Thaís Viana Paiva

Vigilância espaço-temporal com Superfícies Acumuladas

Belo Horizonte, agosto de 2010

Thaís Viana Paiva

Vigilância espaço-temporal com Superfícies Acumuladas

Dissertação apresentada como requisito parcial para obtenção de grau de Mestre em Estatística pela Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Renato Martins Assunção

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

Belo Horizonte, agosto de 2010

Agradecimentos

Agradeço primeiramente a Deus por ter me iluminado durante toda a minha caminhada até essa conquista.

Aos meus pais Alcy e Beatriz, pelo amor e apoio e por serem meus exemplos a serem seguidos. Devo tudo isso à dedicação de vocês. Ao meu irmão Leandro, pelos momentos de descontração e alegria. Aos meus avós tão carinhosos, pelos conselhos sábios e experientes. Ao Galletti, pelo amor, compreensão, paciência e tranquilidade.

Ao meu orientador, Professor Renato Assunção, pelas oportunidades, ensinamentos e explicações durante esses quatro anos, e também por ser o principal motivador para definir o meu destino nos próximos quatro anos.

Aos meus professores no curso de Mestrado em Estatística, pelo conhecimento e convivência durante esse período. Aos funcionários da secretaria de Estatística da UFMG e à FAPEMIG pelo apoio financeiro.

Aos amigos e amigas do Leste, em especial a Érica, Aline, Márcia, Letícia, Gabi e Alessandra, por tanta ajuda e apoio e por escutarem todas as minhas reclamações e histórias. Às amigas do bonde, Ju, Naty, Bárbara, Joana, Suelen e Taci, e às amigas eternas, Amanda, Amanda, Iara e Ana Luisa, pelos momentos de diversão e conversas intermináveis. A todos os amigos e familiares que de alguma maneira contribuiram para a realização desse trabalho. Muito Obrigada!

Resumo

O mapeamento de crimes e doenças fornece informações sobre o padrão espacial e temporal de ocorrência dos eventos. É de interesse o monitoramento dos eventos para a detecção precoce de mudanças no seu padrão espacial. Esta dissertação apresenta um método de vigilância espaço-temporal prospectiva de dados pontuais, verificando se há um cluster emergente. A cada novo evento, o escore local de Knox é calculado e suavizado de maneira a formar uma superfície estocástica. Essas superfícies são então acumuladas sequencialmente até que ultrapassem um limiar estabelecido, quando o alarme soa, identificando a região do provável cluster. As vantagens estão em exigir pouca informação prévia do usuário e em fornecer uma maneira de identificar a localização de possíveis clusters, através da visualização da superfície acumulada. A performance do método foi avaliada a partir de resultados de simulações em diferentes cenários. O método foi aplicado a um conjunto de dados de casos de meningite em Belo Horizonte.

Palavras-chaves: *Sistema de Vigilância, Dados Pontuais, Espaço-Temporal, Escore de Knox Local, Superfícies Acumuladas*.

Abstract

Mapping crime and disease events provides information about the spatial and the temporal pattern of these events. To guide actions based on such information, it is necessary to use a statistical method to identify when there is a change in the pattern of events. We developed a space-time prospective surveillance method when the data are point events, monitoring if there is an emerging cluster. At each new event, a local Knox score is calculated and spatially spread to form a stochastic surface. The surfaces are accumulated sequentially until it overcomes a specified threshold, when an alarm goes off, identifying the region of the probable cluster. The advantages are to require little prior knowledge from the user and to provide a way to identify the locations of the possible clusters, through the visualization of the cumulative surface. A simulation study is presented for different scenarios and a dataset of meningitis cases in Belo Horizonte was monitored.
Keywords: *Surveillance, Point Pattern, Space-Time, Local Knox Score, Cumulative Surfaces.*

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A consideração simultânea dos padrões espaciais e temporais da ocorrência dos eventos é importante para identificar clusters ou conglomerados espaço-temporais. Definimos um cluster espaço-temporal como uma região geograficamente pequena em relação à região em estudo e que concentra um número excessivo de eventos durante um período limitado de tempo.

Os métodos de detecção de clusters espaço-temporais são, em sua maioria, retrospectivos. Esses métodos procuram por evidências da presença de um cluster em um banco de dados de eventos já ocorridos. O teste de detecção de conglomerados espaço-temporais mais popular foi desenvolvido por Knox (1964), e ele testa se há interação espaço-temporal. Ou seja, ele testa se casos próximos no espaço tendem a estar próximos no tempo também. Mantel (1967) estendeu o teste de Knox, usando as distâncias espaciais e temporais entre os pares de eventos ao invés dos indicadores binários de proximidade. Jacquez (1996) propôs um teste para interação espaço-temporal de k vizinhos mais próximos e compara os resultados com os testes de Knox e Mantel, mostrando que o seu teste tem maior poder que os outros. Esses três testes podem ser viciados se a população de risco subjacente muda diferencialmente. Kulldorff (1999) propôs uma modificação para solucionar esse problema, mas ela exige informação da população sob risco.

Ao analisar eventos como crimes e doenças, um objetivo importante é detectar um cluster emergente, através da vigilância prospectiva. O desafio é desenvolver métodos de vigilância eficientes e que detectem rapidamente os clusters, além de minimizar o número de alarmes falsos. O interesse é focado em ações que possam mitigar os efeitos dos clusters se realizadas rapidamente.

Recentemente, pode-se observar um grande desenvolvimento de métodos prospectivos puramente temporais, especialmente na área epidemiológica. Os métodos mais usados para fazer vigilância prospectiva epidemiológica foram resumidos e avaliados em Sonesson and Bock (2003).

Höhle (2007) desenvolveu um pacote para o software R para fazer a vigilância prospectiva para dados epidemiológicos.

Poucos métodos para vigilância prospectiva espaço-temporal foram desenvolvidos até o momento. Um desses métodos foi apresentado por Kulldorff (2001) e calcula uma estatística de scan espaço-temporal contando o número de eventos dentro de um cilindro de raio e altura pré-estabelecidos. Neill et al. (2005) também desenvolveram um método de detecção de clusters emergentes usando a estatística de scan espaço-temporal. Uma extensão do método de scan foi proposto por Kulldorff et al. (2005), que não precisa de informações sobre a população em risco. Estes métodos são baseados em idéias de testes de hipóteses, que não são adequados no contexto de vigilância prospectiva (Woodall et al., 2008). Rodeiro and Lawson (2006) utilizam modelos Bayesianos espaço-temporais para detectar um aumento do risco em mapas de doenças com dados de áreas. Diggle et al. (2005) também apresentam um modelo de vigilância para processos pontuais, monitorando se probabilidades preditivas, determinadas espacial e temporalmente, ultrapassam um limiar pré-estabelecido. A principal desvantagem desses métodos é que eles exigem muito do usuário em termos de modelagem estocástica e podem ter um custo computacional elevado ao se fazer vigilância espaço-temporal, já que é necessário reajustar o modelo à medida que novos eventos ocorrem.

Rogerson (2001) combina a estatística de Knox sob uma perspectiva local com somas acumuladas, de forma a detectar onde e quando ocorre uma interação espaço-temporal. Esse método, usando a estatística de Knox local, foi avaliado por Marshall et al. (2007), comparando os efeitos dos parâmetros e dos dados nos resultados, e observou-se que o método de Rogerson não teve um bom desempenho na detecção de clusters. Um método mais recente de vigilância prospectiva espaço-temporal é proposto por Assunção and Correa (2009), que utilizam martingales para fazer a vigilância de dados pontuais. Uma desvantagem desse método é assumir um cluster de formato espacial circular.

O método de superfícies acumuladas foi originalmente desenvolvido por Simões and Assunção (2005) e tem como objetivo a detecção prospectiva de clusters de eventos pontuais no espaço e no tempo. A cada novo evento que é registrado, calcula-se um escore local de Knox, definido por Rogerson (2001), que leva em consideração o número de vizinhos próximos no espaço e no tempo. No entanto, as fórmulas utilizadas no cálculo do escore de Knox local foram redefinidas

para a abordagem prospectiva e serão apresentadas neste trabalho. O escore local é distribuído no espaço através de uma densidade de kernel, formando uma superfície estocástica. As superfícies são acumuladas sequencialmente à medida que os eventos ocorrem. Os picos dessas superfícies identificam áreas que presenciaram recentemente um número de eventos maior que o esperado. Se a superfície acumulada ultrapassar um limiar determinado, um alarme é soado e, um ou mais clusters localizados são identificados.

Com a utilização das superfícies é possível identificar os locais dos possíveis clusters no momento em que o alarme soa. Também é possível visualizar os níveis de interação espaço-temporal em toda a região, e captar a emergência de clusters de diferentes formatos. Analisando-se os valores dos escores locais de cada evento, também é possível determinar quais eventos provém do cluster. Na Figura 1.1, podemos ver o resultado da aplicação do método para dados de meningite em Belo Horizonte. O gráfico apresenta a superfície acumulada até o momento em que o alarme soou, com os eventos que foram registrados e a região do possível cluster destacada.

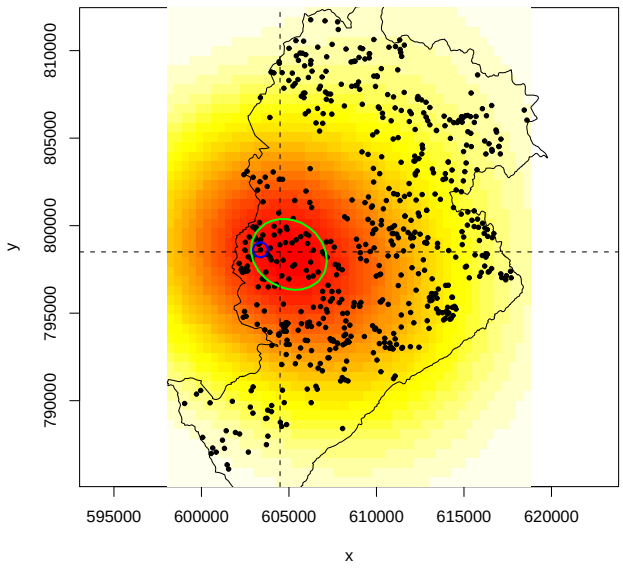


Figura 1.1: Visualização da superfície acumulada em níveis de contorno, para os casos de meningite

Outra vantagem muito importante da metodologia apresentada é que ela não exige uma modelagem *a priori* dos padrões espacial e temporal dos dados, o que torna o método acessível para usuários com pouco conhecimento estatístico, podendo ser utilizado diretamente por órgãos responsáveis por tomar medidas necessárias na ocorrência de um cluster.

1.1 Objetivos

O objetivo desta dissertação é a apresentação do método desenvolvido de vigilância espaço-temporal com superfícies acumuladas, juntamente com as fórmulas corrigidas para a abordagem prospectiva, bem como avaliar a performance do método a partir de resultados de simulações e a aplicação a um banco de dados.

1.2 Organização do Trabalho

Esta dissertação está organizada da seguinte maneira: o Capítulo 2 concentra a descrição da metodologia e os resultados obtidos, no formato de um artigo a ser submetido para uma revista internacional. Na Seção 2.2, o escore local de Knox, originalmente definido por Rogerson (2001), é descrito e as fórmulas são corrigidas para uma vigilância prospectiva. A metodologia de Superfícies Acumuladas é apresentada na Seção 2.3 e os resultados para as simulações estão na Seção 2.4. O método foi aplicado a um banco de dados reais e os resultados obtidos estão na Seção 2.5. Finalmente, na Seção 2.6, nós discutimos os resultados e conclusões encontrados.

Capítulo 2

Prospective space-time surveillance with geographical identification of the emerging cluster

2.1 Introduction

The simultaneous consideration of spatial and temporal patterns of occurrence of events is important to identify spatial-temporal clusters. We define a spatial-temporal cluster as a region geographically small in relation to the study area and concentrating an excessive number of events in a limited period of time.

The methods for detecting spatial-temporal clusters are mostly retrospective. These methods look for evidence of the presence of a cluster in a database of events that have already occurred. The interest is to understand the process that generated the events, and identify potential risk factors. The most popular test to detect spatial-temporal clusters was developed by Knox (1964). It tests if there is space-time interaction. That is, it tests whether cases near each other in space tend to be close in time as well. The Knox test is based on the number of pairs of events that are simultaneously close in space and time, with proximity being defined as a binary indicator variable based on numerical thresholds. Mantel (1967) extended the Knox test, using the spatial and temporal distances between pairs of events instead of binary indicators of proximity. Jacquez (1996) proposed a test for space-time interaction of the k nearest neighbors and compared the results with the Knox and Mantel tests, showing that his test had a greater power. The problem is that these three tests can be biased if the underlying population at risk changes differentially. Kulldorff (1999) proposed a modification to the Knox test to fix this problem, with a drawback of requiring information about the underlying population.

When dealing with events such as crimes and diseases, an important goal is to detect an emerging cluster through prospective surveillance. The challenge is to develop efficient surveillance methods that rapidly detect clusters as soon as possible after their emergence, while controlling the number of false alarms. The interest is focused on actions that can mitigate the effects of the clusters if they are performed quickly enough.

There is a large number of purely temporal prospective methods, especially in epidemiology. The methods most used to make epidemiological prospective surveillance are summarized and evaluated in Sonesson and Bock (2003). Höhle (2007) developed a package for R to make epidemiological prospective surveillance.

Very few spatial-temporal prospective surveillance methods have been developed so far. One method was presented by Kulldorff (2001) and it calculates a space-time scan statistic counting the number of events within a cylinder of radius and height pre-established. Neill et al. (2005) also developed a method to detect emerging clusters using a space-time scan statistic. An extension for the scan method was proposed by Kulldorff et al. (2005), which does not require information about the population at risk. However, these methods are based on ideas of hypothesis testing, which are not appropriate in the context of prospective surveillance (Woodall et al., 2008). Rodeiro and Lawson (2006) used spatial-temporal Bayesian models to detect an increase in disease risk on maps with data areas. Diggle et al. (2005) also present a model to monitor point process, predicting spatially and temporally localised excursions over a pre-specified threshold. The main disadvantage of these methods is that they require a lot from the user regarding stochastic modelling and they can have a high computational cost when making space-time surveillance, since it is necessary to refit the model when new data arrive.

Rogerson (2001) combines the Knox statistic under a local perspective with accumulated sums in order to detect when a space-time cluster is emerging. This method, using the local Knox statistic defined by Rogerson (2001), was evaluated by Marshall et al. (2007). They compared the effects of parameters and data on the results, and concluded that the method of Rogerson did not have a satisfactory performance. A more recent prospective spatial-temporal surveillance method is proposed by Assunção and Correa (2009), using a martingale approach to monitor point patterns. The disadvantage of this method is that it assumes a spatially circular shaped cluster.

The Cumulative Surface method was developed by Simões and Assunção (2005) and aims at

the prospective detection of clusters of space-time point patterns. For every new event that occurs, it is calculated a local Knox score, with the definition from Rogerson (2001) modified. This local score takes into account the number of events close in space and time. This local score is distributed in the space through a kernel density, forming a stochastic surface. The surfaces are accumulated sequentially as the events occur. The peaks of these surfaces identify areas where a number of events greater than expected recently happened. If the surface exceeds a certain threshold, an alarm is sounded, and one or more localized clusters are identified.

There are several advantages of using surfaces. In contrast with other methods of surveillance, it is possible to identify the locations of probable clusters, not only to signal its emergence. Using surfaces also enables to visualize the levels of space-time interaction all over the study region, and capture the emergence of clusters with different shapes. Analyzing the scores for each event, it is also possible to determine which events comes from the cluster.

Furthermore, the methodology is based on sequential inferencial procedures, unlike other methods, such as Kulldorff (2001) based on hypothesis testing. With this we avoid the problem of controlling multiple successive tests.

Another very important advantage of our methodology is that it does not require prior modeling of the spatial and temporal patterns of data. This makes the method accessible to users with little statistical knowledge, so it can be used directly by public agencies responsible to take necessary actions in the occurrence of a cluster.

2.2 Review of the local Knox method

Rogerson (2001) defined a local Knox statistic, decomposing the global measure of space-time interaction developed by Knox (1964) into local scores associated with each individual event. For each pair of events, we will define a dummy variable indicating if the two events are at a distance smaller than a critical space radius r_s. Following the notation of Rogerson (2001), let n_s be the total number of pairs of events close in space. Similarly, we will define a binary variable indicating that the waiting time between two events is smaller than a critical time r_t and let n_t be the total number of such pairs. We will denote by n_{st} the total number of pairs of events that are close in space and time. The test of Knox (1964) is based on this statistic.

Rogerson (2001) also defines $n_s(i)$, $n_t(i)$ e $n_{st}(i)$ as the number of events close in space, close

in time and close in space and time from the i-th event, respectively, with $i = 1, \ldots, n$. The Knox statistic n_{st} can be decomposed in terms of local statistics $n_{st}(i)$, since $n_{st} = {}^1/_2 \sum_{i=1}^{n} n_{st}(i)$.

The local Knox score defined by Rogerson (2001) is:

$$z_i = \frac{n_{st}(i) - E\{N_{st}(i)\} - 0.5}{\sqrt{Var\{N_{st}(i)\}}} \tag{2.1}$$

where $N_{st}(i)$ is the random variable associated with the observed value $n_{st}(i)$. To determine the null distribution of $N_{st}(i)$, Rogerson (2001) assumes that each random permutation of the times, keeping the spatial coordinates fixed, is equally likely. He also considers that the i-th event can be any of the n events observed, and his time is also permuted. So, the location i can be assigned to any of the times from $j = 1, \ldots, n$. The resulting distribution is a weighted sum of hypergeometric distributions, each one corresponding to the times that can be assigned to location i. With this, he obtained:

$$E\{N_{st}(i)\} = \frac{2n_t n_s(i)}{n(n-1)} \tag{2.2}$$

$$Var\{N_{st}(i)\} = \frac{\left[2(n-1)n_t - \sum_{j=1}^{n} n_t(j)^2\right] n_s(i)\,(n-1-n_s(i))}{n(n-1)^2(n-2)} \tag{2.3}$$

Simões and Assunção (2005) and Marshall et al. (2007) found that (2.3), the variance proposed by Rogerson, was inaccurate and corrected it:

$$\begin{aligned} Var\{N_{st}(i)\} &= \left[\sum_{j=1}^{n} n_t(j)^2\right] \frac{n_s(i)}{n(n-1)^2}\left[-\frac{(n-1-n_s(i))}{n-2} + n_s(i)\right] \\ &\quad + \left[\frac{2n_t n_s(i)}{n(n-1)}\frac{n-1-n_s(i)}{n-2} - \left(\frac{2n_t n_s(i)}{n(n-1)}\right)^2\right] \\ &= \frac{\left[2(n-1)n_t - \sum_{j=1}^{n} n_t(j)^2\right] n_s(i)\,(n-1-n_s(i))}{n(n-1)^2(n-2)} + \frac{n_s(i)^2 \sum_{j=1}^{n}\left[n_t(j) - \frac{2n_t}{n}\right]^2}{n(n-1)^2} \end{aligned} \tag{2.4}$$

Marshall et al. (2007) notes that the variance (2.3) proposed by Rogerson is an approximation for (2.4) but it will be always less or equal than the correct formula, since (2.4) is the sum of (2.3) with a non-negative term.

However, there is a serious problem in using these local scores in a prospective method. Since the goal is to make prospective surveillance, we are interested in measuring the local space-time interaction as each new event occurs. The statistic $n_{st}(i)$ is calculated for each new event that arises and, at this moment, there are only those events that happened before the event in question. Future events have not occurred yet and can not be used in the calculation of $n_{st}(i)$.

As defined by Rogerson (2001), the local scores are suitable for detection of clusters in retrospective analysis, but are not appropriate for a prospective use. The computation of $n_{st}(i)$ in (2.2) considers as possible neighbors of the i-th event both the events that occurred before the i-th, and those that occurred later. Consequently, the calculation of moments is not done correctly. This explains the disappointing results of the prospective Rogerson method found by Marshall et al. (2007). If the local score is defined appropriately under the prospective context, as explained ahead, the Rogerson method presents a very good performance (Piroutek, Assunção, and Paiva, 2010).

To correct the Rogerson method, we consider that the i-th event is the last one, so there are no events with times greater than it. We will define $N^*_{st}(i)$ as the random variable of the number of events close in space and time of the i-th event, considering only the events that happened before it. It is also necessary to define:

$n^*_s(i)$ number of events that are close in space of the i-th event, considering only those events that happened before it;

$n^*_t(i)$ number of events that are close in time of the i-th event, considering only those events that happened before it;

$n^*_{st}(i)$ number of events that are close simultaneously in space and time of the i-th event, considering only those events that happened before it;

To find the null distribution of $N^*_{st}(i)$, we permute the observed times across the spatial coordinates, but now we will keep the i-th event fixed. The resulting distribution will be only one hypergeometric distribution, instead of the weigthed sum of Rogerson (2001).

The distribution can be visualized considering the drawing of arrows corresponding to the times of events. The arrows are divided into two types: type A are the ones that are close in time

of the i-th event, and the remaining ones are of type B. So we will have a total of $(i-1)$ arrows, of which $n_t^*(i)$ are of type A. We take a sample of size $n_s^*(i)$ from the $(i-1)$ arrows to distribute among the events that happened close in the space of the i-th event. Figure 2.1 is an example of a set of points, with the spatial coordinates represented in the horizontal axis and the times of occurrence of the events represented on the vertical axis, with the time of the i-th event highlighted. The dotted arrows represent the events of type A, those who are close in time of the i-th event. The events with the locations within the circle are those who are neighbors in space of the i-th event.

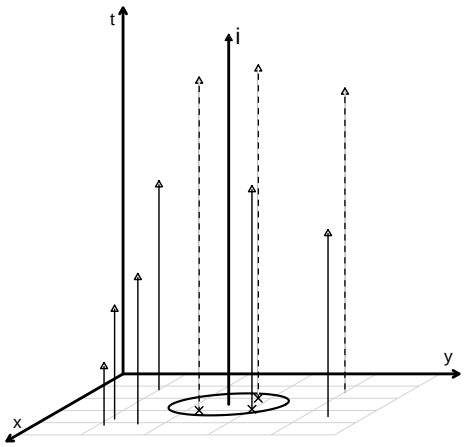


Figure 2.1: View of the neighborhood of i-th event

Hence, $N_{st}^*(i)$ follows a hypergeometric distribution with parameters $(i-1)$, $n_t^*(i)$ e $n_s^*(i)$. That is,

$$P\left\{N_{st}^*(i) = k\right\} = \frac{\binom{n_t^*(i)}{k}\binom{i-1-n_t^*(i)}{n_s^*(i)-k}}{\binom{i-1}{n_s^*(i)}}$$

The mean and the variance of $N_{st}^*(i)$ are given by:

$$E\left\{N_{st}^*(i)\right\} = \frac{n_s^*(i)n_t^*(i)}{i-1} \tag{2.5}$$

$$Var\left\{N_{st}^*(i)\right\} = \frac{n_s^*(i)n_t^*(i)(i-1-n_s^*(i))(i-1-n_t^*(i))}{(i-2)(i-1)^2} \tag{2.6}$$

It is not possible to compare the variances (2.4) and (2.6), since they are based in $n_{st}(i)$ e $n_{st}^*(i)$, respectively, amounts defined in different ways.

2.3 Methodology

2.3.1 Cumulative Surfaces

The Cumulative Surface method is based on the local Knox scores defined in (2.1), with mean and variance given by (2.5) and (2.6), respectively. For each event, it is counted the number of events that are close to him in space and that happened a little earlier in time. The local Knox score is then calculated by standardizing this number according to (2.1).

The score is spread around the position of the i-th event through a two-dimensional kernel function. The function used in the method is the bivariate Gaussian density function defined as:

$$K^*(x, y) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x^2 + y^2)\right\} \tag{2.7}$$

The kernel functions modify the functions $K^*(x, y)$, transferring them to a new center and changing its concavity with the bandwidth parameter τ. Namely, the positive z_i scores are spread around the coordinates (x_i, y_i) of the i-th event, forming surfaces $w_i(x, y)$ given by:

$$w_i(x, y) = z_i^+ K_i(x, y) = \frac{z_i^+}{\tau^2} K^*\left(\frac{x - x_i}{\tau}, \frac{y - y_i}{\tau}\right) \tag{2.8}$$

where $z_i^+ = \max\{0, z_i\}$.

Note that

$$\iint w_i(x, y)dxdy = z_i^+$$

To form the surface of the i-th event denoted by w_i, the region is divided into a grid with size specified. This grid is determined by dividing the range of the events on the axis x and y in the size established. Coordinates (x, y) are then the points of the grid where the surface is calculated.

At the point (x_i, y_i) the surface has its maximum values equal to $w_i(x_i, y_i) = z_i^+ K^*(0, 0)/\tau^2 = z_i^+/2\pi\tau^2$. As the distance between the grid point (x, y) and the event coordinates (x_i, y_i) increase, $K^*(x, y)$ becomes $1/2\pi$ multiplied by a value between 0 and 1 getting smaller. The value of the surface goes to zero for locations away from the event.

Figure 2.2 shows, for a random set of points, how the score z_i^+ is spread around its position, forming the surface corresponding to the i-th event.

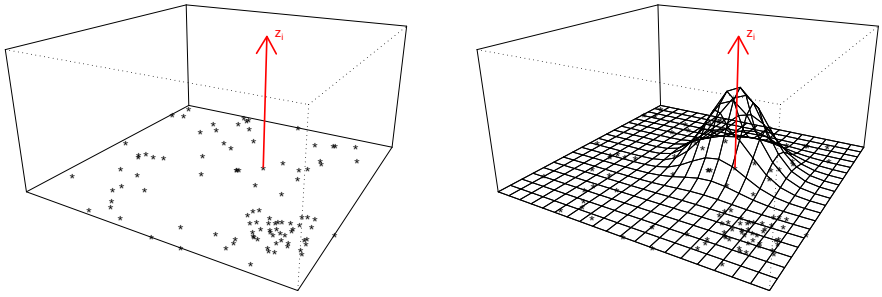


Figure 2.2: Three-dimensional visualization of the score z_i and its respective surface, for a random set of points

The bandwidth τ affects mainly the concavity of the surface. Small values for this parameter make the surface decays rapidly and abruptly, while higher values make the surface decays more slowly.

The bandwidth can be specified as a fixed value in order to get an specific view of the surfaces, if there is knowledge about the range and scale of the data coordinates. Otherwise, the bandwidth value can be estimated automatically. This is done through simulations, generating coordinates uniformly distributed between the minimum and maximum coordinates of the events. At each simulation, it is calculated a value of τ using the following formula, suggested by Härdle (1990) for a random variable Z and sample size n:

$$\tau = \frac{1.06}{n^{1/5}} \min\left\{sd(Z), \frac{iqr(Z)}{1.34}\right\} \tag{2.9}$$

where $sd(Z)$ is the standard deviation and $iqr(Z)$ is the interquartile range of the variable Z. In our case, the variable Z will be replaced by the events' coordinates (x, y), the value of $sd(Z)$ will be the average between $sd(\mathbf{x})$ and $sd(\mathbf{y})$ and the interquartile range $iqr(Z)$ will be the average between $iqr(\mathbf{x})$ and $iqr(\mathbf{y})$. At the end of the simulations, the estimated bandwidth will be the average of the values obtained at each simulation.

As the scores z_i take into account the information about the neighbors in space and time, we define the window m as the number of events that is necessary to wait until the scores are stabilized. The reason is that in the first events there is little information about the pattern of occurrence of the other events. In Figure 2.3, we can see an example of a time series plot of z_i^+ scores. This scores were calculated for 500 events with spatial coordinates uniformly distributed in the region

$[0, 10] \times [0, 10]$ and times also uniformly distributed in $[0, 10]$. In this case, a value of the m^* window equal to or greater than 100 would be enough for the escores to stabilize.

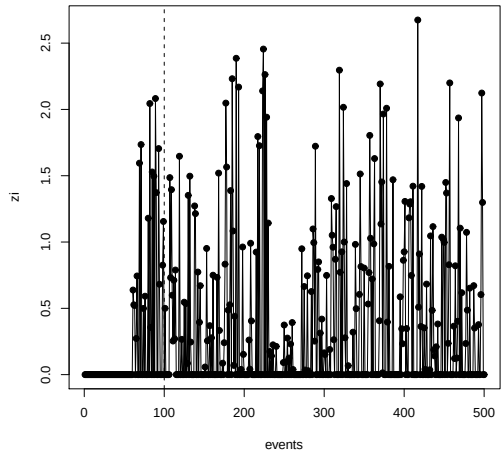


Figure 2.3: Time series of the scores z_i^+ obtained in an illustrative example

We also define another window m^* as the number of surfaces of events prior to the i-th event that are going to be accumulated. This window is defined because the oldest events do not help to detect an emerging cluster, causing only a noise in the variance.

At each i-th event, we accumulated iteratively the last m^* surfaces $w_j(x, y)$:

$$\begin{aligned} S_i(x,y) &= \sum_{j=i-m^*+1}^{i} w_j \\ &= S_{i-1}(x,y) - w_{i-m^*}(x,y) + z_i^+ K_i(x,y) \quad \text{with } i = 1, 2, \ldots, n. \end{aligned} \tag{2.10}$$

2.3.2 Determining the threshold h

The threshold h must be determined as a value of the distribution of the maxima of the cumulative surfaces, such that it is under control the false alarm probability. As we are looking at the maxima of stochastic surfaces that are accumulated over time, it is not immediate how to find a theoretical distribution to determine the threshold h. However, we use two approaches to find this value, using in both the maxima of the maxima of the surfaces $\max_i\{\max_{(x,y)} S_i(x, y)\}$ obtained by permutations of the events' times.

Let us define one more window m^{**} as the number of successive accumulated surfaces $S_i(x, y)$, that will be used to control the probability of false alarms under the assumption that there is no cluster. So, the maximum of the maxima of the surfaces will be searched only in the cumulative

surfaces corresponding to the last m^{**} events, that is $i = n - m^{**} + 1, \ldots, n$.

Controlling the probability of false alarms with the window m^{**} corresponds to determine the *ARL*, Average Run Length. It is the expected number of events to happen until the alarm goes off falsely, used in other surveillance methods like Rogerson (2001) and Assunção and Correa (2009).

The first step in the method is to test if the number of events in a training dataset is greater than or equal to the sum of the three windows, i.e.

$$n \geq m + m^{*} + m^{**}$$

If this condition is met, then the data set is large enough so that the scores z_i are stable and we can use the windows m^{*} and m^{**}.

At each permutation, the spatial positions are kept fixed and times are permuted. That is, we generate new configurations of events under the assumption that there is no space-time interaction. These simulated configurations follow the purely spatial and purely temporal pattern of the data. The surfaces are then accumulated and the maxima of the maxima of the surfaces of the last m^{**} events are stored. From now on, we will refer to the these values as the maxima of the surfaces.

In the first approach, we fit a Gumbel distribution to the maxima of the surfaces, and the threshold h_{teo} is the value obtained through its distribution function. The reason for this is a theorem of Piterbarg (1996). In the second approach, the threshold h_{emp} is obtained as the percentile of the empirical distribution function of the maxima. We detail below the two approaches.

Theoretical threshold

The stochastic process $\{S(\mathbf{x}),\ \mathbf{x} \in T\}$, where $T \subset \mathbb{R}^k$ and $\mathbf{x} \in T$ represents a location, is a Gaussian field if, for any $k \geq 1$ and any locations $x_1, \ldots, x_k \in T$, the vector $(S(x_1), \ldots, S(x_k))^T$ is normally distributed (Rue and Held, 2005). Let $S(\mathbf{x})$, with $\mathbf{x} \in \mathbb{R}^n$, be a homogeneous Gaussian field - i.e., with constant moments over the region - with zero mean, covariance function $\rho(\mathbf{x}) = \int_{\mathbf{k}} e^{i\mathbf{k}\mathbf{x}} \Psi(\mathbf{k}) d\mathbf{k}$, $\rho(\mathbf{0}) = \sigma^2$, and spectrum $\Psi(\mathbf{k})$.

When $n = 2$ and $\mathbf{x} = (x, y)$, we can view the Gaussian field as a surface, with oscillations forming peaks, or waves, throughout the region. The wavelength is the distance between the points where the wave begins and where it ends, in the propagation direction. The crest length of the wave is the distance between the points where the crest begins and where it ends, in the perpendicular

direction of the propagation. In this case, let the unit volume be $|V| = \lambda_0 \lambda_c$, where λ_0 is the mean wavelength and λ_c the mean crest length. That is, the unit volume is the mean volume that a wave occupies.

Piterbarg's theorem (1996, Theorem 14.1) determines the asymptotic extreme value distribution for the maximum of a Gaussian homogeneous field in $\mathbb{R}^n$. Let T be the subset $T \subset \mathbb{R}^n$ with volume $|T|$, or with size $N = |T|/|V|$, where $|V|$ is the unit volume defined above. According to the theorem, we have:

$$P\left(\max_{\mathbf{x}\in T} S(\mathbf{x}) \leq \sigma u\right) \sim \exp\left[-(2\pi)^{\frac{n-1}{2}} e^{-u^2/2} H_{n-1}(u) N\right] \tag{2.11}$$

where H_n are Hermite polynomials with respect to the standard Gaussian density ($H_0(u) = 1$, $H_1(u) = u$, $H_2(u) = u^2 - 1, \ldots$). When N increases, i.e. the number of waves within the subset T increases, the distribution tends asymptotically to a Gumbel distribution:

$$G(u) = \exp\left(-\exp\left(-a(u - u_0)\right)\right) \tag{2.12}$$

where $u_0 \approx x_0 + \dfrac{(n-1)\log(x_0)}{x_0}$, $a = u_0 - \dfrac{(n-1)}{u_0}$, and $x_0 = \sqrt{2\log N + (n-1)\log(2\pi)}$.

In our case, we are considering the cumulative surface as the homogeneous Gaussian field as of the i-th event. To apply the approximation (2.12), we need N, the expected number of waves in the subset T, which is not simple to calculate in our method. Thus, we use a theoretical approximation to the distribution of the maxima of the surfaces to determine the threshold h.

Let us consider a Gumbel distribution with parameters α and β, with the following density and distribution functions:

$$f(x) = \frac{1}{\beta} \exp\left\{-\frac{(x-\mu)}{\beta}\right\} \exp\left\{-\exp\left[-\frac{(x-\mu)}{\beta}\right]\right\} \tag{2.13}$$

$$F(x) = \exp\left\{-\exp\left[-\frac{(x-\mu)}{\beta}\right]\right\} \tag{2.14}$$

To verify the adequacy of our assumption about the Gumbel distribution, we analyzed the maxima of the maxima of the cumulative surfaces $\max_i\{\max_{(x,y)} S_i(x,y)\}$, obtained from 1000 permutations of data generated under the null hypothesis of no clusters. Figure 2.4 is the QQ-Plot of the observed quantiles of the maxima found. In this figure, we plot in the vertical axis the order

statistics of the maxima of the maxima of the surfaces, obtained from 1000 permutations, and in the horizontal axis the theoretical quantiles of a standard Gumbel distribution, with parameters $\mu = 0$ and $\beta = 1$.

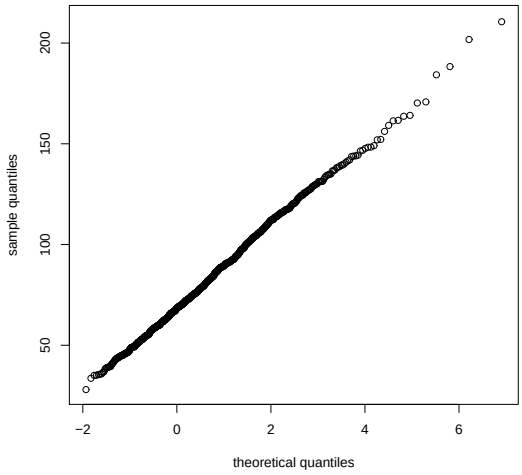


Figure 2.4: QQ-Plot of the theoretical quantiles of the Gumbel distribution versus the observed quantiles of the maxima of the surfaces, in an illustrative example

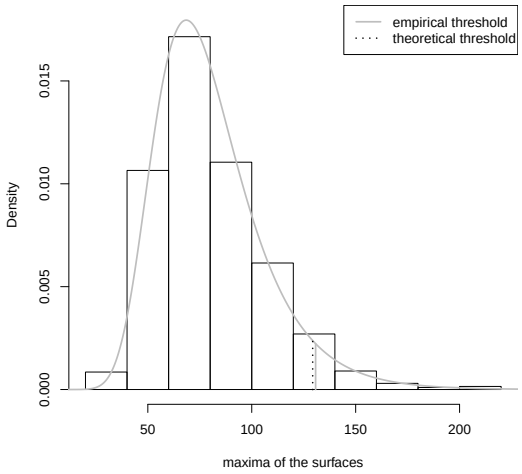


Figure 2.5: Histogram of the maxima of the surfaces, with the Gumbel density and the thresholds highlighted, in an illustrative example

Since the points plotted in the QQ-Plot of Figure 2.4 are very close to a straight line, one can consider that the observed and theoretical quantiles are linearly related. This suggests that, in the scenario where there is no space-time interaction, the observed values seems to follow a Gumbel distribution with different parameters of location and scale. As we will use this distribution of the maxima of the surfaces to find a threshold h_{teo} under the null hypothesis, it is sufficient to check that this approximation is valid in this scenario.

The maximum likelihood estimators of the parameters μ and β of the maxima's distribution can not be found analytically. We find them numerically, with initial values obtained by the method of moments. With $\hat{\mu}$ and $\hat{\beta}$, we can determine, from the distribution function (2.14), the threshold h_{teo} as the value such that the probability that at least one of the last m^{**} cumulative surfaces overcome it is equal to one minus the distribution function at the point h_{theo}, which is equal to α,

the probability of false alarm.

$$P\Bigg(\max_{x,y}\{S_{n-m^{**}+1}(x,y)\} > h_{teo} \;\; or \;\; \max_{x,y}\{S_{n-m^{**}+2}(x,y)\} > h_{teo} \;\; or \;\; \ldots \;\; or \;\; \max_{x,y}\{S_n(x,y)\} > h_{teo}\Bigg) = 1 - F(h_{teo}) = 1 - \exp\left\{-\exp\left[-\frac{(h_{teo}-\hat{\mu})}{\hat{\beta}}\right]\right\} = \alpha \qquad (2.15)$$

In Figure 2.5, we see how the density of a Gumbel distribution with parameters $\hat{\mu}$ and $\hat{\beta}$ fits well the histogram of the maxima of the cumulative surfaces found in 1000 permutations under the assumption that there are no clusters. Moreover, we can see that the threshold values h_{teo} and h_{emp}, which will be described in the next section, are very close.

Empirical threshold

To determine the threshold h_{emp}, we will use again the window m^{**} to control the probability of false alarm. A threshold h is determined in order that the probability that at least one of the last m^{**} surfaces' maxima is higher than h is equal to α. This is the probability of a false alarm, given by

$$P\Bigg(\max_{x,y}\{S_{n-m^{**}+1}(x,y)\} > h_{emp} \;\; or \;\; \max_{x,y}\{S_{n-m^{**}+2}(x,y)\} > h_{emp} \;\; or \;\; \ldots \;\; or \;\; \max_{x,y}\{S_n(x,y)\} > h_{emp}\Bigg) \;=\; \alpha \qquad (2.16)$$

At the end of B random permutations of the events' times, we have B maxima of surfaces' maxima. The threshold h_{emp} is the $(1-\alpha)$-th percentile of the maxima distribution.

It is important to note that, since that we are observing maxima of surfaces, we are dealing with a very asymmetric distribution with heavy tail. Because of that, a sample percentile may not be adequate to estimate a theoretical percentile, if the sample size is not large enough.

2.3.3 Overview of the surveillance method

The cumulative surface method was implemented in C language, divided in two parts. The first part uses a training dataset to make the times' permutations and determine the threshold value. Then, this value is used in the second part, where the surveillance is made.

For the first part, the necessary input is a training dataset, as well as the values for the windows m, m^* and m^{**}, the critical distance r_s and the critical time r_t, the false alarm probability α, the

bandwidth τ, and the grid size of the surface. The output is the value of the thresholds h_{teo} e h_{emp}. For a dataset of 500 events, the average running time of this part is 3.5 minutes, in a machine with 2.80 GHz and 2 Gb RAM.

After the threshold h is determined, it is not necessary to repeat this step when new events are added to the database. Permutations must be done again only if its judged that the overall events' incidence pattern has changed.

In the second part of the method, the necessary input is the dataset, along with the parameters already fixed m, m^*, m^{**}, r_s, r_t, α, τ and the grid size. In our software, we allow the surveillance to be run only for the last events in the input dataset. For that, the user specifies the number a of new events to be monitored. At each new event, the method accumulates the surfaces taking into account the window of m^* previous events.

$$\begin{aligned} S_i(x,y) &= \sum_{j=i-m^*+1}^{i} w_j \\ &= S_{i-1}(x,y) - w_{i-m^*}(x,y) + z_i^+ K_i(x,y) \\ &\text{with } i = n-a+1,\ldots,n. \end{aligned} \tag{2.17}$$

The alarm goes off immediately after the i-th event if $S_i(x,y) > h$ for some position (x,y). If this happens, the method outputs the surface accumulated until this moment and the z_i values. With the view of the surface, it is possible to visualize other areas that may also have high levels of space-time interaction. We identify as the locations of possible clusters the regions where the surface was above the threshold h. With the same computer configuration and 500 events, the average running time of the second part is less than one minute.

The plan for the future is to implement the two parts as an surveillance method in an R package, to allow that it becomes acessible to different users.

2.4 Simulations

To evaluate the quality of the Cumulative Surface method, we analyzed simulation results in different scenarios. First, it is generated a database of 500 events, with the spatial coordinates x and y in the region $[0,10] \times [0,10]$ and the time coordinates between $[0,10]$, distributed according to the scenario in question. These events are then used to determine the thresholds h_{emp} and h_{teo},

permuting the temporal coordinates 1000 times, keeping the spatial coordinates fixed, as described in Section 2.3.2.

In all scenarios, we use the critical space radius r_s equal to 2 and critical time r_t equal to 1. Thus, when data follow a uniform distribution in $[0, 10]^3$, the expected number of events in the critical region is equal to 6.28. The windows m and m^{**} are set equal to 100. These values were determined proportionally to the total number of events, such that the scores are stabilized and the probability of false alarm is controlled as 0.05 for the last m^{**} events. The window m^*, which corresponds to how many previous events are accumulated to build the surface associated to the i-th event, took the values 10, 25 and 50. In this way, we can analyze its impact on the proportion of false alarms. The grid size to evaluate the continuous surface was set as 50×50. To obtain a better visualization of the surfaces, with distinguishable peaks, the bandwidth was fixed as 0.25.

After determining the threshold value h, we can proceed to the second part of the surveillance. For each scenario, we ran 1000 simulations. At each simulation, data are generated according to the scenario pattern and the surveillance was carried out.

Let s be the number of total simulations and s_A be the number of simulations in which the alarm went off. Let $I_C(j)$ be the indicator variable signing whether the alarm went off correctly on the j-th simulation. It only makes sense when there are clusters, and it indicates if the alarm went off for a event that belongs to the cluster or occurred near it. Similarly, let $I_F(j)$ be the indicator variable of a false alarm in the j-th simulation. In the scenario without clusters, it indicates the simulations in which the alarm went off. When there are clusters, it indicates the simulations in which the alarm went off for events before the cluster's emergence or for events that are far from the cluster's location. Finally, let d_j be the delay of events that belong to the cluster that occurred until the alarm went off, in the j-th simulation. With the results of simulations, we will analyze the following measures:

- *FAR - False Alarm Rate*:

$$FAR = \frac{\sum_{j=1}^{s} I_F(j)}{s_A}$$

  When there is no cluster, it will be equal to one, since all the alarms are false.

- *AR - Alarm Rate*:

$$AR = \frac{s_A}{s}$$

It is the proportion of simulations in which the alarm went off. When there are no clusters, it corresponds to α, the false alarm probability.

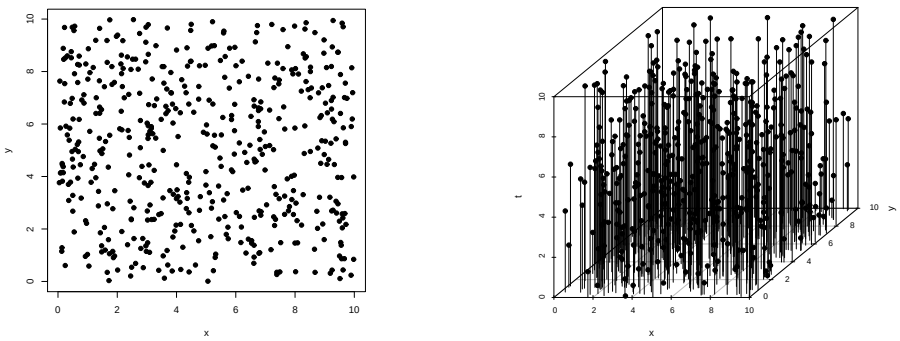
- *CED - Conditional Expected Delay*:

$$CED = \frac{\sum_{j=1}^{s} d_j \, I_C(j)}{\sum_{j=1}^{s} I_C(j)}$$

It is the mean number of events from the cluster that happened before the alarm goes off correctly. When there is no cluster, it is equal to zero.

2.4.1 Scenario without clusters

This is the scenario under the null hypothesis, where there is no cluster. Events are generated with the coordinates uniformly distributed in the region $[0, 10]^3$. The typical data used to perform the permutations can be seen in Figure 2.6.



(a) Spatial coordinates **(b)** Spatial and temporal coordinates

Figure 2.6: View of the coordinates of events generated in the scenario without cluster

The thresholds h_{emp} and h_{teo} are given in Table 2.1(a) and were calculated by using the permutations and the maxima distribution, respectively. We can also see them in Figure 2.7. The threshold values do not differ much, which means that the results of simulations should be similar for h_{emp} and h_{teo}.

Afterwards, the simulations were performed, generating new data sets, also uniformly distributed in $[0, 10]^3$, and all the possible events were monitored. That is, we monitor all events after

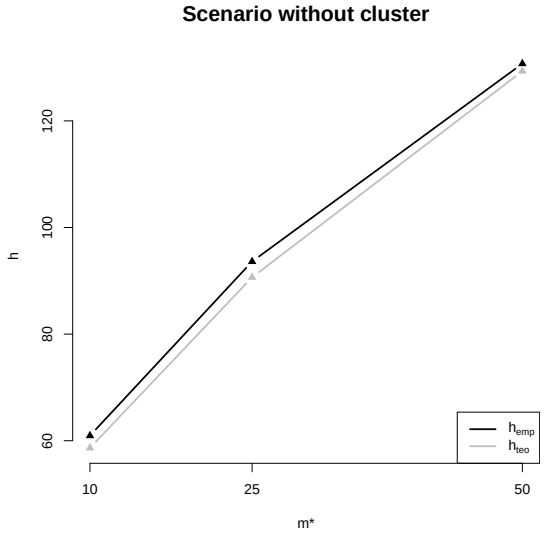


Figure 2.7: Comparative graphs of empirical and theoretical thresholds, for the scenario without clusters

the first m, the window of events needed for the scores' stabilization. Table 2.1(a) shows the results of the simulations for each value of m^*. As we are in the scenario without cluster, *FAR* is equal to 1.0 and *CED* is equal to 0.0.

Scenario without cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*60.943*	*93.612*	*130.735*
FAR	1.000	1.000	1.000
AR	0.111	0.108	0.134
CED	0.000	0.000	0.000
$\mathbf{h_{teo}}$	*58.649*	*90.655*	*129.327*
FAR	1.000	1.000	1.000
AR	0.140	0.134	0.143
CED	0.000	0.000	0.000

(a) Surveillance for all events

Scenario without cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*60.943*	*93.612*	*130.735*
FAR	1.000	1.000	1.000
AR	0.036	0.046	0.054
CED	0.000	0.000	0.000
$\mathbf{h_{teo}}$	*58.649*	*90.655*	*129.327*
FAR	1.000	1.000	1.000
AR	0.047	0.054	0.059
CED	0.000	0.000	0.000

(b) Surveillance for the last 100 events

Table 2.1: Results of the scenario without cluster: (a) surveillance for all events; (b) surveillance for the last 100 events

Even under the assumption that there are no clusters, the proportions *AR* in Table 2.1(a) are larger than α fixed. This is due to the fact that in this case, the surveillance was made for all possible events, i.e. the last 400 events, and the empirical threshold was determined using the last $m^{**} = 100$ events.

To check if the proportions *AR* are according to the expectations, the surveillance was carried out only for the last m^{**} events, choosing the number of new events to be monitored a equal to 100.

In the results in Table 2.1(b), we see that the proportions are very close to the value set of $\alpha = 0.05$.

Thus, even increasing from 100 to 400 the number of events to be analyzed, the proportion of times the alarm went off did not increase in the same proportion, staying within the tolerance level α determined by user.

2.4.2 Scenario with clusters

In the scenarios with clusters, we generated 500 events. A cluster is formed generating a fixed number of events in a small region of space and a short time interval. The remaining events are generated uniformly distributed in the region $[0, 10]^3$. The values of the input parameters were the same as in the previous scenario.

For each scenario with cluster, we analyze the results for three different values of window $m^* =$ 10, 25 and 50, in order to measure the impact of the choice of this parameter in the results. The simulations were made with the thresholds h_{emp} and h_{teo}. In each simulation, the surveillance was made for all possible events, excluding only the initial window m for the scores' stabilization.

The performance of the method was evaluated for different scenarios varying the intensity of the cluster, its size and its format.

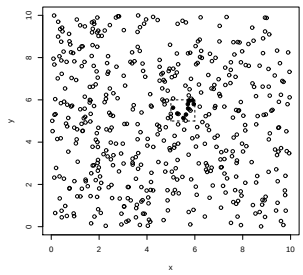
Clusters with different intensities

To see the impacts of different cluster intensities, we fixed the cluster region as $[5, 6]^2 \times [9, 10]$, adding 10, 25 or 50 events besides the underlying uniform intensity. An example of a dataset with the three cluster intensities can be viewed in Figure 2.8.

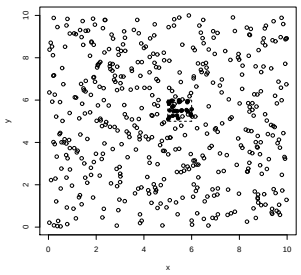
In Table 2.2 we see the thresholds h_{emp} and h_{teo} found for each of the values of m^*. The thresholds found were very close, as we can see in the graphs of Figure 2.9 confirming the adequacy of the theoretical approach of the maxima distribution described in Section 2.3.2. Therefore, the results with the empirical and the theoretical threshold did not differ much.

Moreover, the thresholds are on the same level for the different intensities of the cluster. This behavior is acoording to the expectations, considering that the presence of a cluster with a different intensity should not change the threshold value.

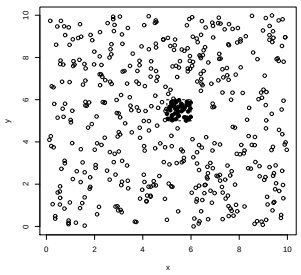
In Table 2.2(a), we can see that in the scenario with a cluster of 10 events, the false alarm rates *FAR* were high, around 27%. The proportions of times the alarm sounded *AR* were low, around 60%, since that there are clusters in all simulations. The mean time to detect the cluster *CED*



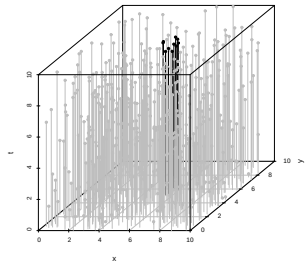
(a) Spatial coordinates, cluster with 10 events



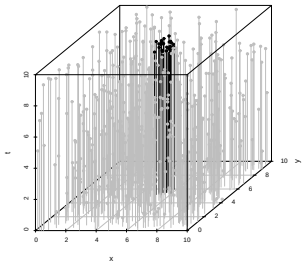
(b) Spatial coordinates, cluster with 25 events



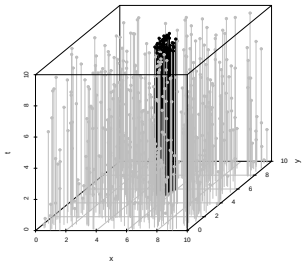
(c) Spatial coordinates, cluster with 50 events



(d) Spatial and temporal coordinates, cluster with 10 events



(e) Spatial and temporal coordinates, cluster with 25 events



(f) Spatial and temporal coordinates, cluster with 50 events

Figure 2.8: View of the events' coordinates generated in the scenario with clusters with different intensities

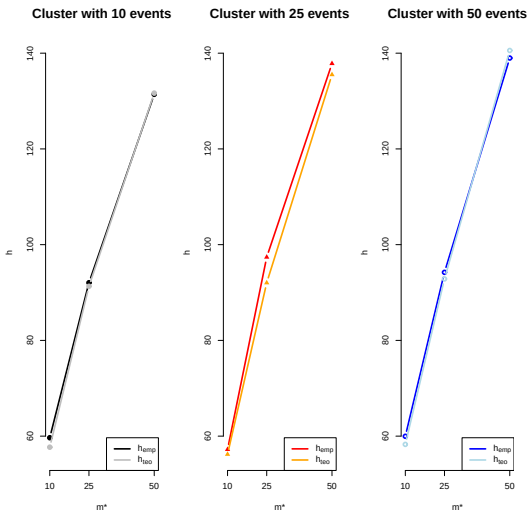


Figure 2.9: Comparative graphs of empirical and theoretical thresholds, for the clusters with different intensities

did not change with the window m^*, remaining around eight events. Since we are in a cluster originally consisted of only 10 events, the alarm sounded around 60% of the time, many of those times falsely.

For the cluster with 25 events, the results were better, according to Table 2.2(b). The values of

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*59.669*	*92.043*	*131.408*
FAR	0.273	0.243	0.282
AR	0.556	0.600	0.577
CED	7.953	7.956	8.145
$\mathbf{h_{teo}}$	*57.693*	*91.297*	*131.655*
FAR	0.300	0.257	0.275
AR	0.603	0.607	0.574
CED	7.502	7.707	8.036

(a) Cluster with 10 events

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*57.219*	*97.391*	*137.832*
FAR	0.170	0.115	0.103
AR	1.000	1.000	1.000
CED	8.997	9.964	11.205
$\mathbf{h_{teo}}$	*56.204*	*92.029*	*135.511*
FAR	0.187	0.137	0.107
AR	1.000	1.000	1.000
CED	8.877	9.644	11.113

(b) Cluster with 25 events

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*59.969*	*94.217*	*138.982*
FAR	0.122	0.099	0.078
AR	1.000	1.000	1.000
CED	8.641	9.889	11.607
$\mathbf{h_{teo}}$	*58.278*	*92.839*	*140.562*
FAR	0.121	0.114	0.079
AR	1.000	1.000	1.000
CED	8.498	9.842	11.661

(c) Cluster with 50 events

Table 2.2: Results in the scenario with cluster, with different intensities: (a) cluster with 10 events; (b) cluster with 25 events; (c) cluster with 50 events

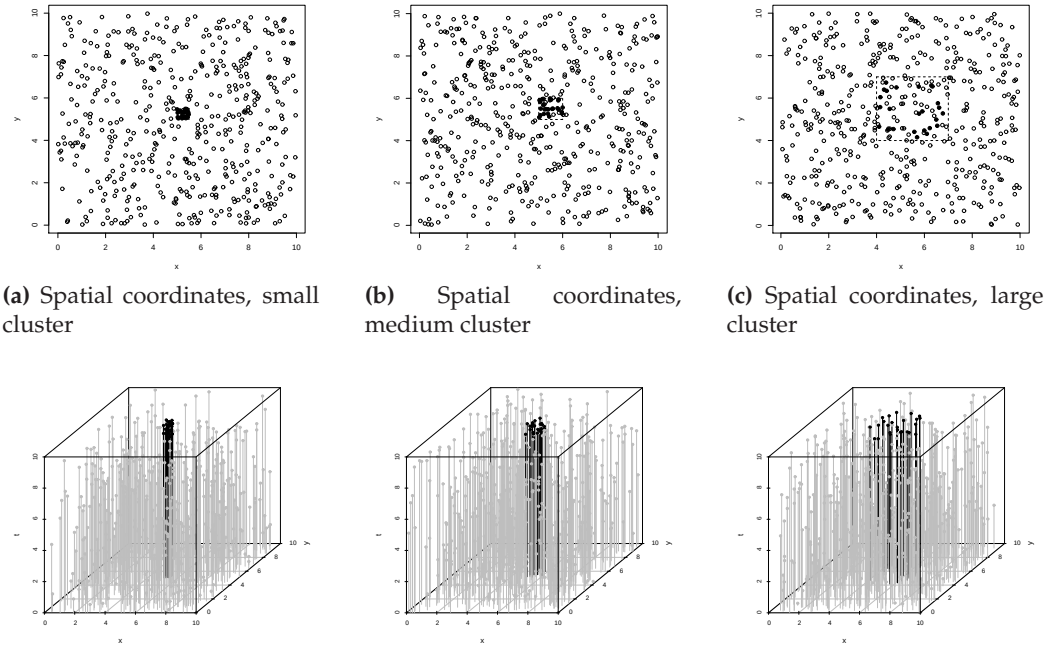
AR were equal to one, indicating that the alarm sounded in all simulations. As the window m^* increases, the false alarm rate decreased, which means that the more information is accumulated, the more the alarm sounds correctly. The mean time expected until the cluster is detected has increased in about one event, with respect to the cluster with 10 events, which means that, proportionally, the alarm sounded more quickly. But there is a trade-off between *FAR* and *CED*, since the *CED* increased directly with the window m^*. This is due to the fact that accumulating more information, the threshold h will increase, decreasing the number of surfaces that would overcome it before the cluster actually starts. Thus, the surface corresponding to the cluster will also take slightly longer to overcome the threshold.

In the scenario with the cluster of greater intensity, as we see in Table 2.2(c), values of *FAR* decreased even more, since the greater the intensity of the cluster, the greater its discrepancy with respect to the rest of the region, and the easier it is to detect it. Again, the alarm sounded in all

simulations. The values of *CED* have not changed with respect to the scenario with 25 events, again, meaning that the cluster is detected more quickly. We must consider whether the gain in reducing *FAR* is better than the delay of some events when detecting a cluster, according to the type of event and the actions to be taken facing an alarm.

Clusters with different extents

To evaluate the detection power for different cluster extents, we fixed the number of events belonging to the cluster as 25 and vary the spatial region of the cluster as $[5, 5.5]^2$, $[5, 6]^2$ and $[4, 7]^2$, while maintaining the temporal coordinates uniformly distributed between $[9, 10]$. The spatial and temporal coordinates of a typical dataset can be viewed in Figure 2.10.



(a) Spatial coordinates, small cluster

(b) Spatial coordinates, medium cluster

(c) Spatial coordinates, large cluster

(d) Spatial and temporal coordinates, small cluster

(e) Spatial and temporal coordinates, medium cluster

(f) Spatial and temporal coordinates, large cluster

Figure 2.10: View of the events' coordinates generated in the scenario with clusters with different extents

The threshold values h_{emp} and h_{teo} are given in Table 2.3. Observing the graph comparing the thresholds in Figure 2.11, we see that the threshold values did not change when we are considering clusters spread in a narrower or wider area. These values are also very close to the thresholds found for the previous scenarios.

Comparing the more concentrated and the medium clusters in Table 2.3(a) and 2.3(b), we see that the method took on average the same time to detect the cluster properly. The false alarm rates

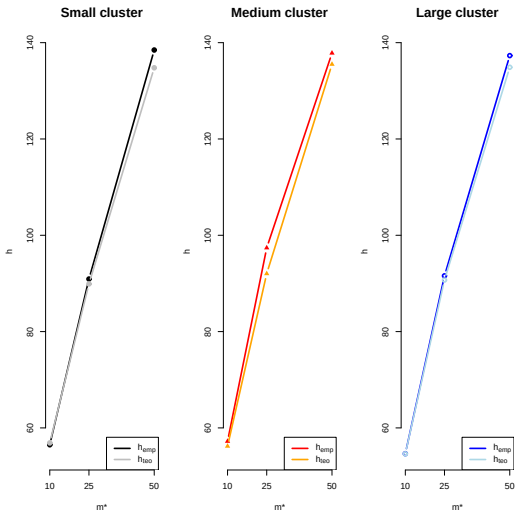


Figure 2.11: Comparative graphs of empirical and theoretical thresholds, for the clusters with different extents

FAR of the small cluster were slightly larger than the ones of the medium cluster.

When we are dealing with a widespread cluster, the mean time to detect it increased, according to Table 2.3(c). This is expected since it is more difficult to identify that a cluster is emerging in such extent area. The false alarm rates increased with respect to the scenario with a medium cluster, when we look at the window m^* equal to 10, but did not change significantly for the other windows. With this, we can conclude that the method is efficient to detect a cluster even in a large region. It is only necessary to choose a reasonable value for the window m^*. For the three cluster sizes the alarm went off in all simulations.

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*56.499*	*90.914*	*138.472*
FAR	0.181	0.157	0.130
AR	1.000	1.000	1.000
CED	8.921	8.535	11.246
$\mathbf{h_{teo}}$	*56.884*	*89.900*	*134.782*
FAR	0.174	0.168	0.131
AR	1.000	1.000	1.000
CED	8.965	9.488	11.069

(a) Small cluster

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*57.219*	*97.391*	*137.832*
FAR	0.170	0.115	0.103
AR	1.000	1.000	1.000
CED	8.997	9.964	11.205
$\mathbf{h_{teo}}$	*56.204*	*92.029*	*135.511*
FAR	0.187	0.137	0.107
AR	1.000	1.000	1.000
CED	8.877	9.644	11.113

(b) Medium cluster

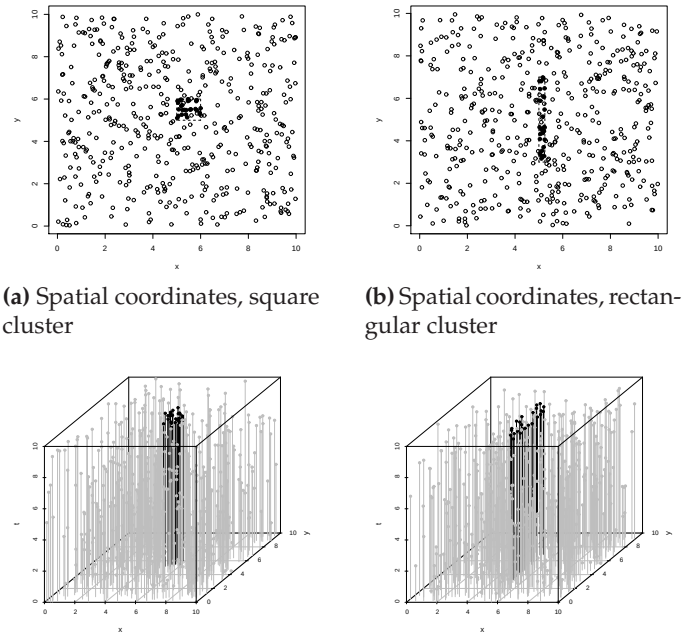
Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*54.650*	*91.582*	*137.312*
FAR	0.212	0.129	0.094
AR	1.000	1.000	1.000
CED	12.024	12.846	14.468
$\mathbf{h_{teo}}$	*54.614*	*90.705*	*134.889*
FAR	0.212	0.133	0.101
AR	1.000	1.000	1.000
CED	12.019	12.753	14.354

(c) Large cluster

Table 2.3: Results in the scenario with cluster, with different cluster extents: (a) cluster in region $[5, 5.5]^2 \times [9, 10]$; (b) cluster in region $[5, 6]^2 \times [9, 10]$; (c) cluster in region $[4, 7]^2 \times [9, 10]$

Clusters with different shapes

As the method uses a cylinder to determine the proximity between events, it is important to examine how the shape of the cluster influences its detection. For this, we compared the results of square-shaped cluster with 25 events in the region $[5,6]^2$ with the results of a cluster also with 25 events, but rectangular, in the region $[5,5.25] \times [3,7]$. In both clusters, times are uniformly distributed in $[9,10]$. The typical dataset can be seen in Figure 2.12.



(a) Spatial coordinates, square cluster

(b) Spatial coordinates, rectangular cluster

(c) Spatial and temporal coordinates, square cluster

(d) Spatial and temporal coordinates, rectangular cluster

Figure 2.12: View of the events' coordinates generated in the scenario with clusters with different shapes

The threshold values are in Tables 2.4(a) and 2.4(b), together with the results obtained for the two different clusters' shapes. Again, there was not great difference between the threshold values found, empirically and theoretically, as can be seen in Figure 2.13.

Comparing the values of *FAR* for different cluster shapes, there are no major differences between the two shapes, which means that the method is also able to identify correctly non-square clusters.

The values of *CED* were greater for the rectangular cluster, as we see in Table 2.4(b). This represents a small increase in detection time when the cluster has a shape different from the cylinder.

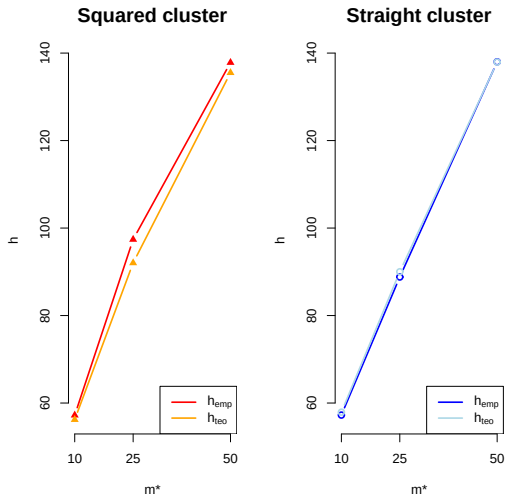


Figure 2.13: Comparative graphs of empirical and theoretical thresholds, for the clusters with different shapes

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*57.219*	*97.391*	*137.832*
FAR	0.170	0.115	0.103
AR	1.000	1.000	1.000
CED	8.997	9.964	11.205
$\mathbf{h_{teo}}$	*56.204*	*92.029*	*135.511*
FAR	0.187	0.137	0.107
AR	1.000	1.000	1.000
CED	8.877	9.644	11.113

(a) Square cluster

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*57.246*	*88.808*	*138.006*
FAR	0.194	0.195	0.115
AR	0.999	1.000	1.000
CED	11.559	11.800	13.787
$\mathbf{h_{teo}}$	*57.912*	*90.012*	*137.947*
FAR	0.173	0.175	0.115
AR	0.999	1.000	1.000
CED	11.691	11.886	13.778

(b) Rectangular cluster

Table 2.4: Results in the scenario with cluster, with different shapes: (a) square cluster in region $[5,6]^2$; (b) rectangular cluster in region $[5,5.25] \times [3,7]$

2.4.3 Scenario with cluster - Non-homogeneous Poisson Process

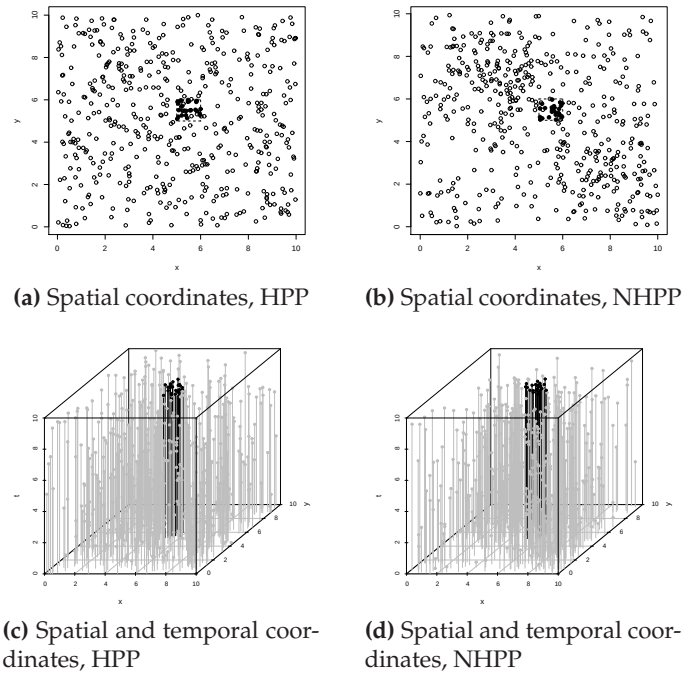
To assess if the method is able to identify a true cluster, even if there are different spatial patterns, for example due to the populational density, we generated events with different patterns of occurrence.

With the number of events belonging to the cluster fixed, the other events are generated according to a Non-homogeneous Poisson Process with intensity given by (2.18).

$$\lambda(x,y) = \phi(x,y;\mu_1,\Sigma) + \phi(x,y;\mu_2,\Sigma) \tag{2.18}$$

where $\phi(x, y; \mu, \Sigma)$ is the density function of a bivariate normal distribution at the point (x, y), with mean $\mu = (\mu_x, \mu_y)$ and correlation matrix Σ. In this case, we assumed that $\mu_1 = (3; 7)$, $\mu_2 = (8; 3)$, $\sigma_x^2 = \sigma_y^2 = 2$ and $\rho = 0$.

To form the cluster, we generated 25 events uniformly distributed in the region $[5, 6] \times [5, 6]$ and time between $[9, 10]$. The temporal and spatial coordinates of a typical dataset can be viewed in Figures 2.14(b) and 2.14(d).



(a) Spatial coordinates, HPP

(b) Spatial coordinates, NHPP

(c) Spatial and temporal coordinates, HPP

(d) Spatial and temporal coordinates, NHPP

Figure 2.14: View of the events' coordinates generated in the scenario with clusters, with the events distributed according to a homogeneous and a non-homogeneous Poisson process

The parameters assumed the same values as in the previous scenarios, including the window m^*. To evaluate the results of the scenario with NHPP, they were compared to results obtained in the scenario with the square cluster with 25 events in the region $[5, 6]^2 \times [9, 10]$, where other events are randomly distributed in the rest of region. Thus, events that do not belong to the cluster follow a Homogeneous Poisson Process with intensity λ constant and equal to 0.475.

In the graph in Figure 2.15, we see that the values of empirical and theoretical thresholds were very similar for the different values of m^*. However, compared with the thresholds of the scenario with HPP, we see that they had a small increase, reflecting the greater heterogeneity in the events' distribution.

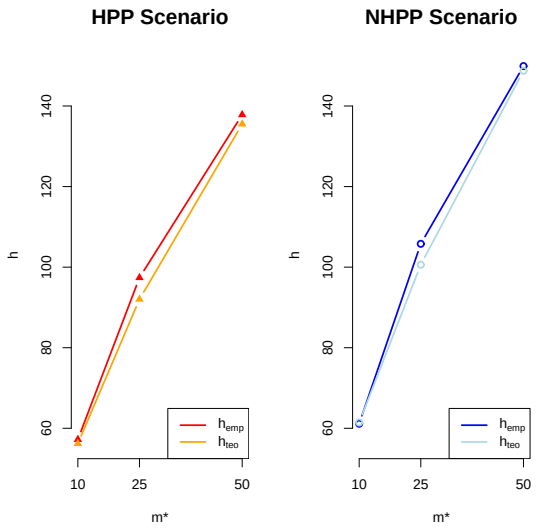


Figure 2.15: Comparative graphs of empirical and theoretical thresholds, in the scenarios with homogeneous and non-homogeneous Poisson process

In Table 2.5, we can see the difference in the results of different scenarios. In both scenarios, the alarm sounded in all simulations. The false alarm rates of NHPP were slightly higher than the rates of HPP, again as expected, since the heterogeneous distribution of the events increase the chance of a random occurrence of a conglomerate of events, not featuring a true cluster. Moreover, the values of *CED* also increased for the non-homoegeneous scenario, but again it was not a significant change with respect to the performance of the method.

Scenario with cluster			
With m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*57.219*	*97.391*	*137.832*
FAR	0.170	0.115	0.103
AR	1.000	1.000	1.000
CED	8.997	9.964	11.205
$\mathbf{h_{teo}}$	*56.204*	*92.029*	*135.511*
FAR	0.187	0.137	0.107
AR	1.000	1.000	1.000
CED	8.877	9.644	11.113

(a) HPP Cluster

Scenario with cluster			
Window m*	**10**	**25**	**50**
$\mathbf{h_{emp}}$	*61.128*	*105.788*	*149.908*
FAR	0.180	0.128	0.160
AR	1.000	1.000	1.000
CED	9.785	10.647	12.021
$\mathbf{h_{teo}}$	*61.439*	*100.608*	*148.726*
FAR	0.176	0.159	0.167
AR	1.000	1.000	1.000
CED	9.797	10.357	11.955

(b) NHPP Cluster

Table 2.5: Results in the scenario with cluster: (a) events distributed according to a homogeneous Poisson process, (b) events distributed according to a non-homogeneous Poisson process

2.5 Application

To apply the method of cumulative surface surveillance to a real dataset, we used a dataset of meningitis cases occurred in Belo Horizonte, MG, Brazil, from 2001 to 2005. The data are the spatial coordinates of occurrence of the meningitis cases, and time is the number of days since the first event, recorded on 01/01/2001. The coordinates of the 1001 events are shown in Figure 2.16.

Parameters were used with the following values: $m = 200$, $m^* = 50$, $m^{**} = 500$, chosen according to the total number of events. The false alarm probability $\alpha = 0.05$ was maintained, as well as the grid size 50×50. The critical distance is 2 km and the critical time is 30 days.

In Figure 2.17, we see a surface $w_i(x, y)$, with bandwidth τ equals to 1 km, and the spatial coordinates of the events plotted. It is observed that, when spreading the z_i score, the bandwidth chosen causes the surface to assume higher values only in a region very near the critical neighborhood, highlighted by the circle in black.

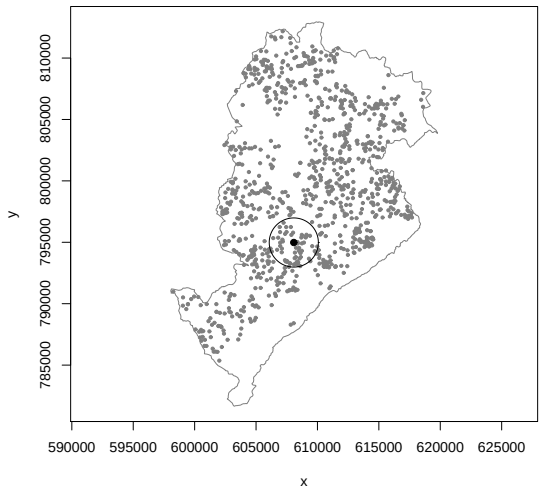


Figure 2.16: Spatial coordinates of the meningitis cases in Belo Horizonte

To determine the threshold h, we made 1000 permutations of the times, keeping the spatial coordinates fixed. In Figure 2.18, we see the histogram of the values of the maxima of the cumulative surface from each permutation, with the density of the Gumbel distribution and the threshold values obtained empirically and theoretically. Since the threshold values h_{emp} and h_{teo} are very close, the results of monitoring won't change, as discussed in Section 2.4. Thus, we chose the threshold value $h_{emp} = 88.344$ to do the surveillance.

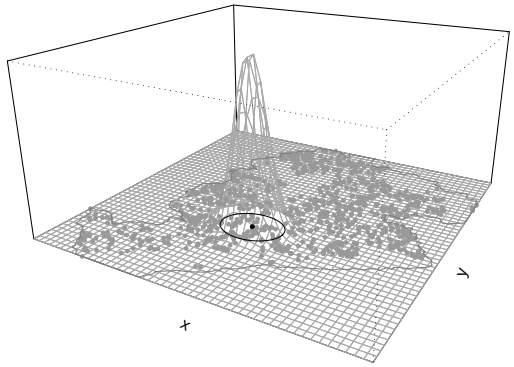


Figure 2.17: View of a surface $w_i(x, y)$, with the spatial coordinates of the meningitis cases in Belo Horizonte

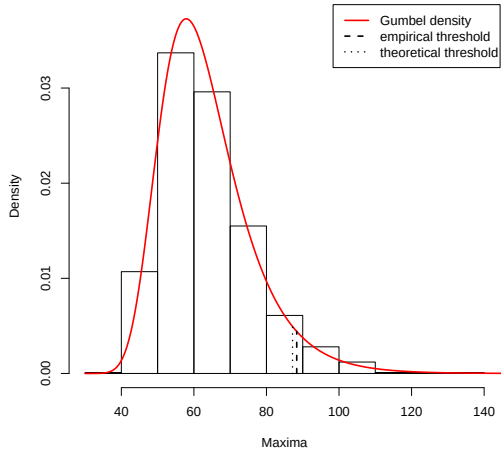


Figure 2.18: Histogram of the maxima of the cumulative surfaces obtained in the permutations, for the meningitis cases

The surveillance was made for all possible events, excluding only the first $m = 200$ events for the sake of scores' stabilization. The alarm sounded when the event 609 was being monitored, with coordinates $(x, y) = (603374.45, 798618.15)$ on 11/10/2003.

In Figure 2.19, we see the cumulative surface when the alarm sounded, with the time corresponding to the threshold highlight in red. In Figure 2.20, we see the cumulative surface in contour levels, where the surface value increases as the colors go from yellow to red, with the events that occurred before the 609-th event (highlighted in blue), and the threshold height corresponding to the highlighted green line, delimiting the region of a possible cluster.

One way to identify, at the time the alarm sounds, which events belong to the cluster is to analyze the value of the z_i scores. In Figure 2.21 we can see the time series of the scores. Highlighted in red is the score corresponding to the 609th event.

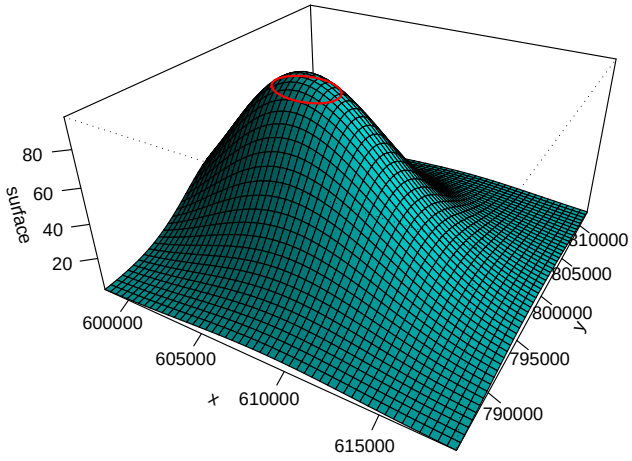


Figure 2.19: Cumulative surface at the moment that the alarm sounded, for the meningitis cases

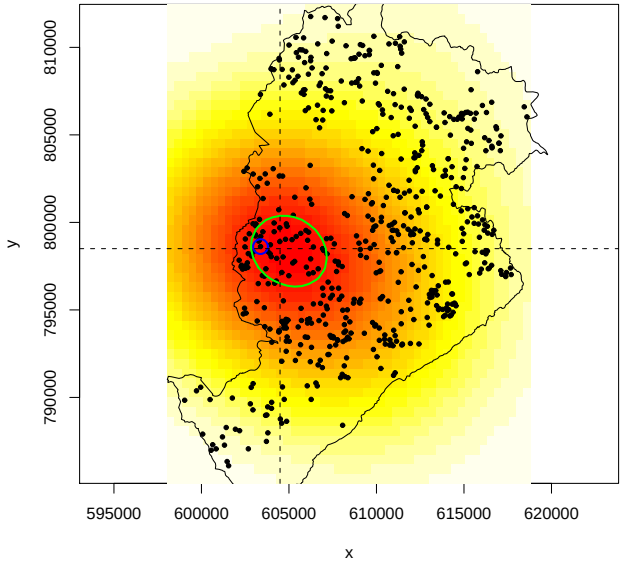


Figure 2.20: View of the cumulative surface in contour levels, for the meningitis cases

In Figure 2.22 we can see again the surface contour levels with the region of the possible cluster highlighted in green. The symbols of the events are plotted with radii proportional to the scores z_i, as well as their colors, ranging from the lowest values in gray to the largest ones in dark blue. In Figure 2.22(a) we see all the events that happened before the alarm went off. In Figure 2.22(b) we see only the 50 last events that happened before the 609-th event. With this second graphic it is possible to identify those events that contributed to raise the surface level at this moment, without the oldest events. Looking at Figure 2.22(b), we can identify a possible emerging cluster on the northwest region of the city, near its left border.

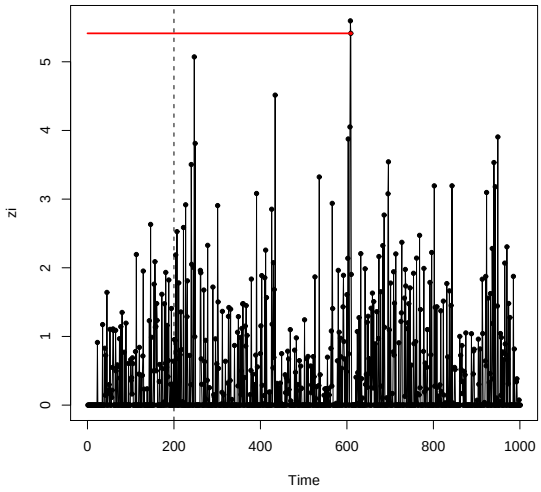
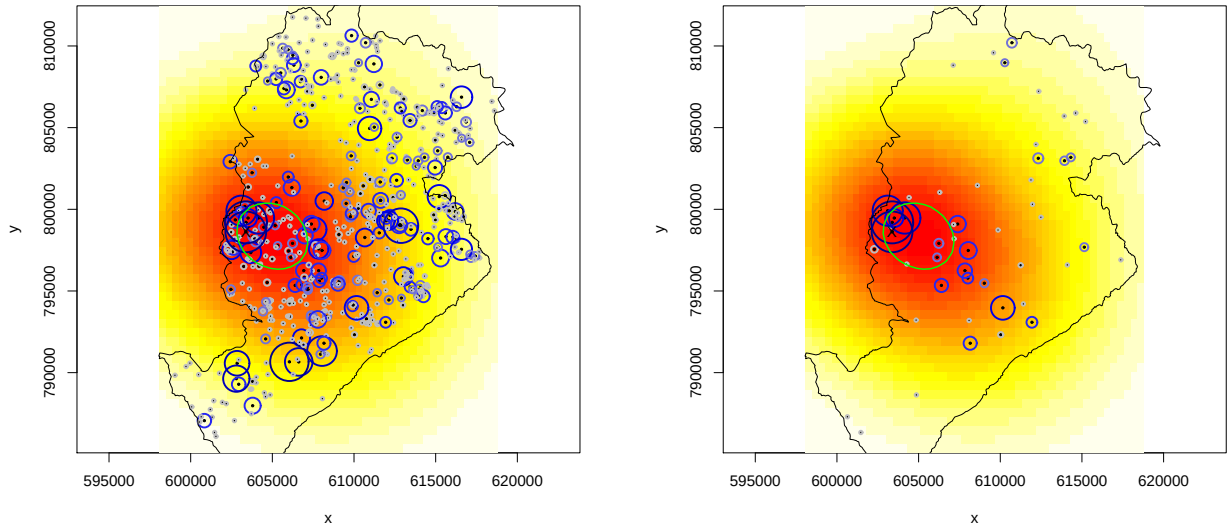


Figure 2.21: Time series of the z_i scores, of the meningitis cases



(a) Plotting of events from the first to the 609-th

(b) Plotting of the 50 events prior to the 609-th

Figure 2.22: View of the events according with the z_i values, for the meningitis cases

We can observe that the alarm didn't sound when there is only a high score, but when the high score was observed in a region where the surface was already at a high level. However, identifying the cluster just as the region where the value of the surface is above the threshold h is not appropriate. It is necessary to analyze simultaneously the z_i scores to identify which events actually led to an increase in the surface level and made the alarm sound, and thus formed a cluster.

2.6 Discussion

To apply the cumulative surface method, it is only necessary to determine the values of some parameters according to the size of the database and range of the study area, requiring no information about the population at risk. This makes the method useful for users with little knowledge of statistics, such as departments of public health and security, with great interest in monitoring spatial-temporal events prospectively, in order to guide actions to combat clusters as they occur.

The simulation results under different scenarios were satisfactory. The false alarm probability stayed within the limit set and the mean time until the cluster is detected correctly was reasonable. This shows that the method detects clusters even with different intensities, extents and shapes. We also had great results for the scenario in which the events have a heterogeneous distribution, like the non-homogeneous poisson process.

The application to a real database shows that the visualization of the cumulative surface, together with the local scores, identifies the region of the cluster and the events that belong to him. However, it is not enough to look for a conglomerate of high values of z_i scores to identify a cluster, since they can be separated by a large time interval. Therefore, the advantage of the method is to accumulate information over time using a window set by the user to localize a large number of events that are happening close in space and in a short interval of time.

Another advantage of the method is to identify the space-time interaction levels all over the study area. For example, assume that the combined surface area exceeds the threshold h in a small region and the alarm goes off, signing this location as a cluster. However, there may be other clusters about to emerge in the map, with high surface values. With the view of the surface, the user is also able to identify these other clusters and take appropriate action. Furthermore, the visualization in contour levels allows a completely flexible shape for the cluster region, it is just necessary to choose an appropriate value for the bandwidth. In the results presented in this paper, we have chosen bandwidth values according to the critical distance. This is due to the fact that we obtained high values estimating the bandwidth automatically, so the surfaces would decay very softly and the peaks corresponding to localized clusters wouldn't be distinguishable. Therefore, determining a better way to estimate an optimal bandwidth is a future research topic.

Referências Bibliográficas

Assunção, R. and T. Correa (2009). Surveillance to detect emerging space-time clusters. *Computational Statistics and Data Analysis 53*(8), 2817–2830.

Diggle, P., B. Rowlingson, and T. Su (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics 16*(5), 423–434.

Höhle, M. (2007). `surveillance`: An R package for the monitoring of infectious diseases. *Computational Statistics 22*(4), 571–582.

Härdle, W. (1990). *Smoothing Techniques*. Springer-Verlag.

Jacquez, G. M. (1996). A k nearest neighbour test for space-time interaction. *Statistics in Medicine 15*(18), 1935–1949.

Knox, E. G. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 13*(1), 25–30.

Kulldorff, M. (1999). The Knox method and other tests for space-time interaction. *Biometrics 55*(2), 544–552.

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

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Med 2*(3), e59.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research 27*(2), 209–220.

Marshall, J. B., D. J. Spitzner, and W. H. Woodall (2007). Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time. *Statistics in Medicine 26*(7), 1579–1593.

Neill, D. B., A. W. Moore, M. Sabhnani, and K. Daniel (2005). Detection of emerging space-time clusters. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 218–227.

Piroutek, A., R. Assunção, and T. Paiva (2010). Space-time surveillance: a critical reavaliation. *Statistics in Medicine*. submitted.

Piterbarg, V. I. (1996). *Asymptotic methods in the theory of Gaussian processes and fields*, Volume 148 of *AMS Transl. of Math. Monographs*. Providence, R.I.

Rodeiro, C. and A. Lawson (2006). Monitoring changes in spatio-temporal maps of disease. *Biometrical Journal 48*(3), 463–480.

Rogerson, P. A. (2001). Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society A 164*(1), 87–96.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications*, Volume 104 of *Monographs on statistics and applied probability*. CRC Press.

Simões, T. C. and R. M. Assunção (2005). Sistema de vigilância para detecção de interações espaço-tempo de eventos pontuais. In *GEOINFO 2005 - VII Simpósio Brasileiro de Geoinformática*, pp. 281–291.

Sonesson, C. and D. Bock (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society A 166*(1), 5–21.

Woodall, W. H., J. B. Marshall, M. D. Joner Jr., S. E. Fraker, and A. G. Abdel-Salam (2008). On the use and evaluation of prospective scan methods in health-related surveillance. *Journal of the Royal Statistical Society, Series A 171*(1), 223–237.