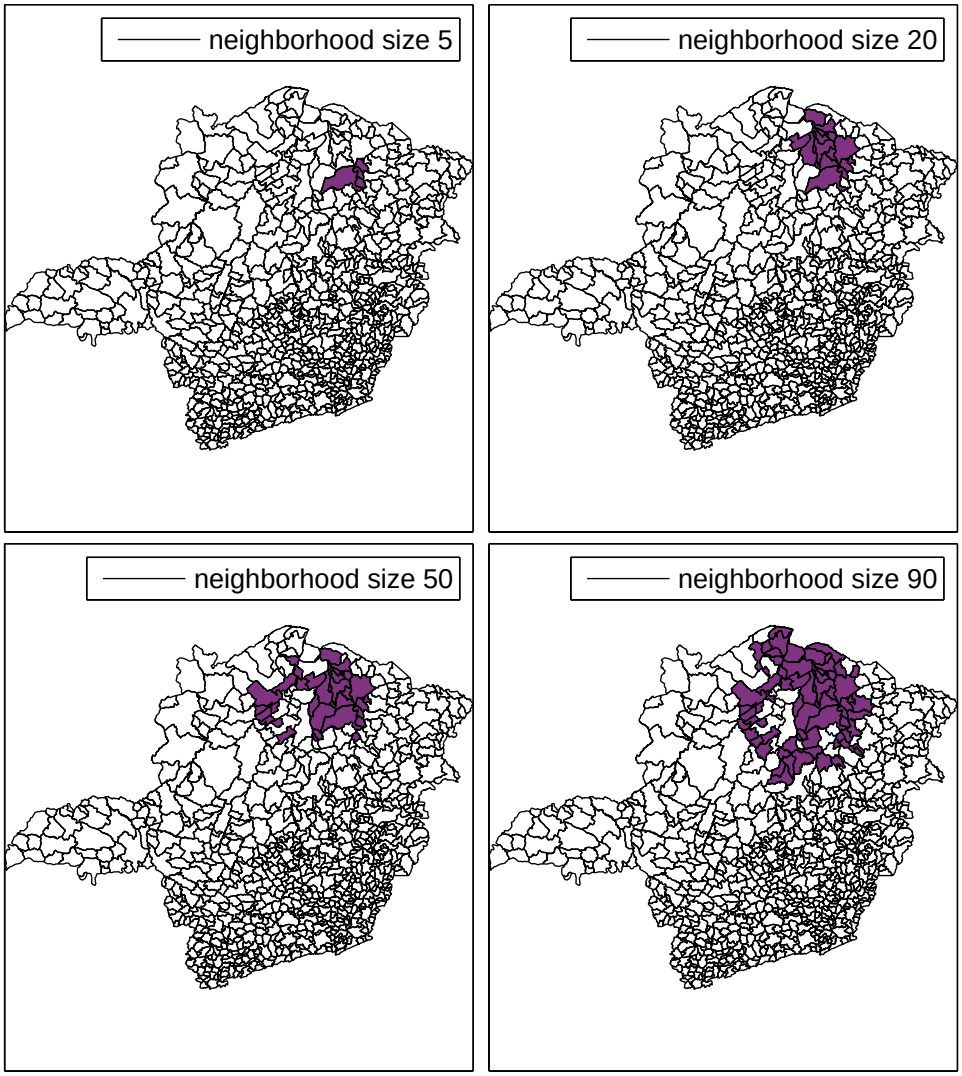

Figure 4.7: Clusters found by dynamic programming method for the data set of Chagas, with neighborhood size $5, 20, 50, 90$.

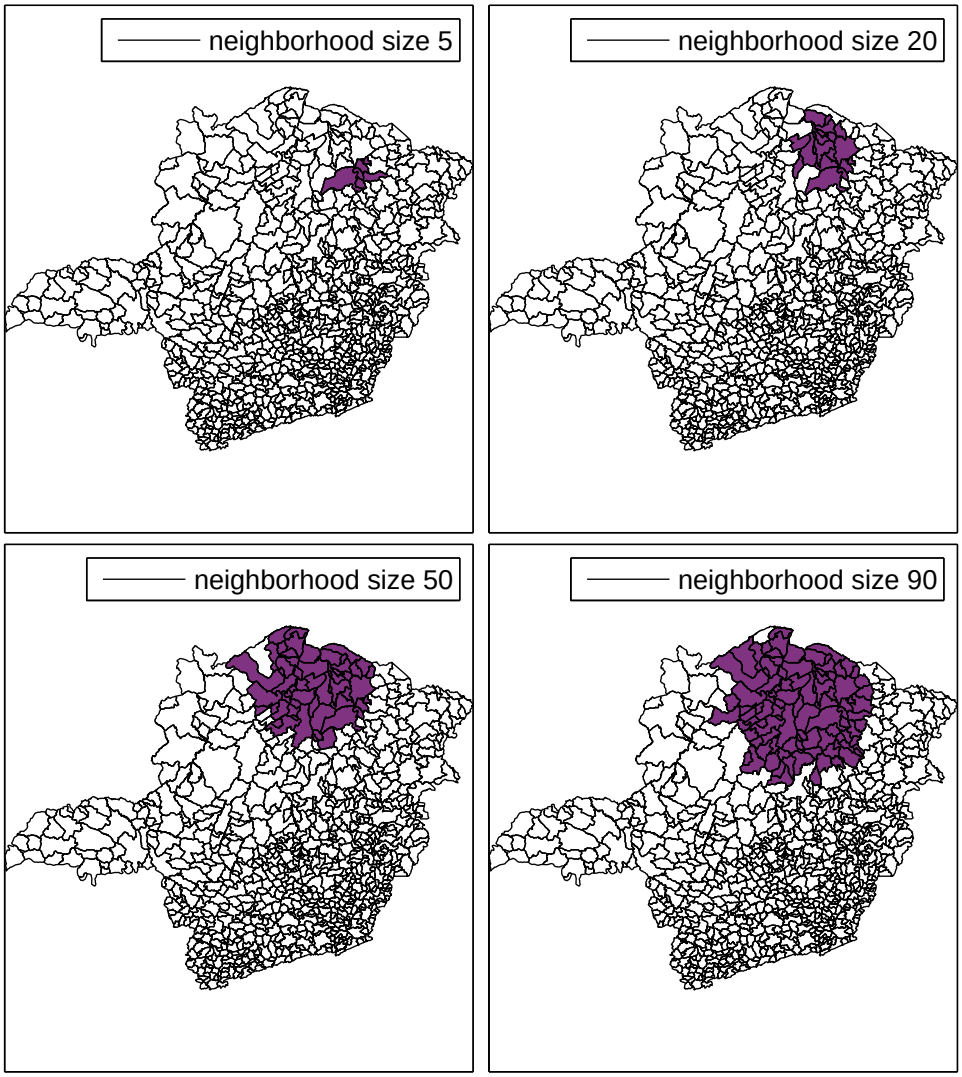

Figure 4.8: Clusters found by classical Kulldorff method for the data set of Chagas, with neighborhood size $5, 20, 50, 90$.

Figure 4.9 shows the Pareto front set obtained by Geographical Dynamic scan with neighborhood size $5, 20, 50, 90$ for the data set of Chagas' disease cases in puerperal women in Minas Gerais state Brazil, consists of $30, 134, 393, 503$ solutions (zones) respectively. Among the solutions (zones) of the Pareto-optimal set, we choose the one that is maximal with respect

to function K. While the Figure 4.10 shows the graphics $C(x) \times N(x)$ of all non-dominated solutions of Problem (3.3) using the dynamic programming algorithm with input data G_i, where G_i is the geographical proximity for each region $i = 1, ..., 853$, respectively for neighborhood size $5, 20, 50, 90$.

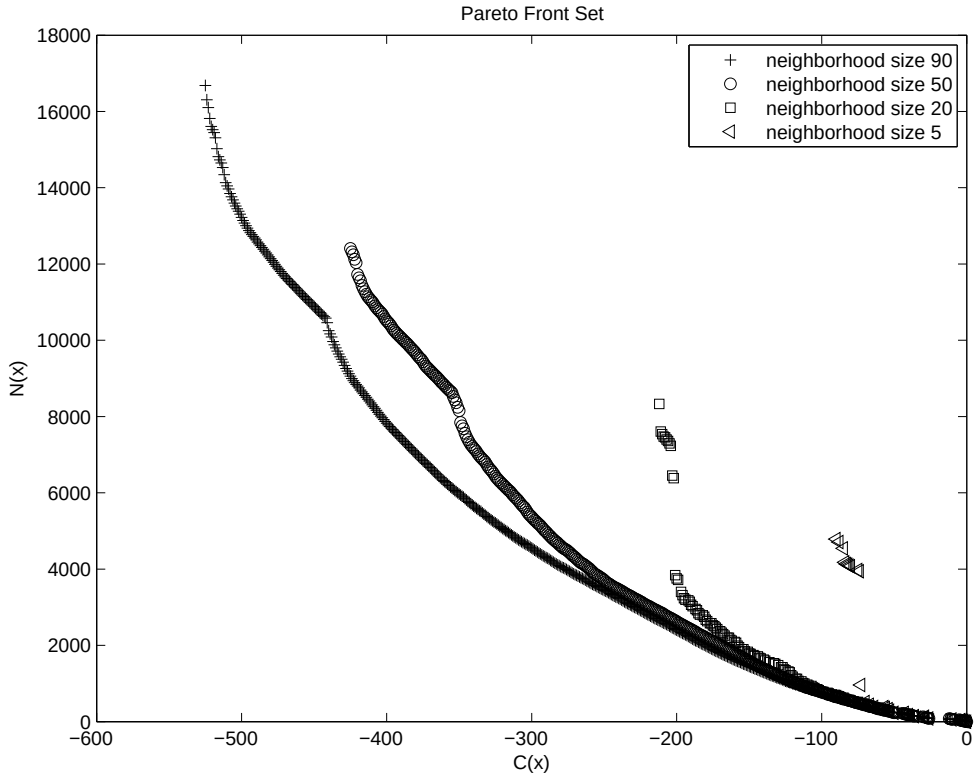

Figure 4.9: Pareto front set obtained by Geographical Dynamic scan, with neighborhood size $5, 20, 50, 90$.

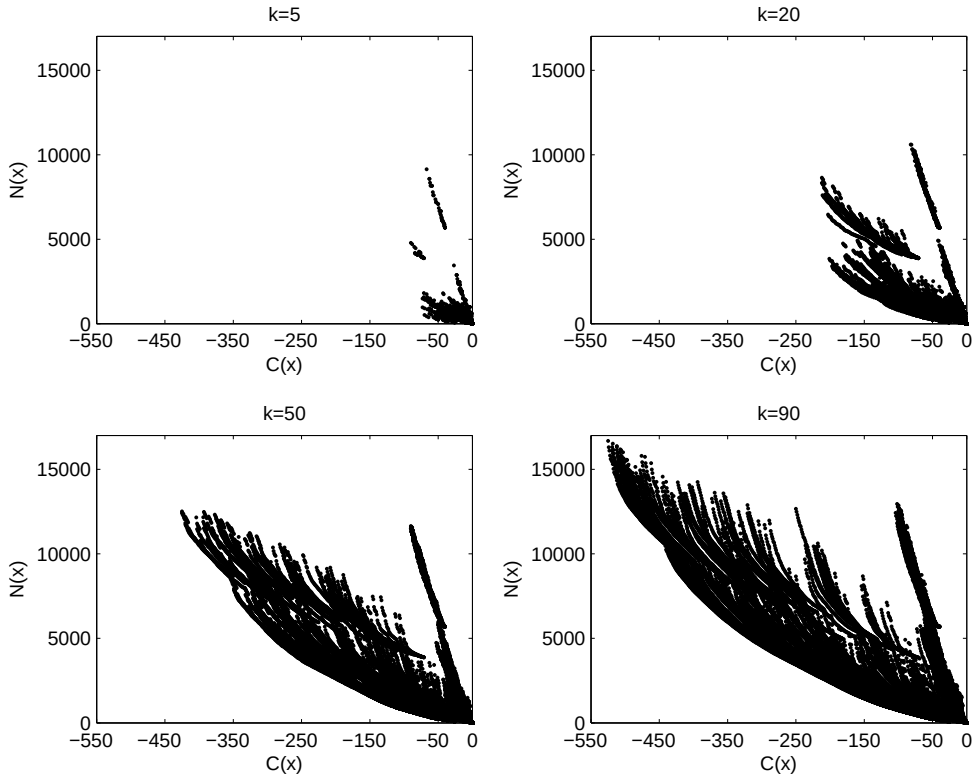

Figure 4.10: All non-dominated solutions considering the geographical proximity for each region $i = 1, ..., 853$, obtained by Geographical Dynamic scan, with neighborhood size $5, 20, 50, 90$.

In addition, a second set of simulations is presented. Next, for each neighborhood size ($k = 10, 20, 30, 40, 50, 60$) fixed, we perform 1,000 null hypothesis Monte Carlo replications to the dynamic programming method. Was computed the number of non-dominated solutions for the 1,000 replications. Table 4.5 and the Figure 4.11 indicates which the number of non-dominated solutions increases linearly with the geometric constraint size k.

In the next step, we evaluate the power, sensitivity and positive predictive value of the geographical dynamic scan algorithm to evaluate the quality of the cluster detection process.

For real data population of the data set of Chagas, three simulated irregularly shaped clusters **A**, **B** and **C** displayed in the Figure 4.12, were used.

Table 4.5: Mean and standard deviation the number of non-dominated solutions in each geometric constraint size k.

# of non-dominated solutions	$k = 10$	$k = 20$	$k = 30$	$k = 40$	$k = 50$	$k = 60$
mean	135.77	176.94	190.13	203.45	217.12	232.72
standard deviation	13.18	11.85	11.07	10.24	9.74	9.63

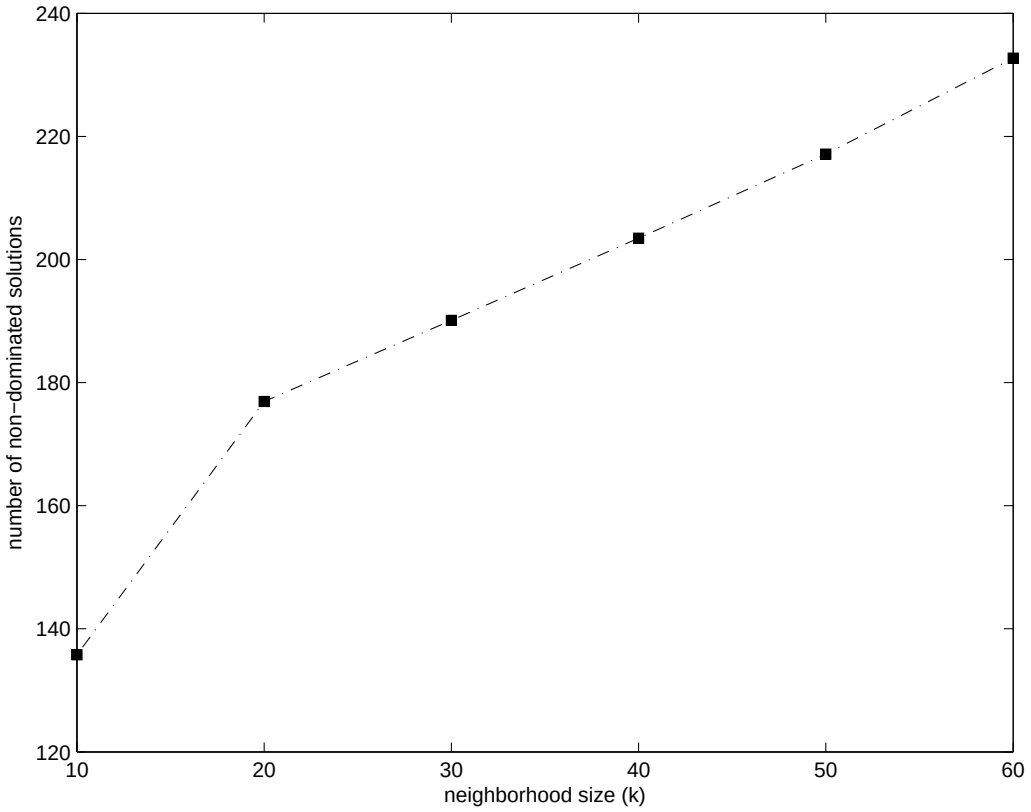

Figure 4.11: Mean of the number of non-dominated solutions versus the geometric constraint size k.

Table 4.6 indicates the number of regions $n(z)$, the number of observed cases z_C and the population z_N for each cluster z. Those clusters will be denoted *real clusters*, in contrast to the *detected clusters* found by the algorithms. For each simulation of data under these three alternative hypotheses, 1,000 cases are randomly distributed according to a Poisson model using a single cluster; we set a relative risk equal to one for every region outside the real cluster and greater than one and identical in each region 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.

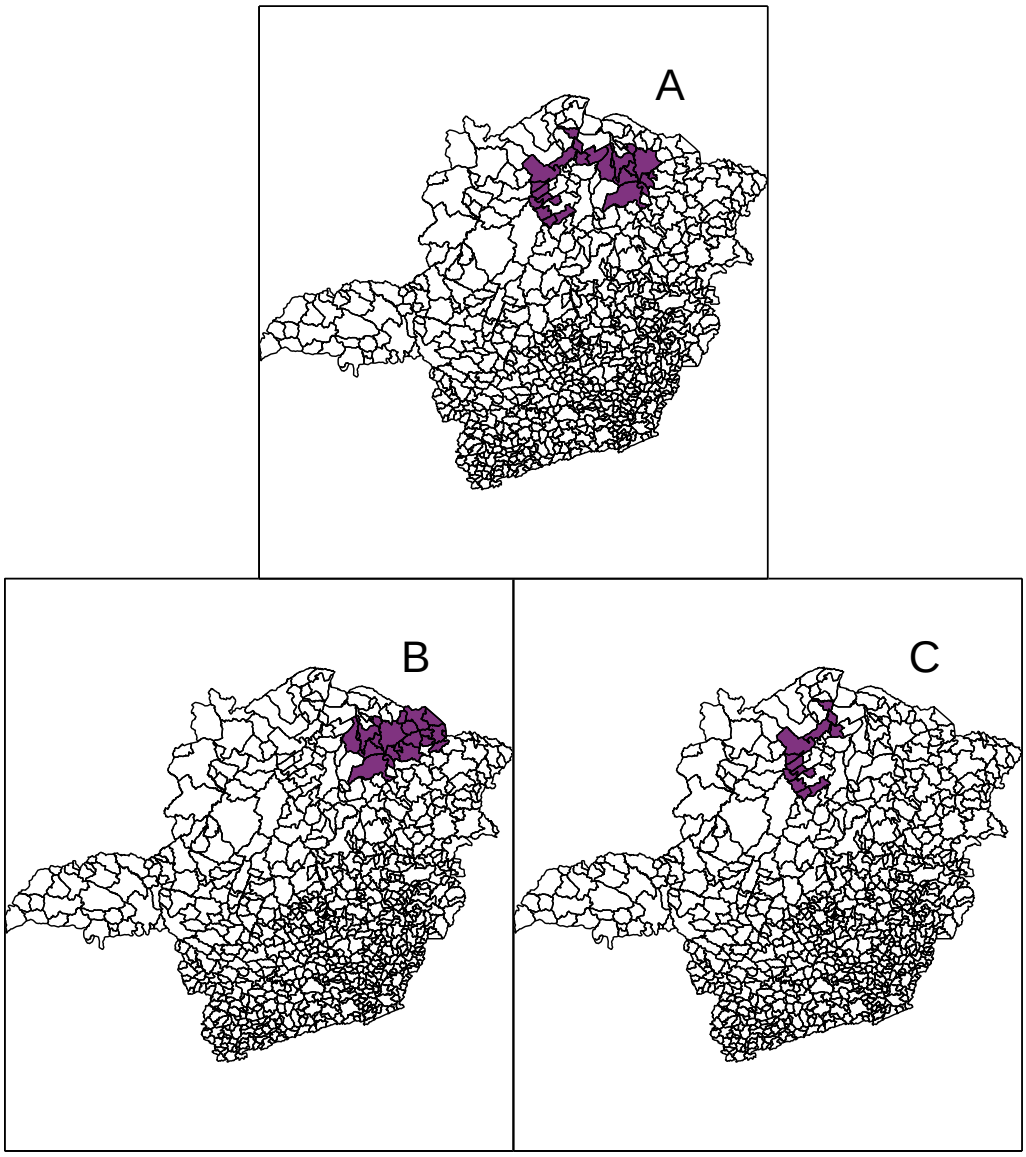

Figure 4.12: Simulated data clusters for data set of Chagas.

Table 4.6: Number of regions $n(z)$, the number of observed cases z_C and the population z_N for the benchmark clusters of Figure 4.12.

Cluster	$n(z)$	z_C	z_N
A	24	226	3938
B	22	156	3566
C	12	79	1661

We perform 1,000 null hypothesis Monte Carlo replications and 1,000 Monte Carlo replications. For each three alternative hypothesis models, for the classical Kulldorff and dynamic programming methods in each neighborhood size ($k = 10, 20, 30, 40, 50, 60$) fixed for the two algorithms. The three measures of power, sensitivity an PPV were computed for the most likely cluster in each replication. The Figures 4.13, 4.14 and 4.15 presents the average power, sensitivity and PPV for the 1,000 replications of each of the three alternative hypotheses clusters **A**, **B** and **C**, for the two algorithms.

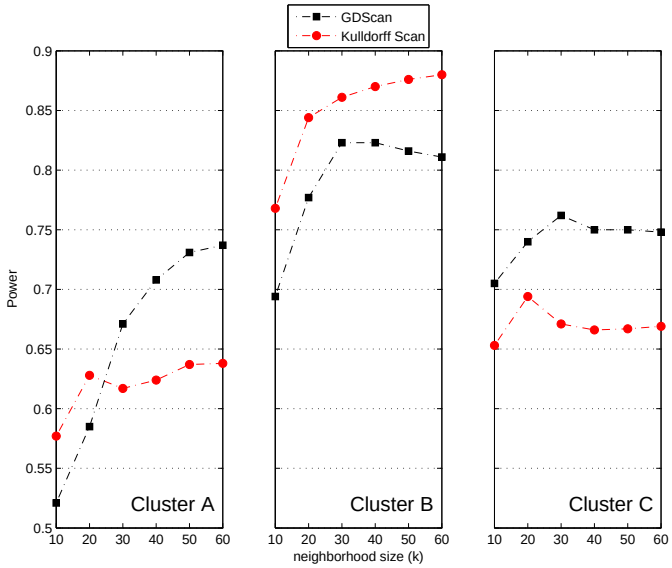

Figure 4.13: Comparison of detection metohds. Average power detection.

In terms of the evaluated measures, the geographical dynamic scan algorithm definitely have a performance superior to the classical Kulldorff method to shaped cluster **C**. This is a good sign, since it shows that the method can detect clusters with irregular geometry. This natural property of the geo-

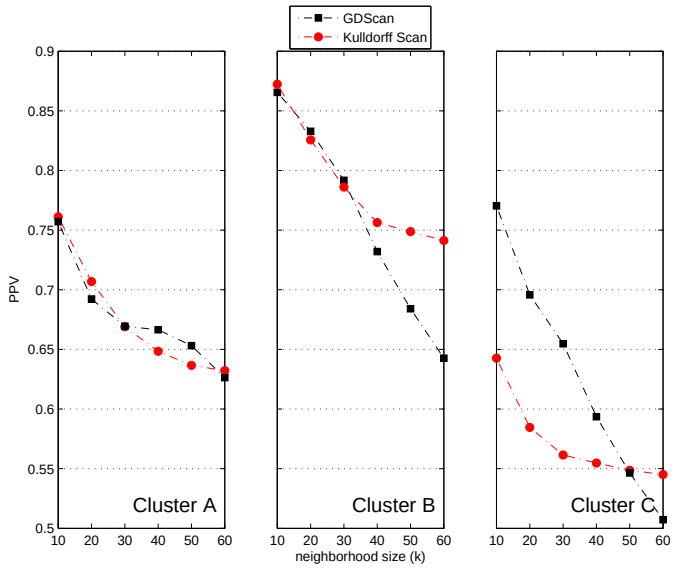

Figure 4.14: Comparison of detection methods. Average positive predicted value.

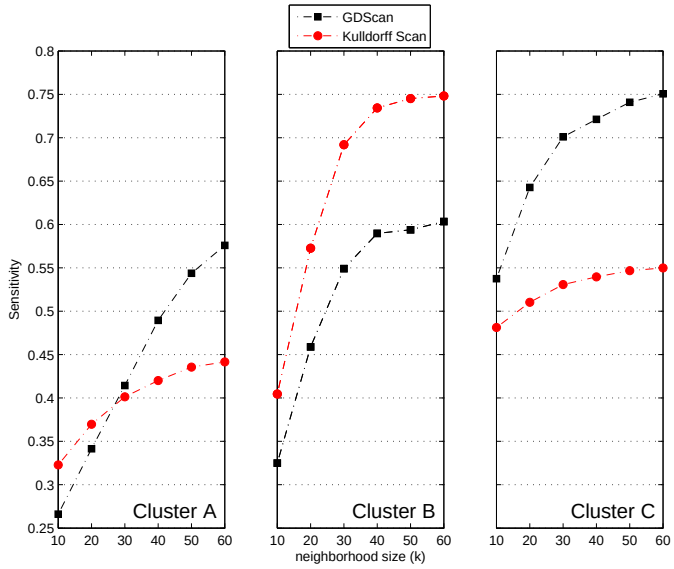

Figure 4.15: Comparison of detection methods. Average sensitivity value.

graphical dynamic scan is also evidenced by the estimated values of PPV, sensitivity and power to the shaped cluster **A**. While, the low value of sensitivity for the shaped cluster **A** in the Kulldorff method, suggests that, on average, underestimates the real cluster, detecting only the circular part of

the cluster. The Kulldorff scan method outperforms the GDScan in terms of power and sensitivity in the form of cluster **B**, but the values of PPV methods were similar.

Chapter 5

Application: A real dataset

We describe an application to cases of dengue fever in the municipality of Lassance in southeast Brazil. We apply the Voronoi Based Scan for the detection of Dengue fever clusters in spatial and space-time coordinates.

5.1 Dengue Fever Clusters

Dengue fever is caused by one of four types of virus, typically transmitted by the mosquito *Aedes aegypti*. Immunity to one strain does not confer lifelong immunity to the other strains. Underreporting is a serious problem with dengue fever data. It is estimated that only 10% of the cases are usually registered at hospitals or health care units (Pessanha, 2010). A pilot project was set in order to obtain more reliable data, with surveillance done at the individual level. Community health agents of the Family Health Program (FHP), (see http://portal.saude.gov.br/portal/saude), performed weekly visits at all residences within the municipality. This already existing program provides guidance for citizens and informs local public health authorities about possible health problems, and is highly regarded in the community. Due to its unique features, the FHP could in principle provide a huge amount of information which would be useful in the surveillance of many diseases, but data almost never is organized beyond local level. In our pilot project, data collected by 13 community health agents in the urban

zone of the municipality of Lassance were compiled by two nurses, and sent for analysis every workweek with the assistance of the Secretary of Health and Epidemiological Surveillance in Lassance. In addition, home location was registered for every resident in the urban part of the city. In the period of six months in 2010, between January $12th$ and June $14th$, a total of 57 cases were reported from a total of 3986 individuals in the population at risk.

The spatial distribution of the observed cases of dengue fever and controls in Lassance City is shown in Figure 5.1. We have included in Figure 5.2 the $\delta(i, j)$ values for the edges of the Voronoi minimum spanning tree along with the drawing of the Voronoi cells in the background (in gray).

Figure 5.1: Spatial distribution of the observed cases of dengue fever (circles) and controls (dots) in Lassance City, southeast Brazil. North is up in the map.

Dengue is not transmitted directly from one person to another. The virus

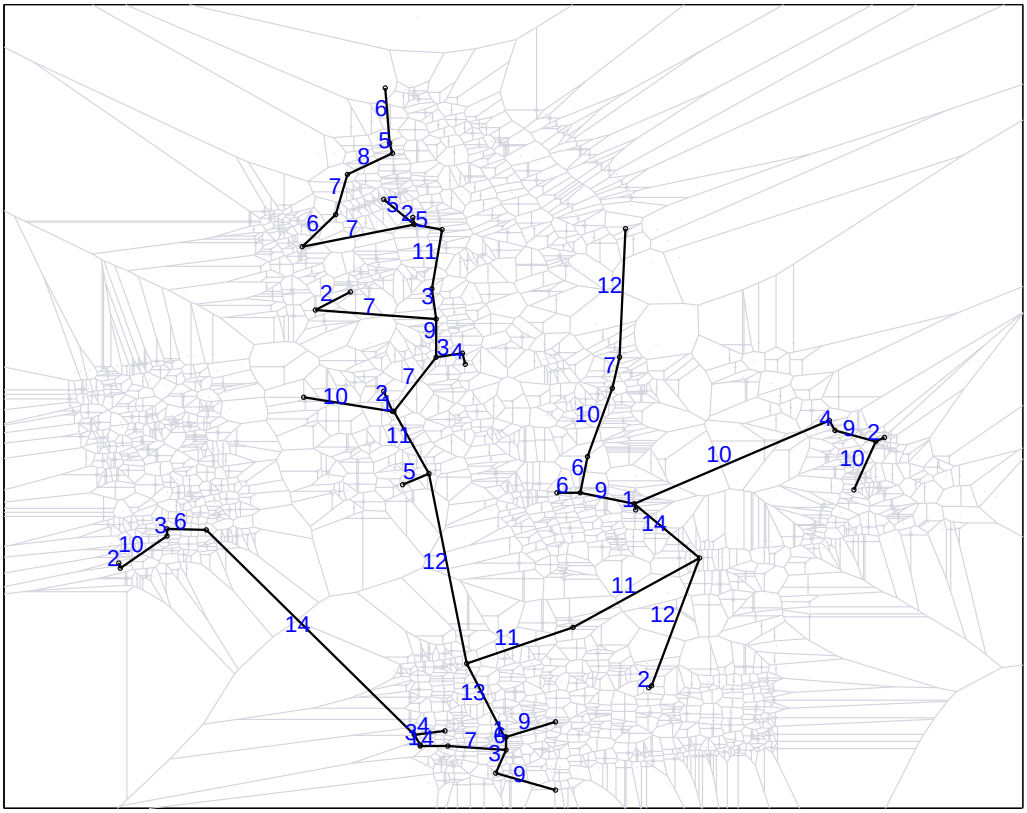

Figure 5.2: Lassance City dengue fever map with assigned weight values for the edges of the Voronoi minimum spanning tree, along with the drawing of the Voronoi cells in the background (in gray).

is transmitted to the mosquito *A. aegypti* after biting an infected individual. The mosquito can carry the virus for 10 to 14 days. In humans, the virus remains in an incubation period that may last from 3 to 15 days. Only after this period the symptoms can be observed. In this way, the study period was divided into 11 intervals of 14 days, as shown in Table 5.1.

5.1.1 Spatial analysis

We relied upon ordinary topographic maps and aerial images provided by Lassance's City Hall, because high resolution Google Earth images were not available (Chang *et al.*, 2009). Those aerial images were manually matched with the existing topographic maps. Data are plotted in the map according

Table 5.1: Study time period subdivided. Each unit represents a period of 14 days.

Time	days observed	cases
1	01-12 to 01-25	03
2	01-26 to 02-08	06
3	02-09 to 02-22	02
4	02-23 to 03-08	07
5	03-09 to 03-22	05
6	03-23 to 04-05	09
7	04-06 to 04-19	04
8	04-20 to 05-03	09
9	05-04 to 05-17	09
10	05-18 to 05-31	02
11	06-01 to 06-14	01

Table 5.2: Match values for spatial clusters Dengue fever data set by using VBScan method

Clusters	LLR	cases	p-value
primary	17.5686	10	0.004
secondary	15.2390	09	0.016

to the exact location of each individual of the population at risk. Data are available as supplementary files. To detect possible clusters, the VBScan method was applied.

The two most likely clusters presented 10 and 9 cases, respectively for the primary and secondary clusters, as shown in Figure 5.3 For the primary cluster a p-value = 0.004 was found, see Table 5.2. Table 5.2 shows that the secondary cluster is also statistically significant. Those p-values are computed from 999 Monte Carlo simulations under the null hypothesis. Hence, we conclude that there is evidence of a geographically significant high risk of dengue fever in some specific regions within the urban area of Lassance City.

Employing the elliptic scan, also with 999 Monte Carlo simulations, the most likely cluster found has only 3 cases, contained within the primary cluster found by VBScan, as marked in Figure 5.3 (p-value= 0.054). The run time for 999 Monte Carlo replications for the Dengue fever cluster was about 187 seconds for the VBScan and 764 seconds for the elliptic scan. This

interesting result arises due to the peculiar features of this problem:

- The population does not follow a random-like spatial distribution; instead, the individuals are roughly aligned according the housing geometry of the streets.
- The neighborhood structure induced by the Euclidean metric, which is used by elliptic scan, becomes very different from the neighborhood structure induced by the Voronoi distance.

Specifically, the population densities, which are considered in the computation of both the scan statistics, are distinct, because the Voronoi distance is calculated along the edges that link the case points, while the density in the

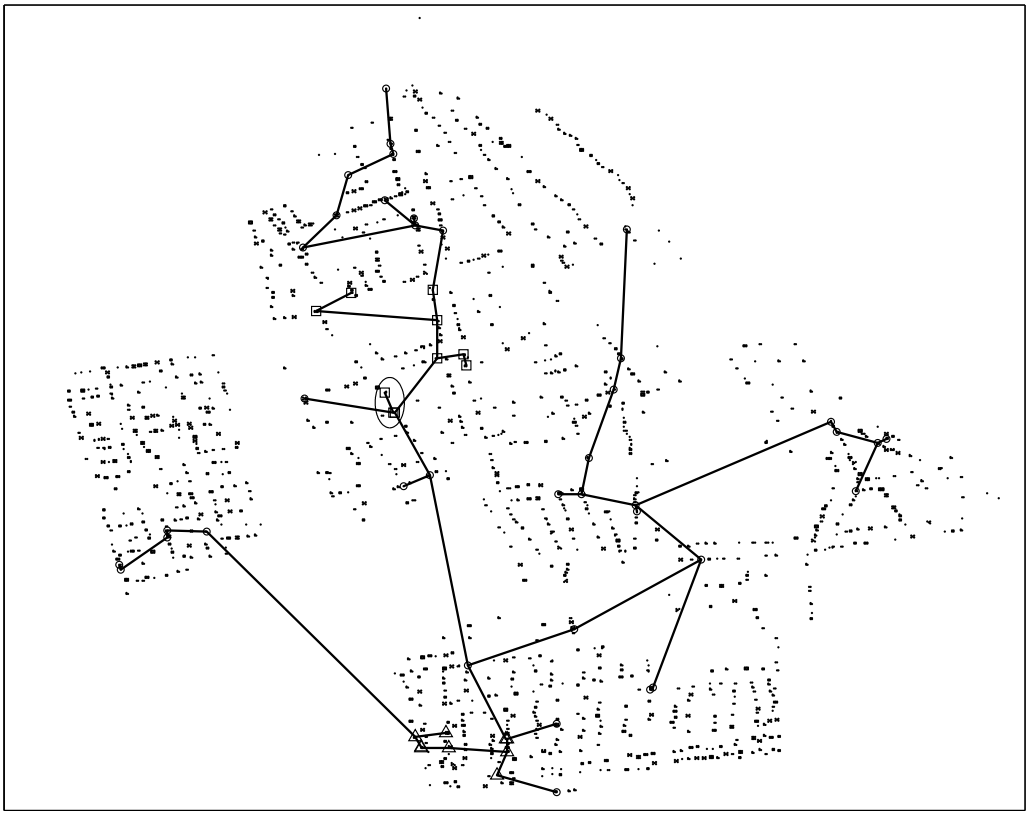

Figure 5.3: Purely spatial primary (squares) and secondary (triangles) dengue fever clusters found by the VBScan, and the primary cluster (within the ellipse) found by the Elliptic Scan.

elliptic scan considers all individuals inside the ellipses. Clearly, this pattern of population spread causes the elliptic scan to consider a greater number of non-infected control cases inside a potential cluster than the VBScan, reducing the power of the Elliptic Scan. It can be noticed, in the primary cluster found by VBScan, that a path used by this algorithm to link a set of cases may avoid the directions in which a large number of non-infected individuals are located. This is due to the definition of Voronoi distance, which exactly assigns larger distances to such paths. The clusters, therefore, may include larger edges (in terms of Euclidean metric) which cross less crowded regions – these are the smaller edges in Voronoi distance – causing the opposite effect in the VBScan detection power.

The primary cluster (indicated by square points in Figure 5.3) has two edges crossing city blocks diagonally, both with assigned value $\delta(i, j) = 7$, as can be seen in Figure 5.2. The longest (in terms of Euclidean distance) edge that links the two northwestern cases crosses a moderately high populated region, as measured by the Voronoi distance, is not an artifact. Although the interior part of the block crossed by this edge has no control individuals, there are many individuals living in its borders, implying that there are several Voronoi cells (bounded by gray lines in the background) inside the block, which in turn makes the diagonally crossing edge intercept several cells in its path. This is a fine example of how the Voronoi distance measures adequately the population density, as a composition of the individual cells (regions of influence) intercepted by the edge's path.

5.1.2 Detecting space-time clusters

The prospective space-time geographical surveillance system proposed here was applied for the detection of dengue fever space-time clusters over the same data set. The time window has a range of $[1, 11]$, in which each unit represents a period of 14 days, as set out in Table 5.1. The results are given in Table 5.3, whose first column indicates the temporal restriction for the construction phase of the minimum spanning tree, influencing the significance of the cluster detection.

Table 5.3: Match values for space-time clusters Dengue fever data set analyzing the periods 1-11, by using VBScan method.

temporal length edge τ	cases	onset time of the disease for the cases	LLR	p-value
1	06	{7,8}	17.3207	0.003
2	07	{5,7,8}	15.0091	0.008
4	06	{7,8}	15.3053	0.019
6	10	{1,2,4,6,8,9}	15.7764	0.024
8	10	{1,2,4,6,8,9}	15.7764	0.024

Table 5.3 shows that all clusters that were found are statistically significant for the time period [01-12 to 06-14]. Again, 999 Monte Carlo simulations were generated under null hypothesis. The two space-time clusters with smaller p-values are part of the secondary spatial cluster, as shown in Figure 5.4 and the values indicated by lines 1 and 2 respectively in Table 5.3.

The cluster that was found as the primary cluster in the purely spatial analysis does not appear as a cluster in the space-time analysis. In the first situation, the cases were spread along the time axis. On the other hand, only a few cases were included in the same cluster, when time is considered. This pattern suggests that, instead of a single space-time cluster of dengue fever, there was a series of several independent re-infections of individuals within the space region of that cluster. This interpretation is consistent with an environmental information: that region belongs to the central part of the municipality, where several public service facilities are located. This means that such a region receives a flow of people from all other regions, which is consistent with the hypothesis of several re-incidences of dengue fever cases in that region in events which are not directly dependent.

On the other hand, the cluster that was found as the secondary cluster in the purely spatial analysis appears as the single detected cluster in the space-time analysis. In this cluster, most of the cases occurred within a small temporal window. Located in a poorer part of the municipality, at the border of the urban area, this region has several environmental factors favoring a large concentration of mosquito larvae, such as deficient sewage installations and garbage collection, accumulated water puddles, and the

Figure 5.4: Space-time clusters of the dengue fever dataset, with temporal constraint parameter values $\tau = 1$(crosses) and $\tau = 2$(squares), matching the values shown in lines 1 and 2 of Table 5.3, respectively.

presence of many vacant lots and houses. Furthermore, the timing of the cluster coincides with the rainiest weeks of 2010. These data are consistent with the hypothesis of a single event epidemics outbreak, with a direct causal correlation between the several cases.

Chapter 6

Conclusions

This thesis addresses the spatial and space-time cluster detection problem. Two algorithms to solve two typical problem for spatial data sets are proposed.

6.1 Summary

We developed and tested a novel algorithm for the detection and inference of space-time clusters for data sets, the Voronoi Based Scan (VBScan). The concept of Minimum Spanning Tree (MST) is adapted with the novel Voronoi distance, which is used to compute the set of potential clusters. This set is then evaluated using the spatial scan statistic, producing the most likely cluster of cases.

The class of problems considered here assumes a point data set to represent the location of individuals in a population, classified either as controls or disease cases, within a limited domain in space-time. The cluster is modeled in space coordinates as a connected graph with tree structure, joining a subset of the disease cases, and in space-time coordinates as a sequence of such trees with space projections that have non-null intersection. A distance measure, named Voronoi distance, is proposed here in order to define a meaningful distance for the construction of a minimum spanning tree (MST) that represents the more likely connections between individuals, in a given graph.

This structure allows the direct application of the scan statistics, with the calculation of the likelihood ratio of the estimated cluster.

The Voronoi distance between any two points may also be interpreted as an approximation to the line integral of the population density function over the segment joining those two points. For this reason, the VMST is the natural extension of the Euclidean MST, taking into account the heterogeneity of the population density. On the other hand, the Euclidean distance is an approximation to the corresponding line integral only when the map is cartogram transformed, in such a way that the population density becomes homogeneous. The Voronoi distance concept is employed once again in our method, after the collection of potential clusters is extracted from the VMST: it is used to estimate the number of control individuals under the region of influence of each one of the case individuals. This allows the definition of the population associated to each potential cluster, which may be evaluated through the spatial scan statistic.

The results of numerical simulations show that the proposed algorithm, space-time VBScan, has higher power of detection, positive predictive value, sensitivity and computational speed than the space-time Elliptic Scan. The flexibility verified of VBScan allows an enhanced ability to deal with the variation of the disease spread along the time dimension.

An application was presented for Dengue fever incidence, with data available at individual level, in the municipality of Lassance, Brazil.

VBScan also includes topological information from the point neighborhood structure, in addition to the usual geometric information. For this reason, it is more robust than purely geometric methods such as the elliptic scan. Those advantages were illustrated in a real setting for dengue fever space-time clusters, where the population spreads along a grid of straight lines according to the street mapping. It is worthy to notice that this kind of geometry of population distribution appears very often in urban environments. In those cases, the employment of VBScan should be recommended.

In the examples that we have analyzed, we observed that the Voronoi distance is very reliable to approximate the population heterogeneity, even for some unusual population distribution patterns, like a city block with zero

individuals living in its interior and many individuals living on its borders.

The ability for the early detection of space-time clusters of disease outbreaks, when the number of points in the dataset is large, was shown to be feasible, due to the reduced computational load of the proposed methodology compared with classical methods. The proposed methodology is shown to present an enhanced power for the detection of space-time disease clusters.

The second proposed algorithm, Geographical Dynamic Scan, was proved to be an efficient method to solve practical spatial cluster detection problem. In particular, we made use of the of the property of the logarithm of the likelihood ratio K and restate the classical spatial cluster detection problem as a bi-objective combinatorial optimization problem, in the same way as done in (Cancado, 2009). In addition, we established a correspondence between the set of non-dominated solutions of the bi-objective combinatorial optimization problem and the solution that maximizes K. We demonstrate that the dynamic programming algorithm used to solve the problem enable efficient unconstrained maximization of the function K for spatial dataset.

The detected clusters may sometimes be disconnected, but this is not a serious disadvantage, provided that there is not a huge gap between its areas. A way to control the presence of those potential gaps is to limit the number of component areas of the cluster, e.g., allowing only clusters which are subsets of a circular zone of moderate maximum size. Even when considering geographic diffusion processes, disconnected clusters may be detected due to the stochastic nature of the process, e.g., when the number of disease cases is small. In this sense, disconnected clusters may be allowed. Disconnected clusters also occur in other cluster detection methods, as in the elliptic scan, when the ellipse is very elongated (Kulldorff *et al.*, 2006).

6.2 Publications

In the following we present the list of the publications related to the theme which were generated based on the results that we have obtained during this thesis development:

- Moreira, G. J. P.; Paquete, Luís; Duczmal, L. H.; Takahashi, R. H. C. Spatial cluster detection by dynamic programming. (pre-print)

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

- Duczmal, L. H.; Moreira, G. J. P.; Burgarelli, D.; Takahashi, R. H. C.; Magalhães, F. C. O.; Bodevan, E. C. A Voronoi based scan for space-time cluster detection in point event data. In: 9th Annual Conference of the International Society for Disease Surveillance (ISDS 2010). Emerging Health Threats Journal, 2011, pp. 18-19.

- Duczmal, L. H.; Magalhães, F. C. O.; Ferreira Neto, S. J.; Moreira, G. J. P.; Duarte, A. R.; Cancado, A. L. F.; Burgarelli, D. Syndromic Surveillance of Dengue Fever in Brazil at the Individual Level. In: Eighth Annual International Society for Disease Surveillance (ISDS) Conference, 2009.

- Moreira, G. J. P.; Takahashi, R. H. C.; Duczmal, L. H. Delineating Spatial Clusters with Artificial Neural Networks. In: Sixth Annual International Society for Disease Surveillance Conference, Advances in Disease Surveillance 2007, 4(3):104.

- Duczmal, L. H.; Moreira, G. J. P.; Ferreira Neto, S. J.; Takahashi, R. H. C. Dual Graph Spatial Cluster Detection for Syndromic Surveillance in Networks. In: Sixth Annual International Society for Disease Surveillance Conference, Advances in Disease Surveillance 2007, 4(3):88.

6.3 Future Work

Following the investigations described in this thesis, a number of projects could be taken up. Some of them are:

- Derive a spatial cluster detection method for areal data, using the concept applied in the VBScan method.

- One potential limitation of our analysis is the spatial mobility of individuals from their residences to workplace, which could impair the geographic delineation of the detected clusters. In a future work we will address this issue, using tools such as the work-flow scan statistic (Duczmal & Buckeridge, 2006).

- Because we make use of an already existing team of community health agents, originally employed for health monitoring in general, Dengue fever surveillance is very cost effective in our setting, and we can focus our effort on mapping, data collection, data integrity issues and analysis. In a future work, we will use additional zoonosis and environmental data, and apply covariate analysis. This will allow better monitoring and forecasting of outbreaks.

Bibliography

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

Cancado, A. L. 2009. *Spatial clusters detection through multiobjective optimization. (in portuguese).* Ph.D. thesis, Universidade Federal de Minas Gerais.

Cançado, ALF, Duarte, AR, Duczmal, LH, Ferreira, SJ, Fonseca, CM, & Gontijo, ECDM. 2010. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International journal of health geographics*, **9**, 1–17.

Chang, A Y, Parrales, M E, Jimenez, J, Sobieszczyk, M E, Hammer, S M, Copenhaver, D J, & Kulkarni, R P. 2009. Combining google earth and gis mapping technologies in a dengue surveillance system for developing countries. *International journal of health geographics*, **8**, 49.

Conley, J, Gahegan, M, & Macgill, J. 2005. A genetic approach to detecting clusters in point-data sets. *Geographical analysis*, **37**, 286–314.

Cucala, L. 2009. A flexible spatial scan test for case event data. *Comput. stat. data anal.*, **53**, 2843–2850.

Demattei, C, & Cucala, L. 2011. Multiple spatio-temporal cluster detection for case event data: an ordering-based approach. *Communications in statistics-theory and methods*, **40**(2), 358–372.

Demattei, C, Molinari, N, & Daurès, J. 2007. Arbitrarily shaped multiple spatial cluster detection for case event data. *Comput. stat. data anal.*, **51**, 3931–3945.

Duarte, AR, Cançado, ALF, Duczmal, LH, & Ferreira, SJ. 2010. Internal cohesion and geometric shape of spatial clusters. *Environmental and ecological statistics*, **17**, 203–229.

Duczmal, L H, & Assunção, R. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational statistics & data analysis*, **45**(2), 269–286.

Duczmal, L H, & Buckeridge, D L. 2006. A workflow spatial scan statistic. *Statistics in medicine*, **25**, 743–754.

Duczmal, L H, Kulldorff, M, & Huang, L. 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of computational and graphical statistics*, **15**(2), 428–442.

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

Duczmal, L H, Cançado, A L F, & Takahashi, R H C. 2008. Geographic delineation of disease clusters through multi-objective optimization. *Journal of computational & graphical statistics*, **17**, 243–262.

Duczmal, L H, Duarte, A R, & Tavares, R. 2009. Extensions of the scan statistic for the detection and inference of spatial clusters. *Pages 153–177 of:* Glaz, Joseph, Pozdnyakov, Vladimir, & Wallenstein, Sylvan (eds), *Scan statistics*. Statistics for Industry and Technology. Birkhäuser Boston.

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

Dwass, M. 1957. Modified randomization tests for nonparametric hypotheses. *Ann. math. statist.*, **28**, 181–187.

Gupta, P, Agarwal, V, & Varshney, M. 2008. *Design and analysis of algorithms*. PHI Learning Private Limited.

Iyengar, V S. 2005. Space-time clusters with flexible shapes. *Mmwr morb mortal wkly rep*, **Suppl 54**, 71–76.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in statistics: Theory and methods*, **26**, 1481–1496.

Kulldorff, M. 1999. Spatial scan statistics: Models, calculations and applications. *Pages 303–322 of:* Glaz, & Balakrishnan (eds), *Scan statistics and applications*. Boston: Birkhauser.

Kulldorff, M. 2001. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the royal statistical society series a*, **164**(1), 61–72.

Kulldorff, M, & Nagarwalla, N. 1995. Spatial disease clusters: Detection and inference. *Statistics in medicine*, **14**(8), 799–810.

Kulldorff, M, Tango, T, & Park, P J. 2003. Power comparisons for disease clustering tests. *Computational statistics & data analysis*, **42**(4), 665–684.

Kulldorff, M, Heffernan, R, Hartman, J, Assunção, R, & Mostashari, F. 2005. A space time permutation scan statistic for disease outbreak detection. *Plos med*, **2**(3).

Kulldorff, M, Huang, M, Pickle, L, & Duczmal, L H. 2006. An elliptic scan statistic. *Statistics in medicine*, **25**(22), 3929–3943.

Kulldorff, M, Mostashari, F, Duczmal, L H, Katherine, Y W, Kleinman, K, & Platt, R. 2007. Multivariate scan statistics for disease surveillance. *Statistics in medicine*, **26**(8), 1824–1833.

Lawson, A, Biggeri, A, & Böhning, D. 1999. *Disease mapping and risk assessment for public health*. New York: John Wiley and Sons.

Naus, J. 1965. The distribution of the size of maximum cluster of points on the line. *Journal of the american statistical association*, **60**, 532–538.

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

Neill, D B. 2009. An empirical comparison of spatial scan statistics for outbreak detection. *International journal of health geographics*, **8**, 20.

Neill, D B. 2010. Fast subset sums for multivariate bayesian scan statistics [abstract]. *Proceedings of the 2009 international society for disease surveillance annual conference*.

Nemhauser, G. L., & Ullmann, Z. 1969. Discrete dynamic programming and capital allocation. *Management science*, **15**(9), 494–505.

Oliveira, Fernando, Duczmal, L H, Cancado, A L F, & Tavares, Ricardo. 2011. Nonparametric intensity bounds for the delineation of spatial clusters. *International journal of health geographics*, **10**(1), 1.

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

Pessanha, J E M. 2010. *Dengue in belo horizonte: a population-based seroepidemiological survey (2006-2007), study of virus vectors (2007). evaluation of the national dengue control plan(2008). (in portuguese).* Ph.D. thesis, Universidade Federal de Minas Gerais.

Robertson, C, & Nelson, TA. 2010. Review of software for space-time disease surveillance. *International journal of health geographics*, **9**, 1–8.

Sahajpal, R, Ramaraju, GV, & Bhatt, V. 2004. Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. *Int. conf. intelligent sensing and information processing.*

Takahashi, K, Kulldorff, M, Tango, T, & Yin, K. 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International journal of health geographics*, **7**(1), 14.

Tango, T, & Takahashi, K. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, **4**(1), 11.

Wieland, S C, Brownstein, J S, Berger, B, & Mandl, K D. 2007. Density-equalizing euclidean minimum spanning trees for the detection of all disease cluster shapes. *Proceedings of the national academy of sciences*, **104**(22), 9404–9409.

Xu, Y, Olman, V, & Xu, D. 2002. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, **18**(4), 536–545.

Yiannakoulias, N, Rosychuk, RJ, & Hodgson, J. 2007. Adaptations for finding irregularly shaped disease clusters. *International journal of health geographics*, **6**(28).