

Taynãna César Simões

**Sistema de vigilância para detecção  
de interação espaço-tempo  
de eventos pontuais  
via superfícies acumuladas**

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

Orientador: Renato Martins Assunção

Universidade Federal de Minas Gerais  
Belo Horizonte, 5 de Abril de 2006

# **Livros Grátis**

<http://www.livrosgratis.com.br>

Milhares de livros grátis para download.

# Resumo

Sistemas de Vigilância têm importantes aplicações em diversas questões práticas. Frequentemente, profissionais de várias áreas de atuação são chamados a intervir de forma rápida e precisa em regiões que pareçam possuir uma taxa de ocorrência de eventos de certa natureza acima do esperado, num dado período de tempo. O aumento repentino na incidência de uma doença grave numa região, por exemplo, pode ser sinônimo da iminência de uma epidemia que deve ser combatida da forma mais eficaz.

A necessidade de realizar estudos de natureza prospectiva na detecção de conglomerados (clusters) espaço-temporais tem sido o foco de alguns pesquisadores. Dentre eles, utilizando uma estatística de detecção local de conglomerados espaço-temporais (escore local de Knox) e métodos de soma acumulada (CUSUM), Rogerson (2001) propõe um sistema de vigilância que detecte clusters existentes no momento da análise (ativos), dentre um conjunto de eventos pontuais de coordenadas  $x$  e  $y$  no espaço e tempo de ocorrência  $t$ .

Resumidamente, Rogerson (2001) estabelece uma estatística de monitoramento, que vai sendo acumulada à medida que um novo evento se torna disponível na análise. Quando uma soma ultrapassa um limiar pré-determinado, um alarme deve ser soado a fim de alertar o usuário de que algo diferente do esperado está ocorrendo, possivelmente a formação de um cluster espaço-temporal.

No entanto, dois problemas verificados nessa técnica são: o método não delimita a região (ões) do(s) cluster(s). Além disso, ele não é capaz de identificar quais os eventos, dentre os mais recentes, pertencem ao(s) cluster(s). Assim, nosso objetivo é principalmente adaptar a técnica apresentada por Rogerson (2001), com o intuito de detectar o cluster eventualmente existente e que teria disparado o alarme, através da visualização da posição em que esse foi formado. A idéia apresentada é monitorar as novas observações, não através de uma soma acumulada, mas através de superfícies acumuladas obtidas através da distribuição dos escores locais de Knox no espaço, com funções de kernel bivariadas. Nós estudamos o desempenho de nosso método com métodos Monte Carlo e ilustramos seu uso com dados de Meningite Meningocócica em Belo Horizonte, no período de 1998 a 2000.

**Palavras-chave:** Sistema de Vigilância, conglomerados espaço-temporais, superfícies acumuladas.

# Sumário

Resumo	ii
<b>1 Introdução</b>	<b>1</b>
1.1 Organização da Dissertação	2
<b>2 Conceitos e Definições Gerais</b>	<b>3</b>
2.1 Processos Pontuais	3
2.2 Conglomerados Espaço-Temporais	4
2.3 Estimacão de Densidade por Kernel	5
2.4 Método de Soma Acumulada (CUSUM)	8
2.5 Vigilância Estatística	11
<b>3 Proposta de Trabalho</b>	<b>13</b>
3.1 Problemas com o método do Rogerson (2001)	13
3.2 Propostas	14
Referências Bibliográficas	14
Anexo 1: Artigo - <i>Revista iP - Informática Pública</i>	17
Anexo 2: Artigo - <i>International Journal of Health Geographics</i>	18

# Capítulo 1

## Introdução

Conhecer a distribuição de eventos no espaço é de interesse em várias áreas do conhecimento. Se os dados são pontuais e os tempos de ocorrência foram registrados, um sistema de vigilância pode ser desenvolvido para detectar conglomerados espaço-temporais. Esse tipo de sistema é de grande utilidade em várias áreas do conhecimento. Na área da saúde, em que a ocorrência de doenças são registradas freqüentemente no espaço e no tempo por exemplo, é de suma importância ter um sistema que detecte o risco repentino de epidemia numa dada região. Um bom sistema de monitoramento deve alertar sobre o aumento acima do esperado do número de ocorrências de um evento no espaço e no tempo, tão rápido quanto possível, minimizando o número de alarmes falsos.

No contexto de sistemas de vigilância, muito se tem visto sobre estudos retrospectivos nos quais a análise é feita para um número fixo de eventos passados. Esses podem ser usados, por exemplo, na estimação da prevalência de uma doença ou para comparar padrões de doenças em diferentes regiões, por teste de hipóteses. No entanto, cada vez mais é de interesse dos pesquisadores os estudos prospectivos em que há uma análise repetida de dados acumulados ao longo do tempo, tratando de uma série de eventos seqüencialmente, com o objetivo de detectar rapidamente qualquer mudança inesperada.

Essa necessidade de realizar estudos prospectivos na detecção de conglomerados espaço-temporais tem sido foco de alguns pesquisadores tais como Raubertas (1989), Rogerson (1997), Järpe (1999), Kulldorff (2001), Kulldorff et al (2005), Rogerson (2001), entre outros. Uma revisão de toda a literatura relacionada ao assunto é apresentada por Sonesson e Bock (2003).

Em particular, trabalhamos com o artigo apresentado por Rogerson (2001), que utiliza uma estatística de detecção de conglomerados espaço-temporais local (estatística local de Knox) e métodos de soma acumulada (CUSUM), para detectar clusters espaço-temporais emergentes. De forma geral, as estatísticas locais padronizadas (escores locais de Knox) que excedem o valor esperado sob a hipótese de não interação espaço-tempo, são acumuladas através de uma soma à medida que os eventos são observados. Caso essa soma exceda um limiar predeterminado, há evidência a favor da hipótese de interação espaço-tempo, indicando

a provável formação de clusters.

No entanto, dois problemas podem ser verificados no método proposto por Rogerson (2001). O primeiro problema é que os eventos que disparam o alarme podem não pertencer ao eventual cluster. O segundo é que o método detecta a presença de clusters, mas não os identifica, dando sua posição e extensão no espaço e tempo. Desta forma, esse trabalho procura adaptar a técnica apresentada por Rogerson (2001), com o intuito de identificar o cluster eventualmente existente e que seria a principal causa do disparo do alarme. Acreditamos ainda que a técnica proposta permite isolar eventos que não pertençam ao cluster, mas que contribuem para fazer o alarme soar. A idéia é monitorar os novos eventos, não através de uma soma acumulada, mas através de superfícies acumuladas. De forma resumida, a cada novo evento, é calculado o escore local de Knox que é distribuído no espaço através de uma densidade de kernel, gerando uma superfície na região do evento. Estas superfícies são acumuladas e geram uma saliência pronunciada em torno de um eventual cluster.

O método desenvolvido por Rogerson (2001), bem como a metodologia proposta, deverão ser aplicados a dados reais referentes à incidência de Meningite Meningocócica em Belo Horizonte, entre os anos de 1998 a 2000. Ressalta-se ainda que, neste texto, o termo conglomerado(s) será, na maioria das vezes, referido como cluster(s), termo em inglês largamente empregado na literatura.

## 1.1 Organização da Dissertação

Este texto é constituído de duas partes. Na primeira, apresentamos uma revisão da literatura, composta por alguns conceitos e definições gerais para detecção de conglomerados espaço-temporais de eventos pontuais, bem como tópicos relevantes no estudo e aplicação de sistemas de vigilância. Abordamos também a proposta de trabalho.

O capítulo 2 aborda conceitos de processos pontuais, conglomerados espaço-temporais, estimação de densidade por kernel, somas acumuladas (CUSUM) e vigilância estatística de processos. O capítulo 3 mostra os problemas verificados no sistema de vigilância apresentado por Rogerson (2001), bem como a proposta de trabalho sugerida para este texto.

Na segunda parte, dois artigos resultantes do trabalho desenvolvido são apresentados:

- Artigo submetido à edição especial da *Revista iP - Informática Pública* (volume 8, número 1, março/agosto de 2006), cujo tema é a geoinformática aplicada ao setor público.
- Artigo a ser submetido ao *International Journal of Health Geographics*. A versão final estará em inglês.

Maior ênfase deve ser dada ao segundo artigo, pois é uma versão revisada do primeiro e apresenta conteúdo mais amplo, com aplicações específicas do método proposto.

Os Anexos 1 e 2 trazem o artigo submetido à *Revista iP - Informática Pública* e o artigo a ser submetido ao *International Journal of Health Geographics*, respectivamente.

# Capítulo 2

## Conceitos e Definições Gerais

### 2.1 Processos Pontuais

Um conceito de grande importância na análise de fenômenos espaciais é a dependência espacial entre as observações. As inferências nesse tipo de dado não são tão eficientes como em amostras independentes. Existe uma perda do poder explicativo, dado que as variâncias maiores para as estimativas levam a níveis menores de significância em testes de hipóteses e a um pior ajuste para os modelos estimados. Assim, considera-se os dados espaciais não como um conjunto de amostras independentes, mas como uma realização de um processo estocástico. Nesse processo, todas as observações são utilizadas conjuntamente para descrever o padrão espacial do fenômeno estudado.

Usualmente, os dados espaciais podem se caracterizar em três grandes grupos: processos pontuais (eventos ou padrões pontuais); variação contínua (superfícies contínuas); e variação discreta (áreas com contagens e taxas agregadas). Em particular, eventos ou padrões pontuais são fenômenos cujas ocorrências são identificadas como pontos localizados no espaço. São exemplos de processos pontuais a localização de crimes, a ocorrência de doenças e a localização de espécies vegetais.

Tecnicamente, processos pontuais são definidos como um conjunto de pontos distribuídos em um terreno, cuja localização foi gerada por um mecanismo estocástico. O conjunto desses pontos é denominado padrão espacial de pontos e um ponto em particular de evento. O objetivo é estudar a distribuição espacial dos mesmos, testando hipóteses sobre o padrão observado: se ele é aleatório, se apresenta aglomerados, regularidade na distribuição ou outras hipóteses de interesse.

Na Figura 2.1, existem dois padrões espaciais de pontos que parecem ser estritamente diferentes. A primeira figura não mostra nenhuma estrutura óbvia e deve ser considerada como um padrão completamente aleatório. Por outro lado, a segunda figura evidencia uma clara formação de aglomerados, que requer alguma explicação apropriada.

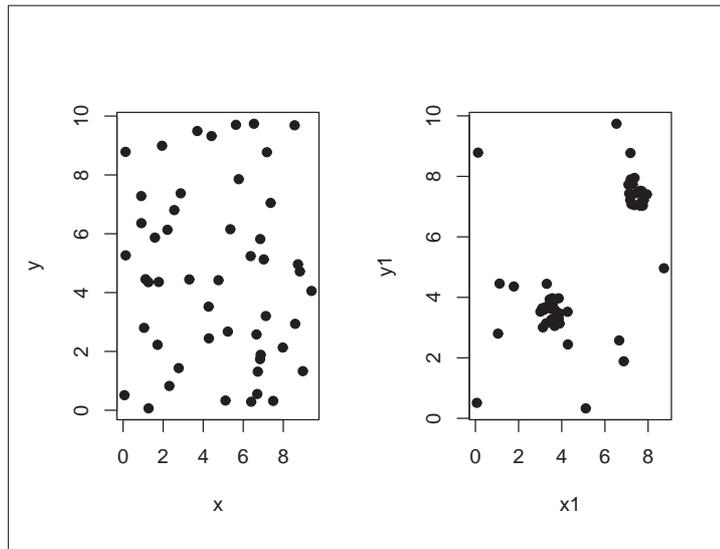


Figura 2.1: Exemplos de processos pontuais.

O interesse maior é encontrar sub-regiões de uma área em estudo com maior probabilidade de ocorrência de eventos ou de maior intensidade. No modelo de completa aleatoriedade, considera-se que os eventos têm igual probabilidade de ocorrência em toda a região e que suas posições são independentes umas das outras. Essa formulação permite estabelecer uma base de comparação entre uma distribuição completamente aleatória e os eventos observados.

## 2.2 Conglomerados Espaço-Temporais

Uma situação de grande interesse prático, é quando o tempo de ocorrência dos eventos é registrado. Assim, o interesse é verificar se espaço e tempo interagem, isto é, se eventos aglomeram no espaço e no tempo simultaneamente.

Muitos estudos avaliam se existe correlação puramente espacial ou puramente temporal de eventos. É comum encontrar substancial variação espacial refletindo a distribuição geográfica não-uniforme da população de risco ou dos fatores ambientais, assim como é usual encontrar aglomerados temporais devido à efeitos sazonais ou tendências de crescimento ou decrescimento acentuado da taxa de ocorrência dos eventos, ao longo do tempo. No entanto, quando as informações tanto de espaço quanto de tempo estão disponíveis, pode-se testar a existência de aglomerados no espaço e no tempo simultaneamente, após ajustar por possíveis variações puramente espaciais ou puramente temporais. O objetivo é testar se casos que estão próximos no espaço tendem a estar também próximos no tempo. Se isto ocorre, pode-se dizer que existem aglomerados espaço-temporais ou que os dados exibem interação espaço-tempo.

As ocorrências de eventos no tempo e no espaço são registradas em vários problemas aplicados. Na área da saúde, por exemplo, pode-se observar o dia e a localização de morte de um indivíduo ou o dia da eclosão e a área geográfica de novos casos de uma certa doença. Na análise de crimes, registram-se delitos por hora e data de ocorrência e a área em que eles ocorreram dentro de uma cidade. Em Ecologia, há o interesse no padrão espacial de espécies de fauna e flora, mas também onde e como distribuições geográficas particulares mudam com o tempo. Em Astronomia, existe interesse na distribuição espacial de estrelas e galáxias, bem como na questão de onde e quando esses padrões espaciais mudam com o tempo; dentre outros exemplos.

Na análise de conglomerados espaço-temporais, os dados em geral consistem de um conjunto de eventos no espaço euclidiano bidimensional, dentro de uma região poligonal no espaço e entre limites temporal superior e inferior. De uma forma geral, busca-se detectar a existência de um padrão de conglomerados espaciais (clusters), através da constatação de um número acima do esperado de casos excessivamente próximos no espaço e no tempo, considerando uma distribuição de referência. Se um padrão de eventos pontuais observados apresentar desvios significativos do comportamento esperado para essa distribuição, há a indicação da existência de uma distribuição espaço-tempo diferente da distribuição de referência, que merece ser objeto de maior análise.

## 2.3 Estimação de Densidade por Kernel

Uma maneira de estimar aproximadamente a densidade de um conjunto de dados é através da construção de um histograma das frequências das observações. No entanto, esse procedimento não estima muito bem os valores mais extremos. Outra maneira de se estimar a densidade de probabilidade é utilizando superfícies de kernel. Resumidamente, o método ajusta curvas com base nos pesos que cada evento tem no conjunto de dados, em relação a observações centrais. Estima-se a densidade em pontos determinados (pontos de referência) usando os pontos empiricamente observados. A Figura 2.2 mostra a estimativa da densidade de kernel de um processo pontual.

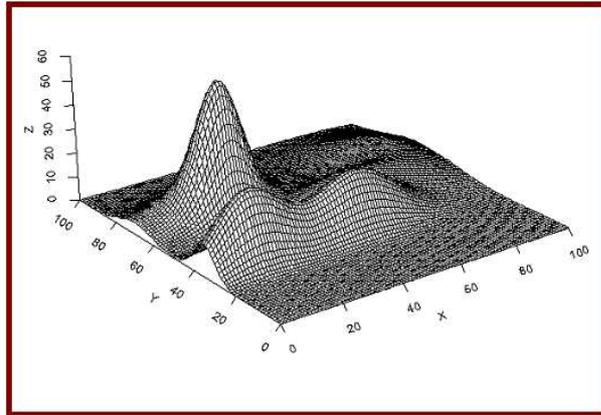


Figura 2.2: Visualização de densidade de kernel.

Para cada evento  $u_i$  existe uma função de kernel centrada em  $u_i$ , dada por  $K_\tau(\bullet - u_i)$ , que é simétrica e integra 1. O estimador de kernel pode ser calculado para pontos  $u$  diferentes do ponto central, através de diferentes funções de probabilidade. Existe um parâmetro  $\tau$  que regula o grau de suavidade das curvas, chamado "largura de banda" (*bandwidth*).

O estimador de kernel pode também ser entendido como uma estimador de intensidade de eventos numa região. Para isso, ajusta-se uma função bidimensional sobre os eventos observados, compondo uma superfície suave cujo valor será proporcional à intensidade de eventos por unidade de área. Esta função realiza uma contagem de todos os pontos dentro de uma região de influência, ponderando-os pela distância de cada um à localização de interesse, como mostrado na Figura 2.3.

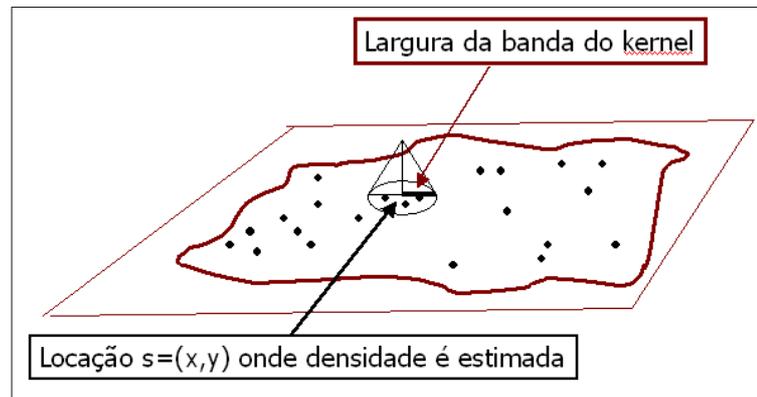


Figura 2.3: Estimador de intensidade de distribuição de pontos - kernel triangular.

A grande desvantagem desses estimadores são a forte dependência do raio de busca (largura de banda) e a excessiva suavização da superfície, que pode em alguns casos, esconder variações locais importantes.

Suponha que  $(u_1, u_2, \dots, u_n)$  são as posições de  $n$  eventos observados em uma região e que  $u$  represente uma localização genérica cujo valor da densidade se quer estimar. O estimador de intensidade é computado a partir dos  $m$  eventos  $(u_i, \dots, u_{i+m-1})$  contidos num raio de tamanho  $\tau$  em torno de  $u$  e da distância  $d$  entre a posição de referência  $u$  e a  $i$ -ésima amostra, a partir de funções cuja forma geral é dada pela equação 2.1.

$$K_\tau(u) = \frac{1}{\tau^2} \sum_{i=1}^m K^* \left( \frac{d(u_i, u)}{\tau} \right), d(u_i, u) \leq \tau \quad (2.1)$$

Os parâmetros básicos são então: (a) um raio de influência ( $\tau > 0$ ) que define a vizinhança do ponto e controla o "alisamento" da superfície gerada; (b) uma função de estimação  $K^*$  com propriedades de suavização do fenômeno. Dessa forma, o valor da função intensidade em cada ponto de referência depende de dois fatores: (a) do número de eventos que caem dentro do círculo de raio  $\tau$ ; (b) do quão distantes os eventos contidos no círculo estão do ponto central.

Quanto mais eventos caírem dentro do círculo e quanto mais próximos do ponto central estes eventos estiverem, maior será o valor da função intensidade neste ponto.

## Raio de influência (largura de banda)

O raio de influência (largura de banda) define a área centrada no ponto de estimação  $u$  que indica quantos eventos  $u_i$  contribuem para a estimativa da função intensidade  $K_\tau$ . Aumentar a largura de banda, implica no aumento da suavização na estimativa. Assim se  $\tau \rightarrow 0$ , há muito ruído na representação dos dados; se  $\tau$  é pequeno, há menos ruído na estimativa; se  $\tau$  é grande, a variação espacial da densidade estimada é bem suave; e se  $\tau \rightarrow \infty$ , a estimativa é tão suavizada que extrapola a forma do kernel escolhido.

Há diferenças de opinião sobre quão ampla a largura de banda deva ser. Uma abordagem simples é dada por Härdle(1999), que sugere uma expressão que minimiza o erro quadrático médio das estimativas de kernel.

O alisamento feito para uma distribuição de valores contínuos de um vetor  $Z$  e raio  $\tau$  é obtido como:

$$\tau = 1.06 * \min \left\{ sd(Z), \frac{iqr(Z)}{1.34} \right\} * n^{\{-1/5\}} \quad (2.2)$$

onde  $sd(Z)$  é o desvio-padrão da variável  $Z$  sendo interpolada,  $iqr(Z)$  é a distância inter-quatílica de  $Z$  e  $n$  é o tamanho de amostra.

## Função de estimação

Cada função  $K^*$  produz diferenças suaves na forma da superfície ou do contorno interpolado. As técnicas pesam diferentemente os pontos no círculo circunscrito em torno da posição de referência. A distribuição normal (gaussiana) pesa os pontos dentro do círculo de forma que pontos mais próximos são pesados mais intensamente comparados aos mais afastados. A distribuição uniforme pesa todos os pontos dentro do círculo igualmente. A função quártica pesa pontos próximos mais do que pontos distantes, mas o decréscimo é gradual. A função triangular pesa pontos próximos mais do que pontos distantes dentro do círculo, mas o decréscimo é mais rápido. O uso de qualquer uma destas funções depende do interesse em se pesar pontos próximos em relação a pontos distantes. Usar uma função de kernel que tenha uma grande diferença nos pesos entre pontos próximos e pontos distantes (por exemplo, o triangular) tende a produzir variações menores dentro da superfície do que as funções que são pesadas mais uniformemente (por exemplo, a distribuição normal, a uniforme ou a quártica). Essas últimas tendem a suavizar mais a distribuição.

## 2.4 Método de Soma Acumulada (CUSUM)

A maneira tradicional de detectar uma situação fora de controle, isto é, detectar um possível desvio do valor esperado é utilizar critérios de tendências (run-test) no contexto dos gráficos de Shewhart. Esses gráficos não acumulam as informações das amostras anteriores mas fazem uso dos valores amostrais plotados sucessivamente para uma tomada de decisão, através dos testes de seqüência. Sua inabilidade em detectar desvios moderados levou ao desenvolvimento de modelos de gráficos de controle que acumulassem informações das amostras coletadas sucessivamente, capazes de detectar desvios pequenos a moderados e de fácil utilização. Um destes gráficos é o de soma acumulada (Cumulative Sum Control Charts - CUSUM) que detecta pequenas mudanças na distribuição da característica em estudo, mantendo grande controle sobre o processo e dando uma estimativa do novo nível da característica monitorada, após uma mudança.

A técnica de soma acumulada pode ser aplicada tanto na construção do gráfico CUSUM para observações individuais como para observações amostrais das médias de subgrupos. No caso de observações individuais, a estatística utilizada é a soma acumulada dos desvios de cada valor individual com relação à medida dada pela hipótese que está sendo testada. Os gráficos CUSUM são mais eficientes que os gráficos de Shewhart para detectar pequenas e contínuas mudanças do processo, da ordem de até 1.5 desvios-padrão, dado que combinam as informações de várias amostras. Além disso, são particularmente mais eficazes com amostras seqüenciais de tamanho  $n = 1$ , ou seja, para cada período de tempo existe apenas uma observação.

## A estatística de Soma Acumulada

O procedimento de soma acumulada começa com o cálculo dos desvios do valor nominal (diferença entre o valor observado da média amostral e o valor médio esperado  $\mu_0$ ). A soma acumulada  $C_i$  para o  $i$ -ésimo período é a soma de todos os desvios do valor nominal desde o período 1 até o período  $i$ , dada por:

$$C_i = \sum_{j=1}^i (X_j - \mu_0) = (X_i - \mu_0) + C_{i-1}, \quad i \geq 1 \quad (2.3)$$

onde  $X_j$  é a  $j$ -ésima observação de um gráfico CUSUM com observações individuais. Se o processo permanece sob-controle para o valor médio esperado  $\mu_0$ , a soma acumulada descreve um caminho aleatório com média zero. Porém, se a média muda para algum valor  $\mu_1 > \mu_0$  ( $\mu_1 < \mu_0$ ), uma tendência positiva (negativa) será vista na soma  $C_i$ . Se os pontos plotados formarem uma tendência para cima ou para baixo, deve-se considerar este fato como uma evidência de que a média do processo mudou e deve-se buscar as possíveis razões para a mudança.

## O Gráfico de Controle CUSUM Tabular

O gráfico de controle CUSUM Tabular é um procedimento que utiliza o algoritmo de soma acumulada para calcular as somas acumuladas unilaterais (a soma é apenas positiva ou apenas negativa). Através do gráfico, as somas são comparadas com um intervalo de decisão  $H$ . Se o valor da soma for maior que este intervalo, o processo é considerado fora-de-controle.

Existem vários métodos para a construção de um gráfico de controle CUSUM Tabular. Segundo Montgomery (2000), seja  $X_i$  cada observação do processo controlado suposto sob-controle. Considera-se que os dados coletados tenham uma distribuição normal com média  $\mu_0$  e desvio-padrão  $\sigma$ .

O CUSUM Tabular utiliza duas estatísticas unilaterais  $C_i^+$  (estatística superior) para detectar mudanças positivas e  $C_i^-$  (estatística inferior) para detectar mudanças negativas. Esses planos de decisão são caracterizados por um único parâmetro denominado intervalo de decisão ou limite de controle, representado por  $H = \pm h$ .

A estatística  $C_i^+$  é a soma acumulada dos desvios positivos, isto é, desvios acumulados de  $\mu_0$  que estão acima do valor alvo.  $C_i^-$  é a soma acumulada dos desvios negativos, ou seja, desvios acumulados de  $\mu_0$  que estão abaixo do valor alvo. Estas estatísticas unilaterais ( $C_i^+$  e  $C_i^-$ ) são denominadas respectivamente por CUSUM superior e inferior e são calculadas como:

$$C_i^+ = \max[0, X_i - (\mu_0 + K) + C_{i-1}^+] \quad (2.4)$$

$$C_i^- = \max[0, (\mu_0 - K) - X_i + C_{i-1}^-] \quad (2.5)$$

onde  $X_i$  é a observação controlada no tempo  $i$ ,  $\mu_0$  é a média da amostra e  $K$  é um valor de referência (valor de compensação ou folga) e é aproximadamente a metade do valor que se tem interesse em detectar rapidamente, entre o valor médio esperado  $\mu_0$  e o valor da média fora de controle  $\mu_1$ . Os valores iniciais  $C_i^+$  e  $C_i^-$  são arbitrariamente iguais a zero. Se, ao longo dos cálculos, forem encontrados valores negativos para  $C_i^+$  ou valores positivos para  $C_i^-$ , é necessário substituí-los por zero.

Se a mudança é expressa em unidades de desvio-padrão, quando  $\mu_1 = \mu_0$ ,  $K$  representa a metade da magnitude desta mudança, ou seja:

$$K = \frac{\delta \sigma}{2\sqrt{n}} = \frac{|\mu_1 - \mu_0|}{2} \quad (2.6)$$

onde  $\delta$  é o tamanho da mudança que se deseja detectar em unidades de desvio-padrão;  $\sigma$  o desvio-padrão;  $\mu_0$  o valor médio esperado e  $\mu_1$  o valor da média fora-de-controle. Quanto menor o valor de  $K$ , menor será a faixa de variação que o gráfico será capaz de detectar e maior será a sensibilidade do gráfico. Se  $C_i^+ > H$  ou  $C_i^- < H$ , então o processo é considerado fora-de-controle.

Para o gráfico CUSUM Tabular padronizado (utilizando o escore  $z_i$ ), o algoritmo de soma acumulada é definido como:

$$C_i^+ = \max[0, z_i - k + C_{i-1}^+] \quad (2.7)$$

$$C_i^- = \max[0, -k - z_i + C_{i-1}^-] = \min[0, k + z_i + C_{i-1}^-] \quad (2.8)$$

O gráfico CUSUM Tabular é projetado pela escolha de valores razoáveis para o intervalo de decisão  $H$  e o valor de referência  $K$ . Montgomery (2000) recomenda que o melhor maneira de selecionar esses valores é defini-los conforme as equações abaixo:

$$K = k \frac{\sigma}{\sqrt{n}} \quad (2.9)$$

$$H^+ = h \frac{\sigma}{\sqrt{n}} \quad (2.10)$$

$$H^- = -h \frac{\sigma}{\sqrt{n}} \quad (2.11)$$

onde  $k$  e  $h$  são constantes (frequentemente  $k = 0.5$  e  $h = 4$  ou  $h = 5$ , respectivamente) e  $\sigma$  é o desvio-padrão dos dados. Estes valores de  $k$  e  $h$  são comumente usados porque produzem um gráfico CUSUM que tem boas propriedades do ARL (número médio de eventos observados até que uma mudança ocorra), com uma mudança de cerca de  $1\sigma$  na média do processo.

## Average Run Length - ARL

Uma forma de avaliar o desempenho do gráfico de controle se relaciona à sensibilidade para detectar desvios na estatística que está sendo monitorada. Essa sensibilidade pode ser medida pelo número de amostras coletadas até que o gráfico sinalize a ocorrência de uma mudança. Para cada amostra coletada, um ponto é plotado no gráfico para monitorar variações nas características de um produto ou serviço. O número de amostras (pontos) desde o início do processo até o instante em que é emitido um sinal de fora-de-controle, excluindo a amostra responsável pela emissão do sinal é o  $RL$  (Run Length) e a média desse número de amostras é o  $ARL$  (Average Run Length). Então, o parâmetro  $ARL$  representa o número médio de amostras necessário para que seja detectada uma mudança no processo.

Um sinal de mudança tanto pode ser um falso alarme, como um sinal de que o processo realmente está fora-de-controle após um desvio médio do valor esperado. Para o gráfico de controle emitir esse sinal é preciso que o tempo necessário seja levado em consideração. Se o processo está sob-controle, este tempo deverá ser aumentado para que a taxa de alarmes falsos seja reduzida. Se o processo estiver fora-de-controle, este tempo deverá ser curto para que a mudança seja detectada rapidamente. O cálculo para o valor do parâmetro  $ARL$  é obtido através da equação  $ARL = 1/P(\text{Alarme falso})$ .

Várias técnicas podem ser usadas para calcular o  $ARL$  do gráfico CUSUM e muitos autores têm usado aproximações adequadas para calcular o valor dos  $ARL$ . Grande parte deles recomendam a aproximação proposta por Siegmund (1985) por causa de sua simplicidade. Para o gráfico CUSUM unilateral (isto é,  $C_i^+$  ou  $C_i^-$ ) com parâmetros  $h$  e  $k$ , a aproximação de Siegmund (1985) é definida como:

$$ARL_0 \approx 2\{\exp(h + 1.166) - h - 2.166\} \quad (2.12)$$

## 2.5 Vigilância Estatística

Um sistema de vigilância monitora mudanças quando novas observações tornam-se disponíveis, no decorrer do estudo. Segundo Sonesson e Bock (2003), vigilância estatística significa um monitoramento de um processo estocástico  $X = \{X(t); t = 1, 2, \dots\}$  com o objetivo de detectar uma mudança importante no processo, em um tempo desconhecido  $\lambda$ , tão rápida e precisamente possível.

A cada instante de tempo  $s$ , deve-se discriminar entre dois estados no sistema monitorado: sob-controle e fora-de-controle. Para isso, utilizam-se observações acumuladas até  $s$  ( $X_s = \{X(t); t \leq s\}$ ). Se  $X_s$  pertence a esse conjunto, então há uma indicação que o processo está no estado fora-de-controle e um alarme é soado. Usualmente, isso é feito usando uma função alarme  $f(X_s)$  e um limite de controle  $h$ , tais que o tempo de um alarme  $t_A$  é escrito como

$$t_A = \min\{s, f(X_s) > h\} \quad (2.13)$$

Usualmente, uma mudança em um parâmetro na distribuição de  $X$  será de interesse. Por exemplo, uma mudança de nível, da variação ou mesmo uma mudança em ambos, ao mesmo tempo.

Diferentes tipos de medidas são utilizadas para avaliar um método de vigilância, caracterizando seu comportamento quando o processo está sob-controle e fora-de-controle. Sob-controle, todos os alarmes são falsos. A distribuição de um alarme falso (falsa indicação de mudança na média) é freqüentemente resumida pelo número médio de observações até que o alarme soe, dado que o processo está sob-controle ( $ARL_0$ ).

$$ARL_0 = E[t_A | \lambda = \infty] \quad (2.14)$$

onde  $\lambda$  é o tempo verdadeiro de mudança no processo e que é desconhecido na prática.

Outra medida normalmente utilizada é a probabilidade de um alarme falso. Dado que o tempo  $t_A$  é uma variável discreta:

$$P(t_A < \lambda) = \sum_{t=1}^{\infty} P(\lambda = t)P(t_A < \lambda | \lambda = t) \quad (2.15)$$

Essa probabilidade depende da distribuição de  $\lambda$  que muitas vezes é desconhecida, por isso essa medida é difícil de ser utilizada.

As medidas de avaliação com respeito a uma mudança verdadeira podem ser feitas de muitas diferentes maneiras. Em uma vasta literatura sobre controle de qualidade, o número médio de observações até que o alarme soe, dado que o processo está fora-de-controle ( $ARL_1 = E[t_A | \lambda = 1]$ ) é muito usado. Isso implica que a mudança ocorreu assim que o monitoramento foi iniciado. Outra medida de avaliação, que quantifica a possibilidade de mudanças posteriores é a espera média condicional ("*Conditional expected delay - CED(t)*"). O  $CED$  depende do instante de tempo da mudança. O tempo médio de espera para um alarme verdadeiro, dado que a mudança ocorreu no tempo  $t$  é expressa por:

$$CED(t) = E[t_A - \lambda | t_A \geq \lambda = t] \quad (2.16)$$

Quando um sistema de vigilância é avaliado, deve-se encarar um *trade-off* (balanceamento) entre alarmes falsos e tempos de espera curtos para observar um alarme verdadeiro. A maneira de tratar isso é usualmente a mesma como na situação de testes de hipóteses, em que o erro Tipo I é fixo e avaliações de poder são realizadas para várias situações. No contexto de vigilância, a situação tem sido tradicionalmente, caracterizar o erro Tipo I pelo  $ARL_0$ . Usualmente, diferentes métodos são comparados para um valor fixo de  $ARL_0$ .

# Capítulo 3

## Proposta de Trabalho

### 3.1 Problemas com o método do Rogerson (2001)

Rogerson (2001) propõe um sistema de vigilância no qual combina métodos de soma acumulada (CUSUM) com uma estatística de detecção de conglomerados espaço-temporais para um conjunto de dados pontuais (Teste de Knox). O resultado é um procedimento para uma rápida detecção de alguma interação espaço-tempo emergente, para um conjunto de eventos pontuais monitorados seqüencialmente. A aproximação conta com uma estatística de Knox local que é útil em análises retrospectivas para detectar quando e onde a interação espaço-tempo ocorre.

Como mencionado anteriormente, Rogerson (2001) propõe um sistema de vigilância que detecte clusters ativos (vivos) através do monitoramento de uma quantidade  $S_i$ , que vai sendo somada à medida que uma nova observação  $i$  se torna disponível na análise. Caso esta soma  $S_i$  ultrapasse um limiar pré-determinado  $h$ , o alarme deve ser soado. No entanto, dois problemas podem ser verificados nessa técnica: a possibilidade de eventos que disparam o alarme não pertencerem ao eventual cluster e a impossibilidade de localização espacial do mesmo.

No primeiro problema, verifica-se que pode ser que a soma acumulada  $S_i$  esteja tão próxima do limiar  $h$ , que um evento que não pertença ao cluster, faça o alarme soar. Numa situação como esta, provavelmente o alarme já estaria na iminência de ser soado, mas não é interessante que o sistema soe justamente no momento que um evento não pertencente ao cluster ocorre. Uma situação como essa possivelmente acontece, dado que eventos que não pertencem ao cluster excedem seu valor esperado, eventualmente.

Quanto ao fato da técnica apresentada por Rogerson (2001) não identificar clusters espacialmente, duas situações são verificadas: a) Ocorrendo o primeiro problema, a localização do evento que disparou o alarme não pode ser usada como um identificador da posição do cluster. b) Mesmo que o alarme seja soado por um evento que pertença ao cluster, não

podemos precisar o tamanho do mesmo ou encontrar os demais elementos que o compõem olhando apenas os eventos responsáveis pela ultrapassagem do limiar. Isso acontece pois os eventos são ordenados apenas pelos tempos de ocorrência no acúmulo da soma.

## 3.2 Propostas

Um dos objetivos do trabalho é adaptar a técnica apresentada por Rogerson (2001), com o intuito de detectar espacialmente o cluster eventualmente existente e que teria disparado o alarme. Além disso, queremos isolar os eventos que não pertencem ao cluster, mas que contribuem para fazer o alarme soar.

A idéia proposta é monitorar os novos eventos não através de uma soma acumulada, mas através de superfícies acumuladas. De forma resumida, a cada novo evento, é calculado o escore local de Knox que é distribuído no espaço através de uma densidade de kernel, gerando uma superfície na região do evento. Essas superfícies são acumuladas ao longo do tempo e geram uma saliência pronunciada em torno de um eventual cluster.

# Referências Bibliográficas

- [1] Figuras geradas pelo software estatístico **R**.
- [2] Rogerson, P.A. (2000). *Monitoring point patterns for the development of space-time clusters*. *Jornal Royal Statistical Society* (2001) **164**, Part 1, 87-96. University at Buffalo, USA.
- [3] Kulldorff, M. (2001). *Prospective time periodic geographical disease surveillance using a scan statistic*. *Jornal Royal Statistical Society* (2001) **164**, Part 1, 61-72. University of Connecticut, Farmington, USA.
- [4] Jarpe, E. (1999). *Surveillance of Spatio-Temporal Patterns: Change of Interaction in a Ising Dynamic Model*. Göteborg University, Sweden.
- [5] Sonesson, C.; Bock, D. (2002). *A review and discussion of prospective statistical surveillance in public health*. Göteborg University, Sweden. *Jornal Royal Statistical Society* (2003) **166**, Part 1, pp 5-21.
- [6] Siegmund, D., O. (1985). *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer.
- [7] Hardle, W. (1999). *Smoothing Techniques*. Louvain-La-Neuve.
- [8] Griffiths, D., F.; Higham, D. J. (1997). *Learning Latex*. SIAM - Society for Industrial and Applied Mathematics. Philadelphia.
- [9] Daley, D., J.; Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer Series in Sstatistics.
- [10] Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press Inc. Londres.
- [11] Manly, B., F., J.; Mackenzie, D., I. (2003). *CUSUM environmental monitoring in time and space*. *Environmental and Ecological Statistics*, **10**, 231-247.
- [12] Kulldorf, M.; Hjalmar, U. (1999). *The Knox Method and other Tests for Space-Time Interaction*. *Biometrics*, **55**, 544-552.

- [13] Rogerson,P.A..(1997). *Surveillance systems for monitoring the development of spatial patterns*. Statistics in Medicine, **16**, 2081-2093.
- [14] Frisen,M..(2003). *Sstatistical surveillance. Optimality and methods*. International Statistical Review, **71**, 403-434.
- [15] Rowlingson,B.,S.; Diggle,P.,L..(1993). *Splancs: Spatial Point Pattern Analysis Code in S-Plus*. Lancaster University, Lancaster, U.K..
- [16] Rowlingson,B.,S.; Diggle,P.,L..(1996). *Splancs Supplement - Spatial and Spatial-Temporal Analysis*. Lancaster University, Lancaster, U.K..
- [17] Camara,G.; Monteiro,A.,M.; Fuks,S.; Camargo,E.; Felgueiras. (2001) *Análise Espacial*. INPE.
- [18] Bailey,T.,C.; Gatrell,A.,C.. (1995). *Interactive Spatial Data Analysis*. Longman Scientific Technical.
- [19] Raubertas,R..(1989). *An analysis of disease surveillance data uses geographic locations of the reporting units*. Statist.Med.. **8**, 267-271.
- [20] Tango, T..(1995). *A class of tests for detecting "general"and "focused"clustering of rare diseases*. Statist.Med.. **14**, 2323-2334.
- [21] Knox, E. G..(1964). *The detection of space-time interactions..* Appl. Statist.. **13**, 25-29.
- [22] Woodall, W. H.; Adams, B. M..(1993). *The Statistical Design of CUSUM Charts..* Quality Engineering.. **5**.
- [23] Montgomery, D. C..(2000). *Introduction to Statistical Quality Control..* 4th Edition, New York : John Wiley, 2000.
- [24] Piterbarg, V. I..(1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields..* AMS Transl. of Math. Monographs, 148, Providence, R.I.
- [25] Socquet-Juglard, H.; Dysthe, K. B.; Trulsen, K.; Fouques, S.; Krogstad, H. E..(2004). *Spatial Extremes and Shapes of Large Waves..* Rougue Waves 2004 Workshop, Brest.
- [26] Coles, S..(1999). *Extreme value theory and applications*.
- [27] Johnson, R. D.; Wichern, D. W..(1998). *Applied Multivariate Analysis*, 4a ed., New Jersey: Prentice Hall.

Anexo 1: Artigo - *Revista iP -  
Informática Pública*

**Anexo 2: Artigo - *International  
Journal of Health Geographics***

# Sistema de vigilância para detecção de interação espaço-tempo de eventos pontuais

Taynãna C. Simões, Renato M. Assunção

Departamento de Estatística – Universidade Federal de Minas Gerais (UFMG)  
Caixa Postal: 702 – 31270-901 – Belo Horizonte – MG – Brasil  
tay\_estatistica@yahoo.com.br, assuncao@est.ufmg.br

**Abstract.** *We introduce a surveillance system to monitor point events located in geographical space and in time. The aim is to detect the space-time clusters as soon they start emerge by triggering an alarm. In this article, the main objective is to adapt a technique proposed by Rogerson (2001), aiming at identifying the eventual cluster that would be the principal cause the alarm going off. Additionally, we believe that our technique isolate events that don't belong to the cluster but that contribute to the alarm warning.*

**Resumo.** *Propomos um sistema de vigilância para o monitoramento de eventos pontuais, a fim de detectar rapidamente a formação de conglomerados espaço-tempo, assim que estes começam a surgir, disparando um alarme. Neste artigo, o objetivo principal é adaptar a técnica apresentada por Rogerson (2001), com o intuito de identificar o cluster eventualmente existente e que seria a principal causa do disparo do alarme. Além disso, acredita-se que a técnica proposta permita isolar eventos que não pertençam ao eventual cluster, mas que contribuam para fazer o alarme soar.*

## 1. Introdução

Detectar aglomerados espaço-tempo de forma rápida, eficiente e em tempo real é uma necessidade em várias áreas do conhecimento. Na área da saúde principalmente, técnicas e ferramentas com este fim têm uma importância particular, pois permitem que intervenções ou ações de controle sejam realizadas de forma eficiente.

O sistema de vigilância, proposto neste artigo, consiste no monitoramento de um conjunto de eventos pontuais, a fim de detectar rapidamente a formação de conglomerados espaço-temporais, assim que estes começam a surgir. O sistema é um método estatístico prospectivo, no qual é feita análise repetida de dados acumulados seqüencialmente ao longo do tempo, com o objetivo de detectar rapidamente uma mudança relevante que ocorra na série de dados.

Utilizando a estatística de detecção de conglomerados espaço-tempo de Knox e métodos de soma acumulada (CUSUM), Rogerson (2001) propõe um sistema de vigilância que detecte clusters ativos (existentes no momento da análise). No entanto, dois problemas verificados nesta técnica são: 1 - eventos que disparam o alarme podem

não pertencer ao eventual cluster; 2 – a técnica é capaz de detectar a presença de clusters localizados, mas não identifica estes clusters, dando sua posição e extensão no espaço e tempo.

Desta forma, nosso trabalho procura adaptar a técnica apresentada por Rogerson (2001), com o intuito de identificar o cluster eventualmente existente e que seria a principal causa do disparo do alarme. Acredita-se ainda que a técnica proposta permita isolar eventos que não pertençam ao cluster, mas que contribuam para fazer o alarme soar. A idéia é monitorar os novos eventos não através de uma Soma Acumulada, mas através de superfícies acumuladas. De forma resumida, a cada novo evento, é calculado um escore de interação espaço-temporal localizado e este escore é distribuído no espaço através de uma densidade de kernel, gerando uma superfície na região do evento. Estas superfícies são acumuladas e geram uma saliência pronunciada em torno de um eventual cluster.

## 2. Vigilância Estatística

Um sistema de vigilância monitora mudanças quando novos eventos tornam-se disponíveis, no decorrer do estudo. A vigilância estatística significa um monitoramento de um processo estocástico  $X = \{X(t); t=1, 2, \dots\}$  com o objetivo de detectar uma mudança importante no processo, em um tempo desconhecido  $\tau$ , tão rápida e precisamente possível.

A cada instante de tempo  $s$ , deve-se discriminar entre dois estados no sistema monitorado: sob-controle e fora-de-controle. Para que isso ocorra, utilizam-se os eventos acumulados até  $s$  (denotado por  $X_s$ ) para formar conjuntos alarme. Se  $X_s$  pertencerem ao conjunto alarme, há uma indicação de que o processo está no estado fora-de-controle e um alarme é soado. Usualmente, isso é feito usando uma função alarme  $p(X_s)$  e um limite de controle  $h$ . O tempo de um alarme  $t_A$  é escrito como

$$t_A = \min\{s, p(X_s) > h\}$$

Diferentes tipos de medidas são utilizados para avaliar um sistema de vigilância, caracterizando seu comportamento quando o processo está sob-controle e fora-de-controle. A distribuição de um alarme falso, por exemplo, é freqüentemente resumida pelo número médio de eventos até que o alarme soe, dado que o processo está sob-controle ( $ARL_0$ ).

$$ARL_0 = E[t_A | \tau = \infty]$$

onde  $\tau$  é o tempo verdadeiro de mudança no processo e que é desconhecido na prática.

Outra medida normalmente utilizada é a probabilidade de um alarme falso:

$$P(t_A < \tau) = \sum_{t=1}^{\infty} P(\tau = t)P(t_A < \tau | \tau = t)$$

Geralmente, a distribuição de probabilidade  $P(\tau = t)$  para  $\tau$  não é conhecida e isto dificulta o uso desta última medida.

Quando um sistema de vigilância é avaliado, deve-se encarar um *trade-off* entre alarmes falsos e tempos de espera curtos para observar um alarme verdadeiro.

### 3. Sistema de Vigilância – Rogerson (2001)

Rogerson (2001) propõe um sistema de vigilância que combina métodos de soma acumulada (CUSUM) com uma estatística de detecção de conglomerados espaço-tempo para um conjunto de dados pontuais (Teste de Knox).

#### 3.1. Teste de Knox

O teste de Knox é um teste baseado na contagem do número de pares de eventos que ocorrem dentro de intervalos críticos pré-especificados de tempo (T) e distância (D).

Dados  $n$  pontos localizados no tempo e no espaço, existem  $n(n-1)/2$  pares distintos que podem ser divididos em uma matriz  $2 \times 2$  (Tabela 1), indicando os pares que estão próximos no espaço ou próximos no tempo.

**Tabela 1. Tabela  $2 \times 2$  para o cálculo do Teste de Knox (Os valores indicados por • são obtidos a partir dos demais).**

Situação	Próximos no espaço	Não próx. no espaço	Total
Próximos no tempo	$n_{st}$	•	$n_t$
Não próx. no tempo	•	•	•
Total	$n_s$	•	$n(n-1)/2$

Sejam  $n_s$  os pares de eventos observados que são próximos no espaço (separados por uma distância menor ou igual a D),  $n_t$  os pares de eventos que são próximos no tempo (separados no tempo por menos que o intervalo crítico T), e  $n_{st}$  os pares de eventos que são próximos em ambos, espaço e tempo. A estatística de teste  $N_{st}$  deve ser comparada com o seu valor esperado, sob a hipótese nula  $H_0$ , de que não há interação espaço-tempo. Caso o número de eventos próximos no espaço e no tempo exceda substancialmente este valor esperado, há uma indicação da existência de interação espaço-tempo.

Esta estatística testa a presença de interação de forma global, sem localizar os clusters, e sem preocupar-se em verificar se os clusters ainda estão presentes ou se já deixaram de existir. Para resolver estes dois problemas, Rogerson (2001) propôs o uso das estatísticas locais de Knox.

#### 3.2. Estatística Local de Knox

Sejam  $n_s(i)$  o número de eventos que são próximos ao  $i$ -ésimo evento no espaço,  $n_t(i)$  o número de eventos que são próximos ao  $i$ -ésimo evento no tempo e  $n_{st}(i)$  o número de eventos que são próximos simultaneamente no espaço e tempo ao evento  $i$ .

Para encontrar a distribuição de  $N_{st}(i)$ , sob a hipótese nula de não interação espaço-tempo, usa-se que cada permutação aleatória dos índices dos tempos, mantendo as posições espaciais fixas, é igualmente provável. Seja  $n_t^j(i)$  o número de pontos que estão próximos no tempo do evento  $i$  quando a este é associado o  $j$ -ésimo valor do

tempo. Mostra-se que a distribuição de  $N_{st}(i)$  é hipergeométrica com parâmetros  $n-1$ ,  $n_s(i)$  e  $n_t^j(i)$ . Considerando todos os valores possíveis de tempo para o  $i$ -ésimo evento, mostra-se que  $N_{st}(i)$  é uma soma ponderada de hipergeométricas. Pode-se mostrar também que:

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

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

$$+ \left[ \frac{2n_t n_s(i)}{n(n-1)} \frac{n-1-n_s(i)}{n-2} - \left( \frac{2n_t n_s(i)}{n(n-1)} \right)^2 \right]$$

Padroniza-se  $N_{st}(i)$ , resultando na seguinte estatística escore  $z_i$  ajustada:

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

A variável  $z_i$  possui distribuição aproximadamente normal, com média zero e variância 1 (denotada por  $N(0,1)$ ), caso o processo esteja sob controle. Caso exista um cluster, espera-se que  $z_i$  tenha distribuição com média maior que zero, se o evento  $i$  fizer parte do cluster. A distribuição de  $z_i$  tende a permanecer aproximadamente igual a  $N(0,1)$  para aqueles eventos que não fazem parte do cluster.

Suponha que  $i-1$  eventos já tenham ocorrido. Pode-se comparar o valor da estatística de Knox após o caso  $i$  (denotado  $K_i$ ) com o valor que seria esperado sob a hipótese nula e condicionado ao valor da estatística de Knox após o evento  $i-1$  e valores observados de  $n_s(i)$  e  $\{n_t^j(i); j=1, \dots, i\}$ :

$$z_i^* = \frac{K_i - E\{K_i | K_{i-1}, n_s(i), n_t(i)\} - 0.5}{\sqrt{Var\{K_i | K_{i-1}, n_s(i), n_t(i)\}}}$$

A informação em  $z_i^*$ , obtida nesta comparação de  $K_i$  com sua esperança condicional, serve para medir a contribuição do evento  $i$  na estatística de Knox. Embora o escore  $z_i^*$  seja interessante por causa dessas propriedades, ele é difícil de ser calculado, pois os dois momentos  $E\{K_i | K_{i-1}, \dots, K_1\}$  e  $Var\{K_i | K_{i-1}, \dots, K_1\}$  não possuem fórmulas analíticas. Por isso, o escore  $z_i$  acaba sendo preferido na prática. É possível mostrar que  $z_i = K_i - K_{i-1}$ .

### 3.3. Métodos de Soma Acumulada (CUSUM)

Os métodos de soma acumulada são muito usados para monitorar processos industriais com o objetivo de detectar rapidamente uma mudança indesejável no processo. Esses métodos são comuns no controle de um processo estatístico seqüencial de uma variável para potenciais desvios da média esperada.

Seja  $X_i$  o valor de uma variável aleatória com distribuição  $N(\mu, \sigma^2)$ , associada ao  $i$ -ésimo evento. Assumindo que não há correlação na série de eventos, a soma acumulada até o evento  $i$  é dada através da relação:

$$S_i = \max(0, S_{i-1} + x_i - \mu - k\sigma); S_0 = 0$$

Assim, a soma acumulada acumula desvios da média que excedem  $k$  desvios padrão, detectando rapidamente qualquer mudança no valor médio de  $X_i$ . Um sinal de "fora-de-controle" (houve uma mudança na média do processo) é soado no primeiro tempo  $i$ , tal que  $S_i$  exceda algum nível de decisão predeterminado  $h$ .

Os parâmetros  $h$  e  $k$  são expressos em termos do desvio-padrão dos eventos. A escolha de  $h$  depende do número médio de eventos até que ocorra uma mudança, sob a hipótese de que o processo esteja sob controle ( $ARL_0$ ). O valor de  $ARL_0$  é fixado pelo usuário e espera-se que ele seja grande. Um bom procedimento de vigilância terá  $ARL_1$  pequeno, onde  $ARL_1$  é o número médio de eventos até o alarme soar a partir de uma mudança do processo.

Em Rogerson (2001), a estatística de soma acumulada correspondente ao evento  $i$  é dada por:

$$S_i = \max(0, S_{i-1} + z_i - k); S_0 = 0$$

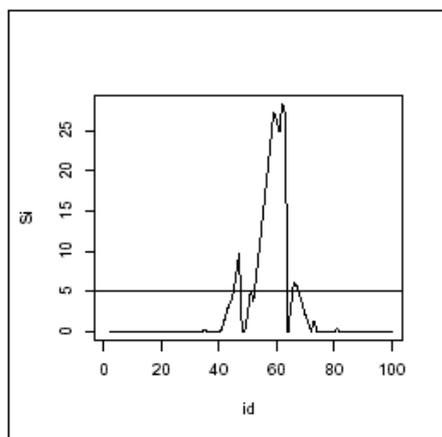
O limiar  $h$  a partir do qual o alarme deve soar é dado pela expressão derivada por Siegmund (1985):

$$ARL_0 \approx 2\{\exp(h+1.166) - h - 2.166\}$$

Para ilustrar o comportamento da técnica proposta por Rogerson (2001), foram gerados 100 pontos com coordenadas  $(x_i; y_i; t_i)$  no espaço e tempo. Os primeiros 80 eventos têm coordenadas espaciais  $x$  e  $y$  geradas de uma distribuição uniforme entre os valores 0 e 1 (denotada  $U(0;1)$ ) e instantes de tempo em que o evento ocorreu  $t$  gerados de uma distribuição uniforme entre os valores 0 e 10 (denotada  $U(0;10)$ ). Assim, os eventos estão distribuídos de forma aleatória dentro de um cubo de base  $(0;1)^2$  e altura 10.

Os outros 20 eventos foram gerados com coordenadas espaciais  $x$  e  $y$  de uma distribuição uniforme entre os valores 0.5 e 0.6 (denotada  $U(0.5;0.6)$ ) e com instantes de tempo  $t$  gerados de uma distribuição uniforme entre os valores 5 e 6 (denotada  $U(5;6)$ ), a fim de formar um cluster nesta região. Os eventos foram ordenados de acordo com o tempo de ocorrência e rotulados por  $i = 1, \dots, 100$  de forma que  $t_i < t_{i+1}$ . Para calcular as estatísticas locais de Knox, usamos os parâmetros críticos  $D = 0.1$  e  $T = 1.0$ . A escolha dos valores dos parâmetros  $D$  e  $T$  é arbitrária e neste exemplo foram escolhidos como 10% da amplitude dos eventos gerados no espaço e no tempo.

Na Figura 1 observa-se o resultado do método, compreendendo um gráfico dos eventos (denotados por  $id$ ) versus a soma acumulada (denotada por  $S_i$ ), mostrando como a soma acumulada pode ser monitorada ao longo da ocorrência dos eventos.



**Figure 1. Visualização do resultado fornecido pelo método do Rogerson (2001) para detecção de interações espaço-tempo.**

Através da Figura 1, verifica-se ainda que o limiar  $h$  a partir do qual o alarme soa tem seu valor próximo a 5 (reta horizontal paralela ao eixo das abcissas) e que é ultrapassado por  $S_i$  pela primeira vez no 45º-ésimo evento. O cluster de tamanho 20 contém os eventos de número 40-47, 50, 51, 54-59, 62, 65, 66. Como o primeiro alarme foi verificado no evento de número 45, nota-se que o mesmo foi soado por um evento que pertence ao cluster. Após o 45º-ésimo evento, a Soma Acumulada retorna a zero e volta a ultrapassar o limiar  $h$  nos eventos de número 50, 53-62, 65 e 66.

#### **4. Problemas com a metodologia de Rogerson (2001)**

Como visto anteriormente, Rogerson (2001) propõe um sistema de vigilância que detecte clusters ativos (vivos) através do monitoramento de uma quantidade  $S_i$ , que vai sendo acumulada através de uma soma, à medida que um novo evento  $i$  se torna disponível na análise. Caso esta soma  $S_i$  ultrapasse um limiar predeterminado  $h$ , o alarme deve ser soado. No entanto, dois problemas podem ser verificados nesta técnica: 1 - eventos que disparam o alarme podem não pertencer ao eventual cluster; 2 - a técnica avalia a presença de clusters, mas não os identifica, dando sua posição e extensão no espaço e tempo.

No primeiro problema, verifica-se que pode ser que a soma acumulada  $S_i$  esteja tão próxima do limiar  $h$ , que um evento com  $z_i > 0$  e que não pertence ao cluster pode fazer o alarme soar. Numa situação como esta, provavelmente o alarme já estava na iminência de ser soado e esta contribuição positiva para  $S_i$  faz o alarme disparar. Isto quer dizer que a localização do evento disparado do alarme não pode ser usada como um identificador da posição do cluster. Esta situação pode acontecer naturalmente, pois os escores  $z_i \sim N(0,1)$  de eventos que não pertençam ao cluster podem exceder  $k$  desvios padrões da média ocasionalmente.

Desta forma, nosso objetivo principal é adaptar a técnica apresentada por Rogerson (2001), com o intuito de detectar o cluster eventualmente existente e que teria disparado o alarme, através da visualização da posição em que este foi formado. A técnica proposta neste artigo também permite isolar eventos que não pertencem ao cluster, mas que contribuem para fazer o alarme soar.

A idéia proposta é monitorar os novos eventos não através de uma soma acumulada, mas através de superfícies acumuladas, que utilizam além do método CUSUM, técnicas de estimação de densidade por kernel.

Suponha que o  $i$ -ésimo evento ocorra na posição  $(x_i, y_i, t_i)$  no espaço-tempo e que  $z_i$  seja o escore de interação espaço-tempo (estatística local de Knox) calculado para este evento. Um valor muito positivo de  $z_i$  é indicativo de um excesso de eventos que ocorreram pouco tempo antes e que estão geograficamente próximos do evento  $i$ . Nós distribuimos este valor de  $z_i$  em torno da posição  $(x_i, y_i)$  do  $i$ -ésimo evento, usando uma superfície de kernel bi-dimensional.

Uma função de kernel bidimensional é obtida a partir de funções de densidade de probabilidade  $K^*(x, y) \geq 0$ , tais que  $\iint K^*(x, y) dx dy = 1$ , com máximo na origem  $(0,0)$ , radialmente simétrica e decrescendo suavemente à medida que nos afastamos da origem. Duas funções muito utilizadas são a função de kernel gaussiana e a função de kernel quártica. A função de kernel gaussiana é definida como

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

O gráfico dessa função está representado na Figura 2.

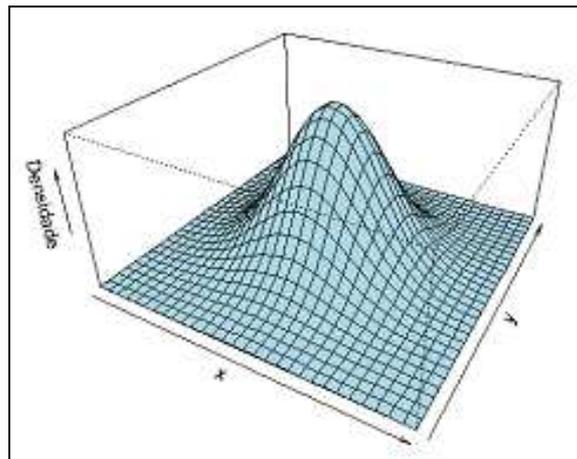


Figure 2. Superfície de kernel usando função gaussiana.

A função de kernel quártica é definida como

$$K^*(x, y) = \begin{cases} \frac{3}{\pi} (1 - (x^2 + y^2))^2 & \text{se } x^2 + y^2 < 1 \\ 0 & , \text{ c.c.} \end{cases}$$

As funções de kernel modificam as funções  $K^*(x, y)$  deslocando-as para um novo centro e mudando sua concavidade na origem, com um parâmetro chamado de largura de banda e denotado por  $\lambda$ . Por exemplo, a função de kernel gaussiana  $K^*(x, y)$ , centrada na posição  $(x_i, y_i)$  do  $i$ -ésimo evento e com largura de banda  $\lambda$  é dada por

$$K(x, y) = \frac{1}{\lambda} K^* \left[ \frac{x - x_i}{\lambda}, \frac{y - y_i}{\lambda} \right] = \frac{1}{\sqrt{2\pi}\lambda} \exp \left\{ -\frac{1}{2} \left( \frac{(x - x_i)^2 + (y - y_i)^2}{\lambda} \right) \right\}$$

De maneira geral, dada uma função base  $K^*(x, y)$  com as propriedades descritas, a função de kernel centrada em  $(x_i, y_i)$  é dada por

$$K_i(x, y) = \frac{1}{\lambda} K^* \left[ \frac{x - x_i}{\lambda}, \frac{y - y_i}{\lambda} \right]$$

O efeito do parâmetro  $\lambda$  é crítico. Um valor muito pequeno irá gerar uma superfície muito descontínua e se for grande demais, a superfície poderá ficar extremamente alisada, mascarando sua forma verdadeira.

Nós usamos a função de kernel, espalhando o escore  $z_i$  em torno da localização espacial do  $i$ -ésimo evento:

$$w_i(x, y) = z_i K_i(x, y) = \frac{z_i}{\lambda} K^* \left( \frac{x - x_i}{\lambda}, \frac{y - y_i}{\lambda} \right) \quad (1)$$

Observe que,  $w_i(x_i, y_i) = \frac{z_i K^*(0,0)}{\lambda}$  e que  $w_i(x, y)$  converge para 0 à medida que a posição  $(x, y)$  afasta-se de  $(x_i, y_i)$ .

Note também que  $z_i = \iint w_i(x, y) dx dy$ , mostrando que, de fato, estamos distribuindo o valor  $z_i$  no espaço usando a função  $w_i(x, y)$ .

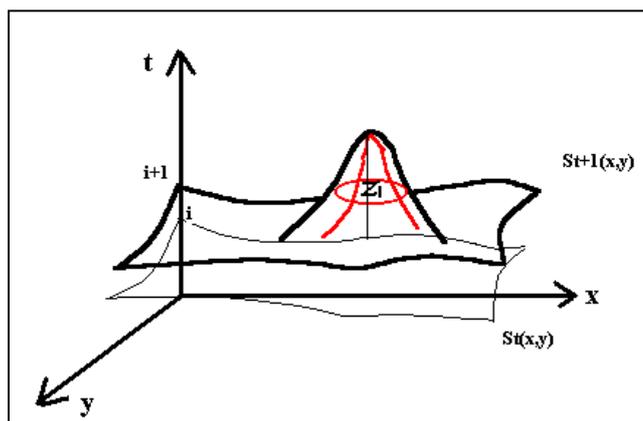
A cada  $i$ -ésimo evento, a superfície é calculada iterativamente como:

$$S_{i+1}(x, y) = \max\{0, S_i(x, y) + z_i K_i\{x, y\}\} = \max\{0, S_i(x, y) + w_i(x, y)\}$$

se  $i \geq 1$  enquanto que  $S_0(x, y) = 0$  para toda posição  $(x, y)$ .

De forma geral, as técnicas de soma acumulada e superfícies acumuladas são bem similares, salvo o fato de a primeira estar somando simplesmente os escores  $z_i$  e a segunda estar multiplicando uma superfície de kernel a este escore, que permite identificar a localização do eventual cluster.

Resumindo, a técnica de soma acumulada é reformulada para ser aplicada a somas de superfícies (superfícies acumuladas). A partir desta técnica, a superfície  $i+1$  seria formada por muitas curvas, de forma que na posição em que estivesse o cluster, se formaria uma curva muito maior, resultante das contribuições dos kernels de cada um dos eventos. A técnica de superfícies acumuladas pode ser visualizada, de forma bem esquematizada, através da Figura 3.



**Figure 3. Visualização da técnica de superfícies acumuladas, para detecção de clusters espaço-tempo.**

Verifica-se, desta forma, que o problema de o alarme ser soado ocasionalmente por eventos que não pertencem ao cluster é solucionado, no sentido que tais eventos contribuirão com uma superfície de kernel na sua localização de origem, não influenciando efetivamente na superfície formada por elementos do cluster.

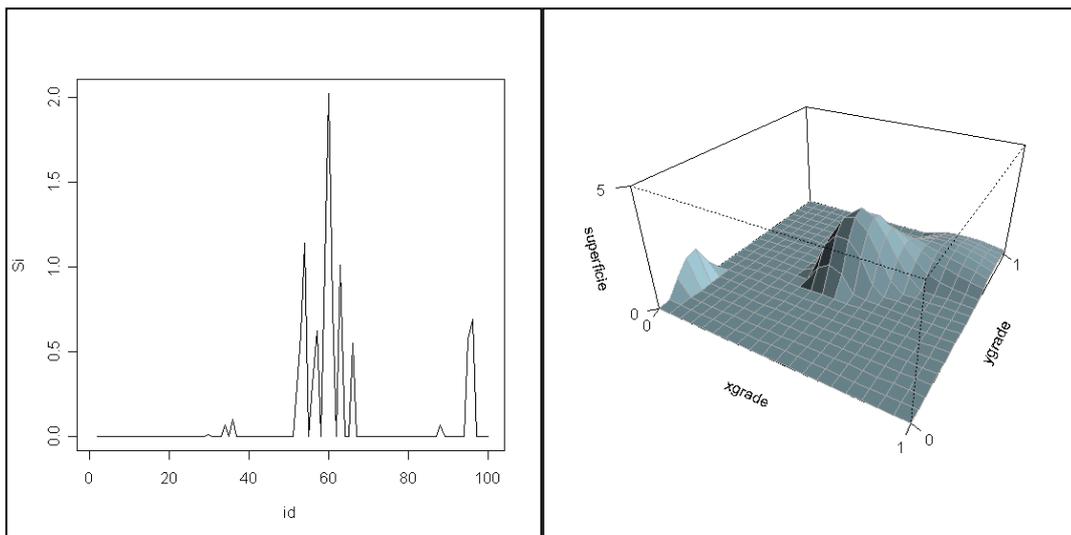
A superfície formada após os  $n$  eventos terem contribuído para a soma deve aparecer com várias protuberâncias resultantes das superfícies de kernel formadas por cada evento, de forma que no intervalo de tempo em que se encontra o cluster, observasse-ia uma protuberância muito maior se comparada às demais, permitindo assim a localização espacial do cluster.

## 5. Simulando a Superfície Acumulada no software R

Para se ter uma idéia do comportamento da superfície acumulada, no caso bidimensional, foram gerados 100 pontos com coordenadas  $(x_i, y_i, t_i)$ . Como antes, 80 eventos possuem coordenadas  $x$  e  $y$  de uma distribuição uniforme  $U(0;1)$  e instantes de

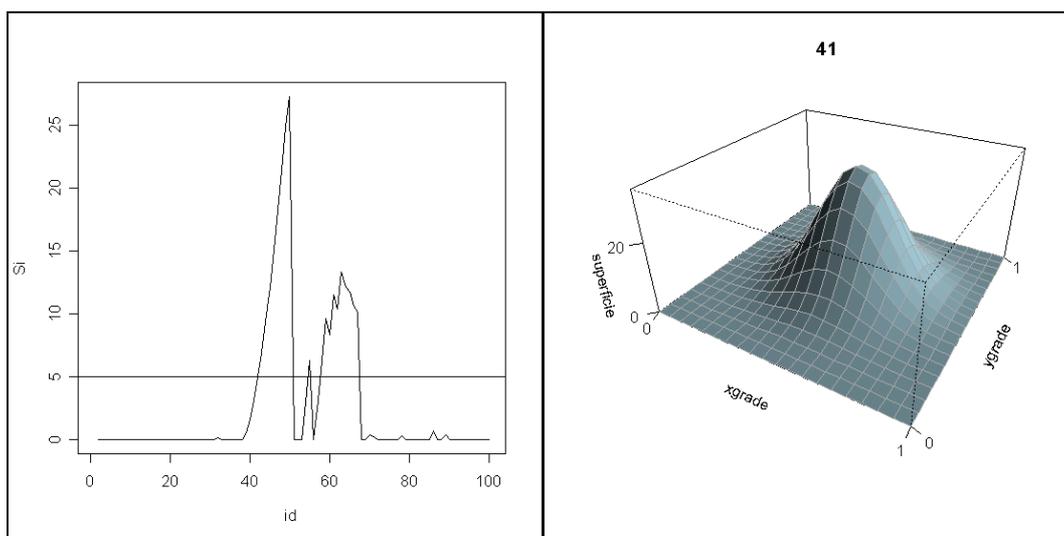
tempo  $t$  de uma uniforme  $U(0;10)$ . Os 20 eventos restantes foram gerados, com coordenadas  $x$  e  $y$  de uma uniforme  $U(0.5;0.6)$  e tempo  $t$  de uma uniforme  $U(5;6)$ , de forma que formassem um cluster nesta região, uma vez que estão próximos no espaço e no tempo. Os parâmetros críticos para o cálculo das estatísticas de Knox são de  $D=0.1$  e  $T=1.0$ . O limiar  $h$  utilizado foi obtido através da fórmula de Siegmund (1985), usando um  $ARL_0 = -n / \log(1-p)$ , onde  $p$  é uma taxa aceitável, predeterminada de alarmes falsos e foi escolhida como 10%. A fórmula para o cálculo de  $ARL_0$  foi obtida, supondo que o número de eventos observados até que um alarme soe sob  $H_0$ , tenha aproximadamente distribuição exponencial.

Os gráficos da Figura 4 mostram as técnicas do Rogerson (2001) e de superfícies acumuladas, sob a hipótese nula de não interação espaço-tempo. Foram gerados apenas os 100 eventos com coordenadas  $(x,y)$  de uma distribuição uniforme  $U(0,1)$  e instantes de tempo  $t$  de uma distribuição uniforme  $U(0,10)$ , ou seja, nenhum cluster foi criado. O gráfico de somas acumuladas mostra que  $S_i$  sai algumas vezes do valor zero, mas não chega a atingir o limiar  $h \cong 5$ , que nem é mostrado por efeito de escala. Desta forma, o alarme não é soado, como era esperado. No gráfico de superfícies acumuladas, vemos apenas algumas elevações resultantes da contribuição das somas das superfícies de kernel dos eventos anteriores.



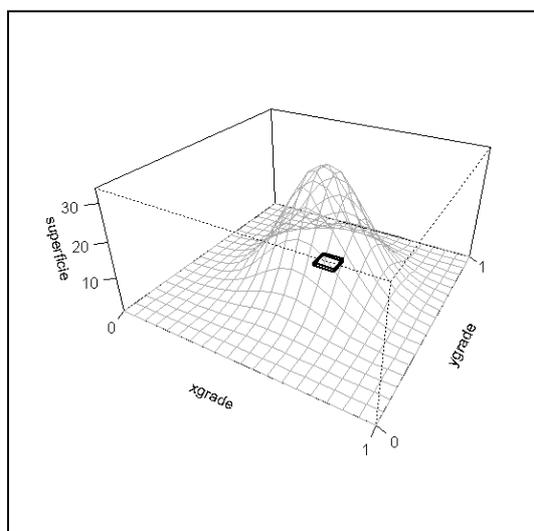
**Figure 4. Gráficos de soma acumulada (Rogerson (2001)) e de superfícies acumuladas, sob  $H_0$ .**

Na Figura 5, foi criado um cluster de tamanho 20, como descrito no início do exemplo. O cluster gerado contém os eventos de número 37- 49, 54, 55, 57 - 59, 61, 63. No gráfico de somas acumuladas  $S_i$ , o limiar  $h$  é ultrapassado pela primeira vez na ocorrência do 41º-ésimo evento, ou seja, o alarme é soado quase no início do cluster. O gráfico de superfícies acumuladas ilustra o método no momento da ocorrência do 41º-ésimo evento que acionou o alarme no método do Rogerson (2001). Observamos uma elevação significativa em relação às demais regiões, levando a considerar a presença de um cluster na região de maior saliência.

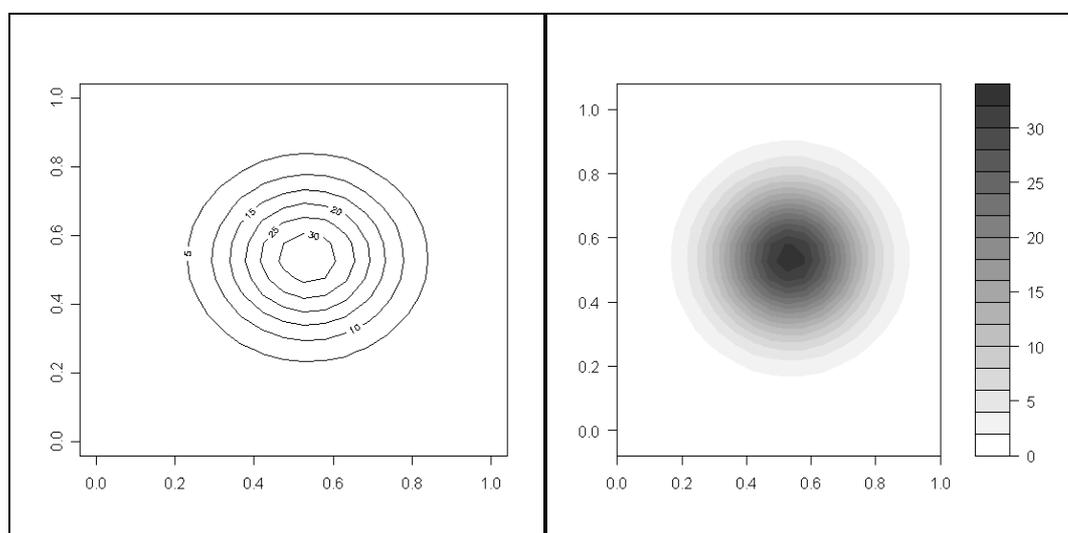


**Figure 5. Gráficos de soma acumulada (Rogerson (2001)) e de superfícies acumuladas, com a existência de um cluster.**

De forma geral, as Figuras 6 e 7 fornecem um mecanismo de visualização do cluster, no momento em que o alarme é soado no sistema do Rogerson (2001). Na figura 6 vemos a posição do cubo compreendendo os eventos do cluster verdadeiro (gerado) na superfície formada no momento de ocorrência do 41º-ésimo evento. Na Figura 7, o primeiro gráfico mostra cortes da superfície (níveis de contorno) e o segundo é um gráfico de imagem, no qual valores maiores da superfície (cores mais fortes) indicam que naquele local existe uma saliência maior se comparada às outras regiões, indicando a localização mais provável do eventual cluster. Além disso, ressaltamos o fato de que em ambos os gráficos da Figura 7, os maiores valores da superfície tenham aparecido entre as posições 0.5 e 0.6 dos eixos x e y, posições estas em que o cluster verdadeiro foi formado.



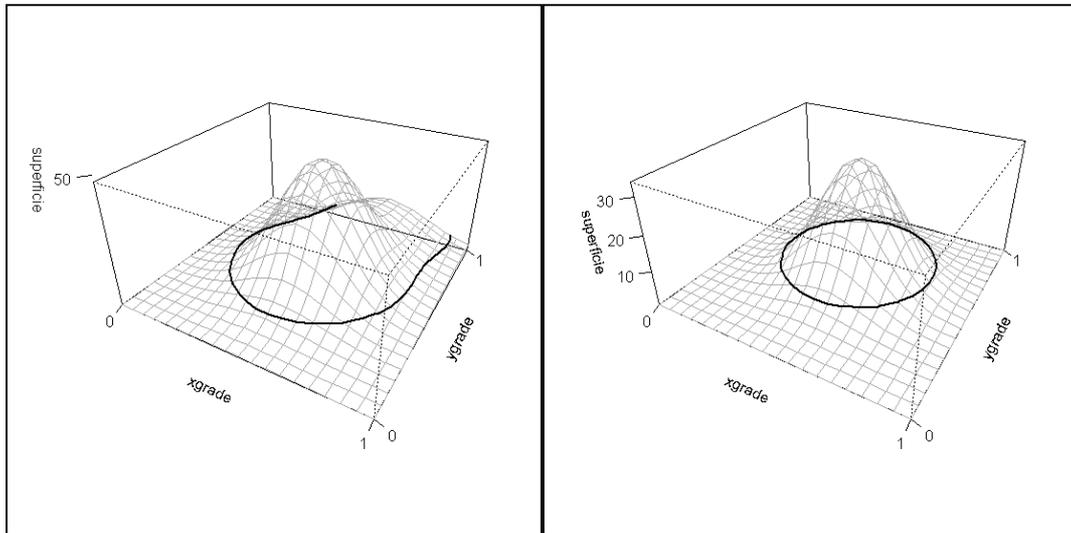
**Figure 6. Gráficos da superfície acumulada, mostrando a posição do cubo representante do cluster gerado.**



**Figure 7. Gráficos de contorno e gráfico de imagem formados no momento do 41º-ésimo evento.**

Apesar da técnica gráfica mostrada dar base a uma boa visualização da posição do cluster, é de interesse encontrar os valores aproximados de suas coordenadas  $x$  e  $y$  no espaço. Para isso, a superfície que contém o ponto de máximo é seccionada por níveis de contorno. Seccionando a superfície que tem o maior pico, obtemos o conjunto de coordenadas  $(x,y)$  tais que  $S_i(x,y) = c$ , onde  $c$  é uma constante que deve ser escolhida de forma a obter as coordenadas mais próximas das coordenadas no cluster.

A Figura 8 ilustra o impacto da escolha da constante  $c$ . No primeiro gráfico, o valor de  $c$  foi escolhido muito próximo do plano, levando à superestimação do tamanho do cluster, uma vez que superfícies vizinhas também são atingidas. No segundo gráfico, a escolha da constante  $c$  foi apropriada, dado que o corte foi feito apenas na superfície que contém o valor máximo e, portanto a região mais provável de conter os eventos do cluster.

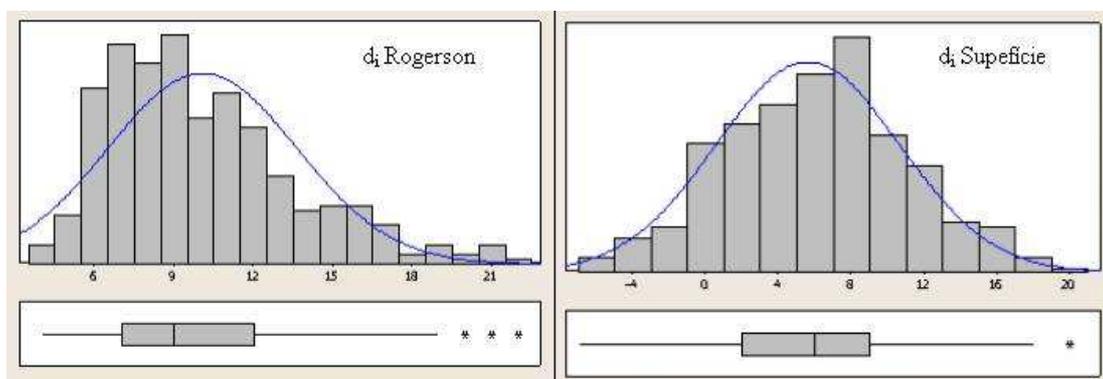


**Figure 8. Gráficos de superfícies acumuladas, com escolhas diferentes da constante  $c$  para o corte das superfícies.**

Uma maneira de avaliar o método de monitoramento é verificando se o alarme foi soado muito antes/depois do cluster verdadeiro ter surgido e através da probabilidade de alarmes falsos, sob  $H_0$ . Para testar a primeira situação, foram feitas 1000 simulações do mesmo cenário acima, utilizando o limiar proposto por Rogerson (2001). Apesar de termos conhecimento de que este limiar é próprio para o uso de séries de observações e não de superfícies, utilizaremos o mesmo, em ambos os métodos, para deixar clara a necessidade de obtermos um limiar diferente quando trabalhamos com superfícies.

Seja  $d_i$  o tempo entre o primeiro alarme e o início do cluster. Se  $d_i > 0$  é possível calcular o número esperado de eventos até que um alarme ocorra, dado que o processo esteja fora de controle, ou seja, dado que existe um cluster ( $ARL_1$ ).

A Figura 9 apresenta a distribuição dos  $d_i$ 's nos métodos do Rogerson (2001) e de superfícies acumuladas. Note que a amplitude dos valores no primeiro nos dá uma indicação de que o tempo de espera por um alarme verdadeiro é maior no método do Rogerson que no segundo método.



**Figure 9. Histogramas dos  $d_i$ 's nos dois métodos.**

Esta colocação pode ser confirmada através da Tabela 2, que mostra além de informações descritivas dos  $d_i$ 's nos métodos do Rogerson (2001) e de superfícies acumuladas, tais como a média, o desvio-padrão, o mínimo (Min), máximo (Max), 1º, 2º (Mediana) e 3º Quartis, os tempos médios de espera por um alarme verdadeiro dado que exista um cluster, em ambos os métodos.

**Tabela 2. Informações relevantes obtidas nas simulações (k=0,5).**

Informações sobre os $d_i$ 's (k=0,5)								
Método	Média	Desvio-padrão	Min	1ºQuartil	Mediana	3ºQuartil	Max	ARL <sub>1</sub>
Rogerson	10,066	3,580	4,000	7,000	9,000	12,000	22,000	0,000
Superfície	5,603	5,048	-7,000	2,000	6,000	9,000	20,000	0,766
Teste de Mann-Whitney: $w = 140074,5$ , $p\text{-valor} = 0,000$								

Através da Tabela 2, verifica-se que, em geral, o alarme soa mais rápido no método de superfícies acumuladas, uma vez que o tempo médio de espera até que um alarme verdadeiro soe, dado que exista interação espaço-tempo (ARL<sub>1</sub>), é menor que no Rogerson (2001). Além disso, esta diferença de tempos de espera dos  $d_i$ 's positivos, entre os dois métodos, é estatisticamente significativa uma vez que o teste de *Mann-Whitney* foi significativo, ao nível de 5%.

No entanto, outra maneira de avaliar um sistema de vigilância é através da probabilidade de alarmes falsos, sob  $H_0$  (sem cluster). Esta probabilidade é calculada como sendo a razão entre o número de vezes que o alarme soa dentro o total de simulações.

A Tabela 3 mostra os resultados da probabilidade de alarmes falsos sob  $H_0$  para 100 simulações do método de somas acumuladas - Rogerson e do método de superfícies acumuladas, utilizando o mesmo limiar do Rogerson (2001) – Superfície e mais dois limites empiricamente observados: Lim1 e Lim2.

Lim1 foi obtido da seguinte maneira: para uma dada simulação, era calculado o valor máximo que a superfície atinga, a cada evento observado. Assim, para uma única simulação, existiam 100 valores de máximos, correspondentes às etapas de formação das superfícies a cada evento acrescentado. Lim1 é o percentil 90 destes máximos. A probabilidade de alarmes falsos é calculada através da razão entre o número de Lim1's diferentes de zero dentre as 100 simulações.

No Lim2, para uma única simulação, era calculado o percentil 90 da superfície em cada etapa de sua formação. Desta forma, existem 100 valores de percentil 90, dos quais é calculada a média aritmética. Isto é feito para cada uma das 100 simulações, permitindo calcular a probabilidade de alarmes falsos como a razão do número de Lim2's que são diferentes de zero, sobre o total de 100 simulações.

**Tabela 3. Informações da probabilidade de alarmes falsos.**

Método	P(Alarmes Falsos)
Rogerson	0
Superfície	0,77
Lim1	0,87
Lim2	0,97

Através da Tabela 3, vemos que a probabilidade de alarmes falsos sob  $H_0$  para o método de superfícies acumuladas, utilizando o limiar do Rogerson (2001) de 0.77 é muito maior se comparado ao método de somas acumuladas que é zero. No entanto, este resultado é esperado, devido ao fato de estarmos usando um limiar apropriado para séries de observações e não para superfícies, como dito anteriormente. Além disso, o método do Rogerson (2001) apresenta um valor muito abaixo que o esperado de 10% de alarmes falsos. Acreditamos que estes resultados possam ser devido ao fato do limiar abordado por Rogerson (2001) ser inadequado nas duas técnicas.

Para os limites Lim1 e Lim2, verificamos que a probabilidade é muito alta, no entanto, este resultado pode ser devido ao fato do procedimento ter sido repetido poucas vezes. Assim, um número maior de iterações deve ser simulado a fim de avaliar principalmente o comportamento de Lim1, que acreditamos ser o mais adequado dentre os limites sugeridos, para o método de superfícies acumuladas.

A intensidade de pontos no cenário gerado e o número de eventos por unidade de volume dentro do cilindro em estudo pode, em muitos casos, ser baixa ou mesmo nula. Assim, uma alternativa seria reduzir o parâmetro  $k$ , utilizado na soma acumulada, para zero, não deixando assim, que a Soma  $S_i$  se torne zero tão frequentemente. Desta forma, fazendo o ajuste  $k = 0$ , fará com que a soma  $S_i$  não tenha tendência a diminuir, possibilitando um alarme mais eficaz, no sentido da detecção mais rápida do cluster.

Estudos ainda não publicados, realizados no LESTE–UFMG (Laboratório de Estatística Espacial), mostram que em cenários onde a intensidade de pontos por volume é baixa, a influência do  $k$  é grande, trazendo subestimação de falsos alarmes no método e prejudicando os resultados devido ao baixo valor dos  $S_i$ 's. Desta forma, a Tabela 4 mostra os resultados descritivos dos  $d_i$ 's, utilizando  $k = 0$ , bem como as probabilidades de alarmes falsos sob  $H_0$  e os  $ARL_1$ 's, nos métodos do Rogerson (2001) e de superfícies acumuladas.

**Tabela 4. Informações relevantes obtidas nas simulações ( $k = 0,0$ ).**

Método	Informações sobre os $d_i$ 's ( $k=0,0$ )							
	Média	esvio-padrã	Min	1ºQuartil	Mediana	3ºQuartil	Max	$ARL_1$
Rogerson	7,646	2,854	2,000	5,000	7,000	9,000	18,000	0,000
Superfície	5,463	5,095	-7,000	2,000	5,500	9,000	20,000	0,834
Teste de Mann-Whitney: $w = 123847,5$ , $p$ -valor = 0,0001								

De uma maneira geral, a Tabela 4 mostra os mesmos resultados que a Tabela 2, ou seja, o método de superfícies acumuladas apresenta tempo médio de eventos até uma mudança verdadeira, dado que exista um cluster, menor que no método do Rogerson (2001).

## 6. Conclusões

Verifica-se que o método proposto de superfícies acumuladas, com base no sistema de vigilância do Rogerson (2001), dá a indicação de ser um bom identificador do cluster, uma vez que detecta a posição espacial do mesmo, através da visualização das representações gráficas das superfícies. Contribuição esta de grande relevância, uma vez que, na prática, a localização do cluster possa levar a várias tomadas de decisão importantes.

No entanto, a suposição de que o novo método forneça um número menor de alarmes falsos deve ser apurada mais profundamente, dado que por simulações prévias, tenha sido verificado que o limiar proposto por Rogerson (2001) e outros empíricos, não sejam apropriados. Acredita-se que um limiar adequado possa ser encontrado, utilizando-se técnicas de teoria de valor extremo, em particular, de máximo de superfícies aleatórias (campos gaussianos). Portanto, os esforços futuros serão direcionados na procura de um limiar ótimo para a nova técnica proposta.

## 7. Referências

- Rogerson, P.A.(2001) “Monitoring point patterns for the development of space-time clusters”. *Jornal Royal Statistical Society* (2001) 164, Part 1, 87-96. University at Bualo, USA.
- Siegmund, D., O..(1985) “Sequential Analysis: Tests and Condence Intervals”. New York: Springer.
- Härdle,W..(1990) “Smoothing Techniques”. Louvain-La-Neuve.
- Sonesson, C.; Bock, D..(2002) “A review and discussion of prospective statistical surveillance in public health”. Göteborg University, Sweden. *Jornal Royal Statistical Society* (2003) 166, Part 1, pp 5-21.
- Diggle, P., J..(1983) “Statistical Analysis of Spatial Point Patterns”. Academic Press Inc. Londres.
- Frisen, M..(2003) “Statistical surveillance. Optimality and methods”. *International Statistical Review*, 71, 403-434.
- Knox, E., G..(1964) “The detection of space-time interactions”.. *Appl. Statist.* 13, 25-29.
- Montgomery, D., C..(2000) “Introduction to Statistical Quality Control”.. 4th Edition, New York : John Wiley, 2000.

# Vigilância espaço-temporal de eventos pontuais via superfícies acumuladas

Taynãna C Simões<sup>1</sup> and Renato M Assunção\*<sup>1</sup>

<sup>1</sup>Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil

Email: Taynãna C Simões - tay\_estadistica@yahoo.com.br; Renato M Assunção\* - assuncao@est.ufmg.br;

\*Corresponding author

## Abstract

---

**Background:** Nós propomos um sistema de vigilância para monitorar a ocorrência de eventos pontuais no espaço e no tempo. O sistema tem o intuito de detectar uma mudança na incidência dos eventos tão logo quanto possível e de identificar no espaço a região ou regiões onde os clusters emergentes estão localizados. Nosso método usa o índice local de Knox proposto por Rogerson [1] como um indicador do grau de interação espaço-temporal considerando a região ao redor do último evento observado e o passado recente. Nós distribuímos este índice no espaço gerando superfícies estocásticas que são somas acumuladas. O sistema baseia-se no máximo que essas superfícies de somas acumuladas atingem. Se a superfície ultrapassar um limiar pré-determinado, um alarme é disparado. Condicionado no evento do alarme ter soado, propomos um método para obter uma seção transversa da superfície que permite visualizar a localização da região ou regiões onde estão localizados os clusters que dispararam o alarme

**Results:** Estudos de simulação mostram que o método permite identificar bem os clusters, uma vez que detecta as posições espaciais dos mesmos de forma bem aproximada.

**Conclusions:** Nós recomendamos o uso do método de superfícies acumuladas para a identificação do cluster atípico no espaço, dado que esta característica é relevante em diversas situações práticas.

---

## Background

Existe um grande interesse atual em desenvolver sistemas de monitoramento que detectem conglomerados espaço-temporais de forma rápida e eficiente pois esses sistemas podem fornecer uma base para atividades de controle sanitário e vigilância epidemiológica (Lawson e Kleinman [?]). Ocorrendo o risco de epidemia numa dada região não-especificada a priori, é importante ter um sistema de vigilância que alerte rapidamente sobre um aumento significativo na incidência da doença naquele local. Uma das

razões para este interesse recente é o desejo de estender sistemas de vigilância puramente temporais para incluir também a informação espacial, hoje coletada regularmente e facilmente analisada com equipamentos de informação geográfica de uso amigável, como o software de geoprocessamento TerraView, desenvolvido pelo INPE (Instituto Nacional de Pesquisas Espaciais).

Nesse contexto, tem sido cada vez maior a demanda por estudos prospectivos, nos quais são feitas análises repetidas de dados distribuídos espacial-

mente e acumulados sequencialmente ao longo do tempo. A utilização desses estudos na detecção de conglomerados espaço-temporais tem sido o foco de alguns estudos recentes tais como Raubertas [2], Rogerson [3], Järpe [4], Kulldorff [5], entre outros. Uma revisão de toda a literatura relacionada a este assunto é apresentada por Sonesson e Bock [6] e por Lawson e Kleinman [?].

Dois dos mais promissores métodos para eventos pontuais são apresentados por Kulldorff *et al.* [7] e Rogerson [1]. O primeiro desses trabalhos propõe o uso de uma estatística de varredura no espaço e no tempo para o monitoramento prospectivo de doenças. O método localiza a região mais provável de existir um cluster, bem como clusters secundários. Além disso, permite análises puramente espaciais ou puramente temporais. A distribuição sob a hipótese de que não existam conglomerados é obtida por permutação.

Rogerson [1] adota uma abordagem diferente. A cada novo evento que é registrado, calcula-se sequencialmente uma estatística local de interação espaço-temporal, chamada de estatística ou escore local de Knox. Esta estatística pode ser vista como um escore que mede o grau de aglomeração dos eventos mais recentes em torno da posição do último evento. Estas estatísticas locais são acumuladas através de uma soma à medida que os eventos são observados. Caso essa soma exceda um limiar predeterminado, há evidência a favor da hipótese de interação espaço-tempo, indicando a formação de clusters emergentes. Quando este limiar é ultrapassado, dizemos que o alarme foi disparado e espera-se que aconteça uma intervenção. Daqui por diante, este método de Rogerson [1] será denominado de Knox+CUSUM.

Existem dois problemas no método Knox+CUSUM que dificultam seu uso na prática. O primeiro é que, ao trabalhar com uma soma acumulada para detectar os conglomerados emergentes, as posições espaciais dos eventos são ignoradas após o cálculo dos escores locais. A consequência é que o método detecta a emergência dos clusters mas não os identifica no espaço. Isto é, o sistema indica que clusters estão emergindo mas não consegue identificar onde eles estão.

Uma possível solução para esse problema seria o exame dos escores locais de Knox mais recentes, imediatamente antes do alarme ter disparado. O objetivo é identificar aqueles eventos cujos escores locais possuam valores positivos muito grandes, em particular aqueles que levaram à ultrapassagem do limiar.

No entanto, esta possível solução esbarra com o segundo problema do método Knox+CUSUM. Após um ou mais clusters emergirem, existirão escores advindos de eventos pertencentes aos clusters mas também haverão escores associados com eventos que não estão nos clusters mas que ocorrem próximos no tempo com aqueles pertencentes aos clusters. Em particular, o evento que faz a soma acumulada ultrapassar o limiar crítico pode nem mesmo pertencer a cluster algum. Uma dificuldade adicional é que, como os clusters devem ter emergido recentemente e devem estar localizados em pequenas regiões do espaço, provavelmente a maioria dos escores mais recentes deve estar associada com eventos não pertencentes aos clusters.

Nosso trabalho procura corrigir esses problemas do método de Knox+CUSUM. Ele tem o intuito de identificar os clusters emergentes e que seriam a principal causa do disparo do alarme. A identificação resume-se a delimitar a região do espaço geográfico onde os eventos dos clusters estão localizados. A técnica proposta permite também isolar eventos que não pertençam aos clusters mas que contribuem eventualmente para fazer o alarme soar.

A idéia chave é monitorar os novos eventos não através de uma soma acumulada, mas através de uma sequência de superfícies acumuladas. De forma resumida, a cada novo evento, é calculado o escore local de Knox e este é distribuído no espaço através de uma densidade de kernel. O resultado é uma superfície plana exceto na região do evento mais recente onde aparece uma protuberância de altura proporcional ao escore. Estas superfícies são sequencialmente acumuladas e, enquanto não existir um cluster, a superfície acumulada terá pequenas e suaves oscilações de altura. A emergência de um cluster vai gerar escores positivos maiores e as superfícies de cada escore vão somando-se às anteriores. Isto acaba por gerar uma saliência pronunciada em torno do eventual cluster emergente. Se o máximo dessas superfícies acumuladas ultrapassar um certo limiar, o alarme é disparado.

## Methods

### Teste de Knox

Suponha que existam  $n$  eventos pontuais  $(x_i, y_i, t_i)$  localizados no espaço e no tempo com  $i = 1, \dots, n$  e  $t_i < t_{i+1}$ . O teste de Knox é baseado na contagem do número de pares de eventos que ocorrem dentro

de intervalos críticos pré-especificados de tempo ( $T$ ) e distância ( $D$ ).

Sejam  $n_s$  os pares de eventos observados que são próximos no espaço (i.e., separados por uma distância menor ou igual a  $D$ ),  $n_t$  os pares de eventos que são próximos no tempo (i.e., separados no tempo por menos que o intervalo crítico  $T$ ), e  $n_{st}$  os pares de eventos que são próximos no espaço e no tempo. A estatística de teste  $N_{st}$  deve ser comparada com o seu valor esperado sob a hipótese nula de que não há interação espaço-tempo. Caso o número de eventos próximos no espaço e no tempo exceda substancialmente este valor esperado, há uma indicação da presença de interação espaço-tempo.

Essa estatística testa a presença de interação espaço-temporal de forma global. Rogerson [1] adaptou-a para um contexto prospectivo e local.

#### *Estatística Local de Knox*

Dentre os  $n - 1$  eventos restantes, sejam  $n_s(i)$  o número de eventos que são próximos do  $i$ -ésimo evento no espaço,  $n_t(i)$  o número de eventos que são próximos do  $i$ -ésimo evento no tempo e  $n_{st}(i)$  o número de eventos que são próximos simultaneamente no espaço e tempo do evento  $i$ .

Para encontrar a distribuição de  $N_{st}(i)$ , sob a hipótese nula de não interação espaço-tempo, usa-se que cada permutação aleatória dos  $n$  índices dos tempos  $t_1, \dots, t_n$  é igualmente provável (mantendo-se as posições espaciais fixas). Considerando todos os valores possíveis de tempo para o  $i$ -ésimo evento, Rogerson [1] mostra que  $N_{st}(i)$  é uma soma ponderada de distribuições hipergeométricas. Sob  $H_0$ , mostra-se também que:

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

A variância da variável aleatória  $N_{st}(i)$  apresentada no artigo do Rogerson [1] está incorreta. No apêndice, mostramos a dedução da expressão correta da variância  $\text{Var}(N_{st}(i))$ , que é igual a:

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

Padroniza-se  $N_{st}(i)$ , resultando na seguinte estatística escore  $z_i$  ajustada:

$$z_i = \frac{n_{st}(i) - E\{N_{st}(i)\} - 0,5}{\sqrt{\text{Var}\{N_{st}(i)\}}} \quad (3)$$

onde o valor 0,5 no numerador é resultado de uma correção de continuidade para melhorar a aproximação normal para a distribuição de  $z_i$ .

Caso não existam clusters espaço-temporais e  $E\{N_{st}(i)\}$  não seja muito pequeno, a variável  $z_i$  possui distribuição aproximadamente normal, com média zero e variância 1 (denotada por  $N(0,1)$ ). Espera-se que  $z_i$  tenha distribuição com média maior que zero se o evento  $i$  fizer parte de um cluster. Um valor muito positivo de  $z_i$  é indicativo de um excesso de eventos que ocorreram em torno de  $t_i$  e que estão geograficamente próximos do evento  $i$ .

#### *Soma Acumulada (CUSUM)*

Na prática, a cada instante de tempo  $t_n$ , só nos interessa o valor  $z_n$  do escore local de Knox para o evento mais recente (o  $n$ -ésimo) avaliado com os  $n - 1$  eventos anteriores. Estes escores são acumulados numa soma  $S_n = \max(0, S_{n-1} + z_n - k)$  com  $S_0 = 0$ . A soma  $S_n$  acumula desvios da média de  $z_n$  sob  $H_0$  (a qual é igual a zero) que excedam um valor  $k$ , detectando rapidamente qualquer mudança substancial no processo. O alarme é soado no primeiro tempo  $n$  tal que  $S_n$  exceda um limiar predeterminado  $h$ .

O parâmetro  $k$  é frequentemente escolhido como sendo igual à metade do desvio-padrão associado com a variável monitorada. No nosso caso, isto significa  $k = 0.5$ . No entanto, em várias simulações que fizemos, percebemos que o método de Rogerson funciona melhor com  $k = 0$ , em termos de controle de alarmes que soam falsamente. Isso acontece porque a soma acumulada se torna zero menos frequentemente, não tendo tendências de diminuir e possibilitando um alarme mais eficaz, no sentido de detecção de cluster mais rapidamente.

o limiar  $h$  depende do valor do parâmetro  $ARL_0$  que representa o número médio de eventos até que o alarme soe falsamente. O valor de  $ARL_0$  é obtido como  $ARL_0 = -n / \log(1 - p)$ , onde  $p$  é a probabilidade pré-estabelecida de obtermos pelo menos um alarme falso após  $n$  eventos. A relação entre  $h$  e

$ARL_0$  usada por Rogerson [1] foi derivada por Siegmund [8]:  $ARL_0 \approx 2\{\exp(h + 1.166) - h - 2.166\}$ . Estas relações assumem que os valores  $z_n$  são independentes entre si, o que não é estritamente correto quando trabalhamos com eventos espaço-temporais.

Existe um trade-off fundamental na escolha de  $ARL_0$ . Como  $h$  é uma função crescente de  $ARL_0$ , se fixarmos um valor extremamente alto para  $ARL_0$ , a consequência será um limiar  $h$  bastante alto. Isto faz com que o alarme demore muito a soar nos casos em que existir de fato um cluster presente. Assim, devemos escolher o menor valor possível e tolerável para o parâmetro  $ARL_0$  de forma a ter uma detecção rápida quando clusters emergirem. Esta escolha de  $ARL_0$  depende muito do problema e do tipo de evento em estudo.

#### Exemplo ilustrativo do método Knox+CUSUM

Para ilustrar o comportamento da técnica proposta por Rogerson [1], foram gerados 100 pontos com coordenadas  $(x_i, y_i, t_i)$ . Os primeiros 80 eventos têm coordenadas espaciais  $x$  e  $y$  geradas independentemente a partir de uma distribuição uniforme entre os valores 0 e 1 (denotada por  $U(0,1)$ ) e instantes de tempo  $t$  em que o evento ocorreu gerados de uma distribuição uniforme entre os valores 0 e 10 (denotada por  $U(0,10)$ ). Os outros 20 eventos foram gerados com coordenadas espaciais  $x$  e  $y$  de uma distribuição  $U(0.5,0.6)$  e com instantes de tempo  $t$  gerados de uma distribuição  $U(5,6)$ , a fim de formar um cluster.

Os eventos foram ordenados de acordo com o tempo de ocorrência e rotulados por  $i = 1, \dots, 100$  de forma que  $t_i < t_{i+1}$ . Para calcular as estatísticas locais de Knox, usamos os parâmetros críticos  $D = 0.1$  e  $T = 1.0$ . A escolha dos valores dos parâmetros  $D$  e  $T$  é arbitrária e devem ser escolhidos de acordo com a situação trabalhada. Neste exemplo, os mesmos foram escolhidos como 10% da amplitude dos eventos gerados no espaço e no tempo. O limiar  $h$  utilizado foi obtido como proposto por Rogerson [1], com probabilidade  $p = 0.1$  de ocorrer pelo menos um falso alarme em  $n = 100$  eventos sucessivos.

Na Figura 1 observa-se o resultado do método, compreendendo um gráfico da soma acumulada (denotada por  $S_i$ ) no eixo das ordenadas versus os eventos (denotados por  $id$ ) no eixo das abcissas. O limiar  $h$  tem seu valor próximo a 5 (reta horizontal paralela ao eixo das abcissas) e é ultrapassado por  $S_i$  pela primeira vez no 45º evento. O cluster de tamanho 20

contém os eventos de números 40-47, 50, 51, 54-59, 62, 65, 66. Como o primeiro alarme foi verificado no evento de número 45, nota-se que o mesmo foi soado por um evento que pertence ao cluster e após acumular evidência de 5 eventos prévios do cluster. Neste exemplo, não há nenhuma intervenção após o alarme soar e a soma acumulada retorna a zero voltando a ultrapassar o limiar  $h$  nos eventos de números 50, 53-62, 65 e 66.

#### Superfícies acumuladas

Se  $S_i > h$ , o alarme soa. Neste momento, devemos enfrentar dois problemas. O primeiro deles é descobrir onde estão os clusters que fizeram o alarme disparar. O segundo é procurar identificar, dentre os eventos mais recentes, quais deles fazem parte desses clusters. Estes problemas aparecem por que o método das somas acumuladas não usa as coordenadas geográficas dos eventos após calcular os escores  $z_i$ . Para resolver estes dois problemas, nós propomos o uso de superfícies espaciais sequencialmente acumuladas.

Embora esteja associado unicamente com a posição  $(x_i, y_i)$  do  $i$ -ésimo evento, o escore  $z_i$  é um indicador de interação espaço-temporal em torno dessa posição. Assim, procuramos distribuir essa medida de interação em torno da posição do  $i$ -ésimo evento. Para isso, utilizamos uma função de kernel bidimensional que é obtida a partir de funções de densidade de probabilidade bidimensionais  $K^*(x, y) \geq 0$ , tais que  $\int \int K^*(x, y) dx dy = 1$ , com máximo na origem  $(0,0)$ , radialmente simétrica e decrescendo suavemente à medida que nos afastamos da origem. Uma função muito utilizada, e a que adotaremos neste artigo, é a função de densidade gaussiana bivariada definida como

$$K^*(x, y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2 + y^2)\right) \quad (4)$$

As funções de kernel modificam as funções  $K^*(x, y)$  deslocando-as para um novo centro e mudando sua concavidade com um parâmetro chamado de largura de banda e denotado por  $\tau$ . A função de kernel centrada na posição  $(x_i, y_i)$  do  $i$ -ésimo evento e com largura de banda  $\tau$ , é dada por

$$K_i(x, y) = \frac{1}{\tau^2} K^*\left[\frac{x - x_i}{\tau}, \frac{y - y_i}{\tau}\right] \quad (5)$$

O efeito do parâmetro  $\tau$  é crítico. Um valor muito pequeno irá gerar uma superfície com oscilações muito bruscas, enquanto que, se  $\tau$  for grande demais, a superfície ficará muito achatada. Há diferentes propostas sobre como escolher  $\tau$ . Uma abordagem simples que será considerada neste trabalho é sugerida por Härdle [9], página 91, para obter boas estimativas de kernel para a densidade unidimensional de uma variável aleatória contínua  $Z$  a partir de uma amostra de tamanho  $n$ :

$$\tau = \frac{1.06}{n^{1/5}} \min \left\{ sd(Z), \frac{iqr(Z)}{1.34} \right\} \quad (6)$$

onde  $sd(Z)$  é o desvio-padrão e  $iqr(Z)$  é a distância inter-quartilica da variável  $Z$ . No nosso caso, a variável  $Z$  deve ser substituída pelas coordenadas  $(x, y)$  dos eventos, sendo o desvio-padrão na fórmula acima calculado como a média entre  $sd(x)$  e  $sd(y)$  e a distância inter-quartilica calculada como a média entre  $iqr(x)$  e  $iqr(y)$ .

Nós usamos a função de kernel gaussiana (5) e largura de banda como proposto acima (6) para espalhar os valores dos escores  $z_i$  positivos em torno de suas localizações espaciais. Seja  $z_i^+ = \max\{0, z_i\}$  e

$$w_i(x, y) = z_i^+ K_i(x, y) = \frac{z_i^+}{\tau^2} K^* \left( \frac{x - x_i}{\tau}, \frac{y - y_i}{\tau} \right) \quad (7)$$

Observe que a altura máxima da superfície  $w_i(x, y)$  ocorre na posição  $(x_i, y_i)$  e é proporcional ao escore  $z_i^+$ :  $w_i(x_i, y_i) = z_i^+ K^*(0, 0)/\tau^2 = z_i^+ / (2\pi\tau^2)$ . Além disso,  $w_i(x, y)$  converge para 0 à medida que a posição  $(x, y)$  afasta-se de  $(x_i, y_i)$ . Note também que  $z_i^+ = \iint w_i(x, y) dx dy$ , mostrando que, de fato, estamos distribuindo o valor  $z_i$  no espaço, usando a função  $w_i(x, y)$ .

A cada  $i$ -ésimo evento, nós acumulamos iterativamente as últimas  $m$  superfícies  $w_j(x, y)$ :

$$\begin{aligned} S_i(x, y) &= \sum_{j=i-m+1}^i w_j(x, y) \\ &= S_{i-1}(x, y) - w_{i-m}(x, y) + z_i^+ K_i(x, y) \end{aligned} \quad (8)$$

se  $i > m$ . Se  $i \leq m$ , nós definimos  $S_i(x, y) = \sum_{j=1}^i w_j(x, y)$ . Assim, as observações mais antigas

são descartadas com a superfície acumulando as informações presentes nos últimos valores  $z_j^+$ , até um total de  $m$  valores.

A razão para esse descarte é que valores muito antigos não devem ajudar a detectar um cluster emergente. Na verdade, como os escores  $z_i^+$  de eventos muito antigos devem ter distribuição aproximadamente  $N(0, 1)$ , sua presença em superfícies acumuladas mais recentes vai afetar apenas a variância fazendo com que ela seja muito grande. Este "ruído" torna mais ineficiente a detecção eventual de clusters emergentes.

Nosso método dispara um alarme imediatamente após o  $i$ -ésimo evento se  $S_i(x, y) > h$  para alguma posição  $(x, y)$ . O limiar  $h$  deve controlar a taxa de alarmes falsos dentre  $m^*$  eventos sucessivos. Suponha que, dentre  $m^*$  eventos observados sem a presença de nenhum cluster espaço-temporal, deseje-se que a probabilidade de algum alarme falso seja igual ou menor que  $\alpha$ . Seja  $Q_j = \max_{x, y} S_j(x, y)$  e

$$M_i = \max_{x, y} \{Q_{i-m^*+1}, \dots, Q_i\} \quad (9)$$

Então, desejamos que  $P(M_i > h) \leq \alpha$ . Assim,  $h$  é o  $\alpha$ -ésimo quantil da distribuição do máximo de  $m^*$  superfícies acumuladas sob a hipótese de que não existem clusters espaço-temporais. O valor de  $h$  não pode ser obtido analiticamente e por isto nós utilizamos métodos Monte Carlo, como explicamos a seguir.

### Determinando o limiar $h$

Para encontrar a distribuição de  $Q_j$  sob a hipótese de que não existem clusters, utilizamos alguns eventos iniciais. Os valores  $z_i^+$  são baseados no número  $n_{st}(i)$  de vizinhos espaço-temporais dentre os eventos prévios ao  $i$ -ésimo evento. Assim, para os primeiros eventos, esse número deve ser zero devido a presença de poucos eventos prévios. Isso quer dizer que, no período de tempo inicial do estudo, existem poucos eventos próximos no espaço e no tempo, tornando os valores  $z_i^+$  muito instáveis. Após um período maior ou igual ao limiar crítico temporal, os valores  $z_i$  devem começar a estabilizar-se. No entanto, para  $i < m$ , a superfície acumulada  $S_i(x, y)$  será composta apenas da soma das  $i$  primeiras superfícies  $w_j(x, y)$ ,  $j = 1, \dots, i$ . Apenas quando  $i$  for maior ou igual a  $m$  teremos  $S_i(x, y)$  composta sempre da soma de  $m$  superfícies  $w_j(x, y)$ . Assim, para

evitar que a distribuição de  $Q_j$  seja afetada pela presença de valores  $z_j^+$  com  $j$  associado aos primeiros eventos e para garantir que  $S_j(x, y)$  seja sempre composta da soma acumulada de  $m$  superfícies, vamos utilizar os primeiros  $2m$  eventos para obter o limiar  $h$ .

A partir desses  $2m$  eventos, permutamos aleatoriamente os índices dos tempos  $B$  vezes, mantendo as posições espaciais fixas. Aplicamos a técnica das superfícies acumuladas calculando  $S_i(x, y)$ ,  $i = 1, \dots, 2m$  para cada  $j$ -ésima permutação, onde  $j = 1, \dots, B$ . A cada  $j$ -ésima permutação, é calculado o valor máximo atingido pela superfície final  $S_{2m}(x, y)$ . Essa permutação dos tempos garante que estamos obtendo os valores do máximo  $Q_{2m}$  sob a hipótese de não interação espaço-tempo, o que nos permite obter a sua distribuição empírica condicionada nas primeiras  $2m$  posições geográficas e nos  $2m$  primeiros tempos observados.

Para obter  $h$ , devemos considerar a distribuição do máximo de  $m^*$  superfícies  $S_j(x, y)$  sucessivas. Procuramos evitar aumentar ainda mais o descarte de informações prévias utilizando apenas a distribuição do valor  $Q_{2m}$ . Para isto, seja  $\alpha$  a probabilidade de um alarme falso dentre  $m^*$  eventos sucessivos. Seja  $A$  o evento em que pelo menos uma superfície  $S_j(x, y)$ , com  $j = i - m^* + 1, \dots, i$  ultrapassa o limiar  $h$  e  $A_j$  o evento em que a superfície  $j$  ultrapassa o limiar. Assumindo estacionariedade temporal de  $S_j(x, y)$ , teremos  $P(A_j) = p$  e portanto,

$$\begin{aligned} \alpha = P(A) &= P(A_{i-m^*+1} \cup \dots \cup A_i) \\ &\leq P(A_{i-m^*+1}) + \dots + P(A_i) \\ &\leq p + \dots + p = m^*p \end{aligned}$$

Assim, basta tomar  $h$  igual ao quantil associado com a probabilidade  $p = \alpha/m^*$  na distribuição empírica de  $Q_{2m}$  para garantir que  $P(M_i > h) \leq \alpha$ . Esta escolha de  $h$  será denotada por  $h_1$ .

Outra possibilidade para escolher  $h$ , denotada por  $h_2$ , é tomá-lo como sendo o  $\theta$ -ésimo percentil da distribuição empírica, onde  $\theta = 1/ARL_0$  e o valor de  $ARL_0$  é obtido como em Rogerson [1], dependendo apenas do número  $n$  de eventos e da probabilidade  $p$  de alarmes falsos. A expressão para  $\theta$  é justificada dado que se  $\theta = P(S_i > h|H_0)$ , o tempo que o limiar é ultrapassado pela primeira vez é  $RL_0 \sim \exp(\theta)$ , e portanto,  $ARL_0 = 1/\theta$ .

### Exemplo do método de superfícies acumuladas

Foram gerados 200 pontos com posições espaciais de uma distribuição  $U(0;1)$ , instantes de tempo  $t$  com distribuição  $U(0;10)$  e parâmetros críticos de distância e tempo de  $D = 0.1$  e  $T = 1.0$ , respectivamente. Estes 200 eventos foram reservados para obter a distribuição empírica do máximo sob  $H_0$ , com  $m^* = m = 100$  e  $B = 1000$  permutações. Para obter os limiares  $h_1$  e  $h_2$  foi usado  $p = 0.01$  como a taxa de alarmes falsos necessária para o cálculo de  $\alpha$  e de  $\theta$ . Nesse exemplo,  $h_1 = h_2 = 22.29$ . O valor de  $\tau$  foi calculado usando-se (6).

Adicionando-se 100 eventos aos 200 iniciais, dois cenários foram analisados. No primeiro deles, não existe cluster entre as novas observações e os eventos são gerados exatamente do mesmo modo que os 200 primeiros. No segundo cenário, geramos mais 80 eventos como os 200 primeiros e 20 outros para constituir um cluster na região  $[0.5, 0.6]^2 \times [8.5, 9.5]$ . Os eventos são reordenados pelo seu tempo de ocorrência de forma que  $t_i < t_{i+1}$ . Para os dois cenários, utilizamos os métodos Knox+CUSUM e superfícies acumuladas.

No cenário 1, a Figura 3 mostra os resultados dos dois métodos, sob a hipótese nula de não interação espaço-tempo. O gráfico da esquerda mostra que a soma acumulada  $S_i$  sai algumas vezes do valor zero, mas não chega nem próxima do limiar  $h \approx 7.35$ . Desta forma, o alarme não é soado, como desejado sob a hipótese nula de não interação. O gráfico da direita mostra superfícies acumuladas  $S_{300}(x, y)$ . Vemos apenas algumas elevações formadas, resultantes da contribuição das somas das superfícies de kernel dos eventos anteriores. Os valores dos máximos  $Q_{201}, \dots, Q_{300}$  não ultrapassaram o valor de limiar  $h$  das superfícies, indicando que também não há evidência de interação espaço-tempo no método das superfícies acumuladas.

No cenário 2, os 100 eventos mais recentes dentre os 300 gerados contém um cluster formado pelos eventos de números 241 e 285. O alarme foi soado no evento  $i = 259$  no método Knox-CUSUM (ver na Figura 4, gráfico da esquerda) e no evento  $i = 255$  no método de superfícies acumuladas (ver superfície  $S_{255}(x, y)$  no gráfico da direita na Figura 4). Existe uma elevação significativa em relação às demais regiões na superfície  $S_{255}(x, y)$ , levando a considerar a presença de um cluster na região de maior saliência. O limiar é ultrapassado pela primeira vez com o valor 25.81.

### *Avaliando o método de superfícies acumuladas*

Simulando 1000 cenários como os cenários 1 e 2 do exemplo anterior, testamos o comportamento do método para três valores de largura de banda  $\tau$ , três posições em que o cluster começava a surgir e vários tamanhos de cluster.

Um dos valores de  $\tau$  foi escolhido como proposto por Härdle [9], página 91, e é denotado por  $\tau_a$ . Os outros dois valores são  $\tau_b = \tau_a/2$  e  $\tau_c = 2\tau_a$ . Os duzentos eventos iniciais foram usados para encontrar a distribuição sob a hipótese nula para o método de superfícies acumuladas. Os 100 últimos eventos são aqueles usados para testar o desempenho do método. Foram testadas três posições para o cluster. A primeira delas corresponde a um cluster localizado no início do período dos 100 eventos restantes. Ele está localizado entre os instantes de tempo 7.5 e 8.5. O segundo cluster ocorre entre os instantes de tempo 8.5 e 9.5 e portanto fica localizado no meio do período dos 100 últimos eventos. Finalmente, o terceiro cluster fica localizado no final do período, entre os instantes 9.5 e 10.0. Analisamos clusters com 13 tamanhos diferentes de acordo com o número de eventos: entre 5 e 30 eventos, variando de dois em dois.

Apresentamos apenas os resultados para clusters de tamanhos 5, 15 e 21, na Tabela 2. Em 1000 simulações, temos o número de alarmes que soaram (denotado por *soou*); o número de alarmes motivados dentre os que soaram (ou seja, o cluster já tinha começado e denotado por *mot.*); o tempo médio de espera por um alarme (*delay*); e o número de alarmes que soaram antes do cluster começar (*al.falso*). Para avaliar a diferença dessas variáveis entre os diferentes  $\tau$ 's foi utilizado o teste não-paramétrico de Kruskal-Wallis. Para testar a diferença entre alarmes soados e alarmes motivados, usamos o teste não-paramétrico de Mann-Whitney. Sob  $H_0$  (sem a presença de cluster), as proporções de alarmes falsos em 100 superfícies  $S_j(x, y)$  sucessivas com  $\tau_a$ ,  $\tau_b$  e  $\tau_c$  foram de 0.005, 0.000 e 0.004, respectivamente.

A Figura 5 mostra a proporção de alarmes que soaram em 1000 simulações para os três  $\tau$ 's à medida que aumentamos o número de eventos no cluster. Não há diferença significativa entre os  $\tau$ 's (p-valor = 0.729). A proporção de alarmes que soaram aumenta com o aumento do número de eventos no cluster e há alto poder em detectar alarmes para clusters de tamanhos moderados (acima de 15 eventos). A mesma interpretação é feita para a Figura 6 que

mostra a proporção de alarmes motivados dentre os alarmes que soaram, ou seja, a proporção de alarmes soados após os eventos do cluster começarem a ocorrer. A proporção de alarmes que soaram e de alarmes motivados é próxima mostrando que, quando um alarme soa, na imensa maioria dos casos existe de fato um cluster presente. Não houve diferença significativa da proporção de ambos os alarmes em relação à posição do cluster.

Apesar da pequena diferença entre proporções de alarmes soados e alarmes motivados, analisamos também sua diferença, isto é, a proporção de alarmes que soaram antes do cluster começar (alarmes falsos). Há diferença significativa dessa proporção entre os  $\tau$ 's (p-valor=0,000). A proporção de alarmes falsos é maior para  $\tau_b$  comparado a  $\tau_a$  e a  $\tau_c$ , que são equivalentes. Não existe diferença quanto ao tamanho do cluster (p-valor=0.774), ao contrário da posição dos mesmos (p-valor=0,000). As Figuras 7, 8 e 9 mostram a distribuição desses alarmes para cada  $\tau$ . Para os três, há diferença das proporções entre as posições (p-valor < 0.002), sendo maior quando o cluster surge no fim das observações e menor quando está no início, como mostrado pelos Box-plots. No entanto, para  $\tau_b$ , não há diferença entre as posições no início e no meio. Não há diferença entre as proporções de alarmes falsos de cada  $\tau$ , em relação ao tamanho do cluster (p-valor > 0.821).

A Figura 10 mostra o tempo médio de espera por um alarme (*delay*) para os  $\tau$ 's, à medida que aumentamos o número de eventos do cluster. Existe evidência de diferença entre os  $\tau$ 's (p-valor = 0.051). O *delay* é maior para  $\tau_b$  comparado a  $\tau_a$  e a  $\tau_c$  (p-valor=0,020). Note que, independentemente do  $\tau$ , o *delay* aumenta até o tamanho de 15 eventos no cluster e depois diminui (p-valor=0.000).

### **Localização do cluster emergente**

Apesar da técnica gráfica fornecer uma boa visualização da posição do cluster, é de interesse encontrar uma forma mas precisa de delimitá-lo no espaço. Para isso, seccionamos por níveis de contorno a superfície  $S_i(x, y)$  que faz o alarme disparar. Queremos escolher o conjunto de coordenadas  $(x, y)$  tais que  $S_i(x, y) = c$ , onde  $c$  é uma constante que deve ser escolhida de forma a obter uma curva de nível próxima da região onde está o cluster.

No momento em que o alarme é acionado, vemos uma região (ou mais de uma) com saliência significativa em relação às demais regiões em estudo, in-

dicando a mais provável localização do cluster. Esta região de maior protuberância contém o valor de máximo global da superfície gerada. No entanto, a superfície que contém esse máximo global pode estar rodeada por superfícies de máximos locais bem menores que este máximo global. Assim, a constante  $c$  não deve ser escolhida com valor muito próximo do plano, dado que pequenas elevações na vizinhança do máximo global também podem ser seccionadas, superestimando o tamanho do cluster. A Figura 11 ilustra o impacto da escolha da constante  $c$ . No primeiro gráfico, o valor de  $c$  foi escolhido muito próximo do plano, levando à superestimação do tamanho do cluster. No segundo gráfico, a escolha da constante  $c$  foi mais apropriada, dado que o corte foi feito apenas na superfície que contém o valor de máximo global.

Nós supomos que a superfície que contém o ponto de máximo global pode ser aproximada por uma densidade de probabilidade normal bivariada, com vetor de médias  $\mu$  e matriz de variância-covariância  $\Sigma$ . Nessa distribuição, os contornos compreendendo os valores  $(x, y)$  que estão a uma mesma altura têm valor constante da densidade e formam elipses, ou seja, o valor da densidade é o mesmo em superfícies onde a distância quadrática  $((x, y) - \mu)' \Sigma^{-1} ((x, y) - \mu)$  é constante. Estes contornos são denominados contornos de nível. Assim, podemos encontrar os pontos de coordenadas  $x$  e  $y$  que formam uma elipse que engloba  $(1 - \alpha) \times 100\%$  dos pontos da distribuição e que representariam os pontos ao redor do cluster.

Os eixos de cada elipse de densidade constante estão na direção dos autovetores de  $\Sigma^{-1}$  e seus comprimentos são proporcionais aos inversos das raízes quadradas dos autovalores de  $\Sigma^{-1}$ . Logo, os contornos de densidade constante para uma distribuição normal bivariada são elipses definidas por  $(x, y)$  tais que

$$((x, y) - \mu)' \Sigma^{-1} ((x, y) - \mu) = c^2 \quad (10)$$

Segundo Johnson e Wichern [10], essas elipses são centradas em  $\mu$  e tem eixos  $\pm \sqrt{\tau_i} e_i$ , onde  $e_i$  é o autovetor relacionado ao autovalor  $\lambda_i$  e  $\Sigma e_i = \tau_i e_i$ ,  $i = 1, 2$ . Mostra-se que a escolha de  $c^2 = \chi_2^2(\alpha)$ , onde  $\chi_p^2(\alpha)$  é o  $100\alpha$ -ésimo percentil de uma distribuição qui-quadrado com 2 graus de liberdade, leva a contornos que contém  $(1 - \alpha) \times 100\%$  de probabilidade. Ou seja

$$((x, y) - \mu)' \Sigma^{-1} ((x, y) - \mu) \leq \chi_2^2(\alpha) \quad (11)$$

Dado que escolhemos um contorno de nível razoável, a partir de um percentil  $\alpha$  podemos obter as coordenadas aproximadas do cluster como as coordenadas da elipse que engloba  $(1 - \alpha) \times 100\%$  dos pontos da superfície. Para isto, o vetor de médias  $\mu$  é obtido como as coordenadas  $x$  e  $y$  do ponto de máximo global da superfície, no momento que o alarme soa. A matriz  $\Sigma$  pode ser obtida através da matriz de variância-covariância das coordenadas  $(x, y)$  do nível de contorno adotado ( $\Sigma^*$ ) a menos de uma constante  $u$ , ou seja,  $\Sigma^* = u\Sigma$ .

Pela expressão da densidade normal bivariada, vemos que o ponto de máximo é dado por

$$Max = \frac{1}{2\pi \sqrt{\det(\Sigma)}} \quad (12)$$

e portanto podemos fazer

$$A = \det(\Sigma) = \left( \frac{1}{2\pi Max} \right)^2 \quad (13)$$

Como exposto anteriormente,  $\Sigma^* = u\Sigma$ . Aplicando o determinante em ambos os lados, podemos fazer

$$B = \det(\Sigma^*) = k^2 \det(\Sigma) \quad (14)$$

Finalmente, de  $A$  e  $B$ , mostra-se que a matriz  $\Sigma$  procurada é dada por

$$\Sigma = \left( \frac{A}{B} \right)^{1/2} \Sigma^* \quad (15)$$

Dado que exista um cluster no conjunto de dados, uma questão importante é encontrar o contorno de nível que forneça as menores diferenças entre as coordenadas do cluster verdadeiro e as coordenadas da

elipse. Testando todos os decis, em 1000 simulações do cenário descrito no exemplo de aplicação das superfícies acumuladas, chegamos a conclusão que o nível de contorno correspondente ao percentil 30 deve ser utilizado. Isto sugere que não há diferença estatisticamente significativa entre as coordenadas do cluster verdadeiro e da elipse, tomando valores próximos ao percentil mencionado.

#### *Exemplo de visualização do cluster no método Knox+CUSUM*

Aplicamos a idéia de elipses da normal bivariada para encontrar as coordenadas aproximadas do cluster, no momento que o alarme soou no exemplo anterior para o método de superfícies acumuladas (255º evento) e o percentil 30 para o corte da superfície.

Na Figura 12, vemos a superfície  $S_{255}(x, y)$  com a curva de nível e o gráfico de níveis de contorno dessa superfície. No gráfico de imagem da Figura 13, os pontos plotados representam os eventos do cluster e a elipse as coordenadas estimadas pela técnica sugerida de localização espacial do cluster. Vemos que a elipse engloba todos os pontos. A amplitude das coordenadas estimadas e das coordenadas do cluster, nos eixos  $x$  e  $y$ , são mostradas na Tabela 1. Todos os decis foram testados e novamente o percentil 30 forneceu as menores diferenças entre as coordenadas.

#### **Aplicação: Meningite Meningocócica**

Para ilustrar o método proposto, utilizaremos dados referentes a casos de meningite meningocócica, ocorridos entre os anos de 1998 a 2000, nos bairros da cidade de Belo Horizonte, Brasil. Foram registrados 149 casos que continham informações sobre a posição no espaço e o tempo de ocorrência, além de outras informações.

A Figura 14 mostra o Mapa de Kernel com a distribuição da doença na cidade, independentemente do ano de ocorrência.

No método Knox+CUSUM, o alarme soou uma única vez, deflagrado pelo evento de número 91. No método de superfícies acumuladas, a estatística monitorada ficou acima do limiar a partir do evento de número 94. A Figura 15 mostra a superfície formada no momento de ocorrência do 94º evento. Vemos uma protuberância elevada a noroeste da região observada, indicando a localização mais provável de um cluster espaço-temporal. A Figura 16 é uma

imagem da superfície anterior, com as cores associadas às alturas da superfície da Figura 15. Note que a região de maior protuberância da figura anterior é destacada no gráfico de imagem, indicando uma ocorrência excessiva, refletida nas estimativas mais altas da superfície, como visto na legenda.

A Figura 17 mostra o mapa da cidade de Belo Horizonte, apenas com os casos de Meningite Meningocócica no ano de 1999. Os pontos em destaque indicam os eventos que dispararam o alarme no método Knox+CUSUM e no método das superfícies acumuladas. Além disso, as posições aproximadas que circundam o cluster são representadas pela elipse destacada.

## **Conclusions**

O método de superfícies acumuladas identifica bem a posição espacial de clusters espaço-temporais emergentes, no momento que a incidência de eventos pontuais está acima do esperado. A identificação é feita através da visualização de representações gráficas das superfícies acumuladas e obtenção das coordenadas aproximadas dos eventos que circundam o cluster. Consideramos esta uma contribuição de grande relevância pois a localização do cluster, pois leva a tomadas de decisão importantes na prática.

Um dos parâmetros críticos do método é a largura de banda ( $\tau$ ) da função de kernel gaussiana. Supondo que o processo pontual é gerado conforme uma distribuição uniforme, concluímos que a mesma deve ter como limite inferior a medida proposta por Härdle [9]. A probabilidade de alarmes falsos sob  $H_0$  está num limite aceitável para larguras de banda testadas nessa faixa. Além disso, o sistema detecta com alto poder clusters que não sejam formados por um número muito pequeno de eventos. Consideramos que o resultado é satisfatório quando existem pelo menos 15 eventos. A proporção de alarmes motivados é muito próxima da proporção de alarmes soados. Isso significa que o sistema quase não soa falsamente entre o início do tempo observado e o primeiro evento do cluster.

Por fim, dado que os parâmetros básicos do método estejam bem definidos, sugerimos o uso do método de superfícies acumuladas para a identificação do cluster emergente, quando trabalhamos com processos pontuais uniformes. Além das vantagens de identificação espacial do cluster emergente, ressaltamos que o limiar proposto por esse método

tem a vantagem de considerar o fato dos eventos serem correlacionados entre si, diferentemente do método sugerido por Rogerson [1]. Para trabalhos futuros, sugerimos o testar o método para diferentes processos pontuais, como por exemplo, eventos gerados através de uma distribuição gaussiana.

## Acknowledgements

Este trabalho foi feito como parte integrante dos projetos desenvolvidos no LESTE-UFMG (Laboratório de Estatística Espacial) e no CRISP-UFMG (Centro de Estudos de Criminalidade e Segurança Pública). A primeira autora teve o apoio financeiro da Fundação FORD e do CNPq através do projeto SAUDAVEL.

## References

1. Rogerson PA: **Monitoring point patterns for the development of space-time clusters**. *Journal Royal Statistical Society* 2001, **164**:87–96.
2. Raubertas R: **An analysis of disease surveillance data uses geographic locations of the reporting units**. *Statistics in Medicine* 1989, **8**:267–271.
3. Rogerson PA: **Surveillance systems for monitoring the development of spatial patterns**. *Statistics in Medicine* 1997, **16**:2081–2093.
4. Järpe E: **Surveillance of spatio-temporal patterns: Change of interaction in a ising dynamic model** 1999.
5. Kulldorff M: **Prospective time periodic geographical disease surveillance using a scan statistic**. *Journal Royal Statistical Society* 2001, **164**, **Part 1**:61–72.
6. Sonesson C, Bock D: **A review and discussion of prospective statistical surveillance in public health**. *Journal Royal Statistical Society* 2003, **166**, **Part 1**:5–21.
7. Kulldorf M, Heffernan R, Hartman J, Assunção R, Mostashari F: **A space-time permutation scan statistic for disease outbreak detection**. *PLoS Medicine* 2005, **2**:216–224.
8. Siegmund DO: **Sequential analysis: Tests and confidence intervals**. *Journal Royal Statistical Society* 2003, **166**, **Part 1**:5–21.
9. Härdle W: *Smoothing Techniques*. Louvain-La-Neuve: Springer-Verlag 1990.
10. Johnson RD, Wichern DW: *Applied Multivariate Analysis*. New Jersey: Prentice Hall 1998.

## Apêndice

### 0.1 Variância de $N_{st}(i)$

Seja  $n_t^j(i)$  o número de eventos que estão próximos no tempo do evento  $i$  quando a este é associado o  $j$ -ésimo valor do tempo. Rogerson [1] mostra que a distribuição de  $N_{st}(i)$  é hipergeométrica com parâmetros  $n - 1$ ,  $n_s(i)$  e  $n_t^j(i)$ . Então

$$p\{N_{st}(i) = n_{st}(i)\} = \frac{1}{n} \frac{\sum_{j=1}^n \binom{n_t^j(i)}{n_{st}} \binom{n-1-n_t^j(i)}{n_s(i)-n_{st}}}{\binom{n-1}{n_s(i)}} \quad (16)$$

Assim, se o evento  $i$  recebe o  $j$ -ésimo valor do tempo, tem-se que a esperança de  $N_{st}(i)$  é dada por:

$$E(N_{st}(i) \mid i \text{ recebe } j) = n_s(i) \frac{n_t^j(i)}{n-1} \quad (17)$$

$$E(N_{st}(i)) = E \left[ n_s(i) \frac{n_t^j(i)}{n-1} \right] = \frac{n_s(i)}{n-1} \sum_{j=1}^n n_t^j(i) \frac{1}{n} = \frac{n_s(i)}{n(n-1)} \sum_{j=1}^n n_t^j(i) \quad (18)$$

Usando o resultado  $\sum_{j=1}^n n_t^j(i) = 2n_t$ , temos:

$$E(N_{st}(i)) = \frac{2n_t n_s(i)}{n(n-1)} \quad (19)$$

Para o cálculo da variância de  $N_{st}(i)$ , temos:

$$Var(N_{st}(i) \mid i \text{ recebe } j) = n_s(i) \frac{n_t^j(i)}{n-1} \frac{(n-1) - n_t^j(i)}{n-1} \frac{(n-1) - n_s(i)}{n-2} \quad (20)$$

$$Var(N_{st}(i)) = E\{Var\{N_{st}(i) \mid i \text{ recebe } j\}\} + Var\{E\{N_{st}(i) \mid i \text{ recebe } j\}\} \quad (21)$$

Seja a primeira expressão da equação 21, após a igualdade, denominada expressão 1 e a segunda expressão 2. Desenvolvendo a expressão 1, temos:

$$\begin{aligned}
E\{Var\{N_{st}(i) \mid i \text{ recebe } j\}\} &= \frac{1}{n} \sum_{j=1}^n \frac{n_s(i)n_t^j(i)[(n-1)-n_t^j(i)][(n-1)-n_s(i)]}{(n-1)^2(n-2)} \\
&= \frac{n_s(i)(n-1-n_s(i))}{n(n-1)^2(n-2)} \sum_{j=1}^n n_t^j(i)(n-1-n_t^j(i)) \\
&= \frac{n_s(i)(n-1-n_s(i))}{n(n-1)^2(n-2)} \left[ (n-1) \sum_{j=1}^n n_t^j(i) - \sum_{j=1}^n (n_t^j(i))^2 \right] \quad (22)
\end{aligned}$$

Desenvolvendo a expressão 2, temos:

$$\begin{aligned}
Var\{E\{N_{st}(i) \mid i \text{ recebe } j\}\} &= E \left[ \left( \frac{n_s(i)n_t^j(i)}{n-1} \right)^2 \right] - \left[ \frac{2n_t n_s(i)}{n(n-1)} \right]^2 \\
&= \frac{n_s(i)^2}{n(n-1)^2} \left( \sum_{j=1}^n n_t^j(i)^2 \right) - \left[ \frac{2n_t n_s(i)}{n(n-1)} \right]^2 \quad (23)
\end{aligned}$$

Unindo as expressões 22 e 23:

$$\begin{aligned}
Var(N_{st}(i)) &= \frac{n_s(i)(n-1-n_s(i))}{n(n-1)^2(n-2)} \left[ 2(n-1)n_t - \sum_{j=1}^n (n_t^j(i))^2 \right] \\
&\quad + \frac{n_s(i)^2}{n(n-1)^2} \left( \sum_{j=1}^n (n_t^j(i))^2 \right) - \left[ \frac{2n_t n_s(i)}{n(n-1)} \right]^2 \\
&= \frac{2n_t(n-1)n_s(i)(n-1-n_s(i))}{n(n-1)^2(n-2)} - \frac{\sum_{j=1}^n ((n_t^j(i))^2)[n_s(i)(n-1-n_s(i))]}{n(n-1)^2(n-2)} \\
&\quad + \frac{n_s(i)^2}{n(n-1)^2} \left( \sum_{j=1}^n (n_t^j(i))^2 \right) - \left[ \frac{2n_t n_s(i)}{n(n-1)} \right]^2 \\
&= \left[ \sum_{j=1}^n (n_t^j(i))^2 \right] \left[ \frac{n n_s(i) - n_s(i) - n_s(i)^2}{n(n-1)^2(n-2)} + \frac{n_s(i)^2}{n(n-1)} \right] \\
&\quad + \frac{2n_t(n-1)n_s(i)(n-1-n_s(i))}{n(n-1)^2(n-2)} - \left[ \frac{2n_t n_s(i)}{n(n-1)} \right]^2 \quad (24)
\end{aligned}$$

Assim, a variância de  $N_{st}(i)$  é dada por:

$$\begin{aligned}
\text{Var}(N_{st}(i)) &= \left[ \sum_{j=1}^n (n_t^j(i))^2 \right] \frac{n_s(i)}{n(n-1)^2} \left[ \frac{n-1-n_s(i)}{n-2} + n_s(i) \right] \\
&\quad + \left[ \frac{2n_t n_s(i)}{n(n-1)} \frac{n-1-n_s(i)}{n-2} - \left( \frac{2n_t n_s(i)}{n(n-1)} \right)^2 \right]
\end{aligned} \tag{25}$$

## Tables

**Table 1 - Amplitudes das coordenadas do cluster localizado e do cluster verdadeiro.**

Amplitude	Cluster localizado	Cluster verdadeiro
coordenadas x	(0.3604884 - 0.6921432)	(0.5019981 - 0.5885624)
coordenadas y	(0.3856045 - 0.7722902)	(0.5021474 - 0.5979194)

Table 1: Amplitudes das coordenadas do cluster localizado e do cluster verdadeiro.

**Table 2 - Simulando a técnica de superfícies acumuladas, usando a distribuição uniforme para a gerar os eventos e diferentes parâmetros. Resultados entre 1000 simulações.**

Tamanho do cluster = 5 eventos												
Tau	Posição: Início				Posição: Meio				Posição: Fim			
	soou	mot.	delay	al.falso	soou	Motivado	delay	al.falso	soou	Motivado	delay	al.falso
a	36	34	4.735	2	39	35	4.771	4	29	26	4.115	3
b	22	22	4.727	0	18	18	4.333	0	21	19	4.316	2
c	77	75	4.747	2	58	54	4.611	4	35	30	4.000	5

Tamanho do cluster = 15 eventos												
$\tau$	Posição: Início				Posição: Meio				Posição: Fim			
	soou	mot.	delay	al.falso	soou	Motivado	delay	al.falso	soou	Motivado	delay	al.falso
a	974	971	8.696	3	969	966	8.813	3	987	986	7.898	1
b	945	944	9.043	0	942	942	9.211	0	987	987	8.152	0
c	984	980	8.266	0	983	974	8.372	9	987	979	7.828	8

Tamanho do cluster = 21 eventos												
Tau	Posição: Início				Posição: Meio				Posição: Fim			
	soou	mot.	delay	al.falso	soou	Motivado	delay	al.falso	soou	Motivado	delay	al.falso
a	996	996	8.178	0	996	989	8.132	7	997	991	7.364	6
b	997	997	8.511	0	996	994	8.481	2	998	994	7.539	4
c	991	990	7.809	0	993	982	7.832	11	997	988	7.389	9

Table 2: Simulando a técnica de superfícies acumuladas, usando a distribuição uniforme para a gerar os eventos e diferentes parâmetros. Resultados entre 1000 simulações.

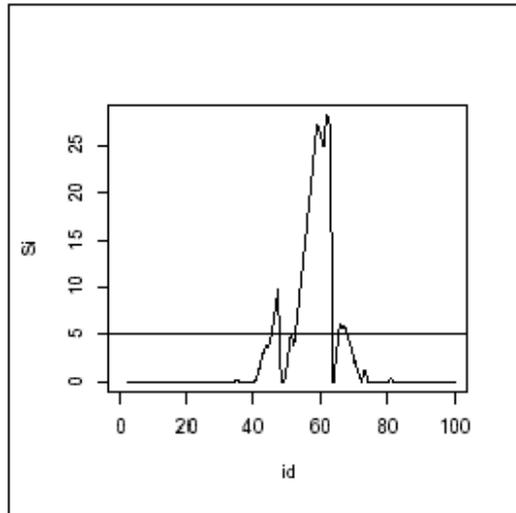


Figure 1: Visualização do método Knox+CUSUM.

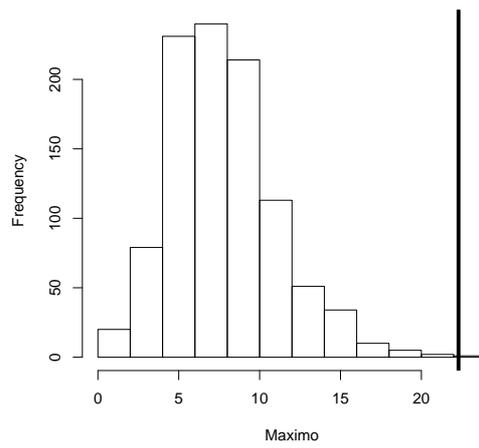


Figure 2: Distribuição do máximo sob  $H_0$  e identificação do limiar (reta vertical).

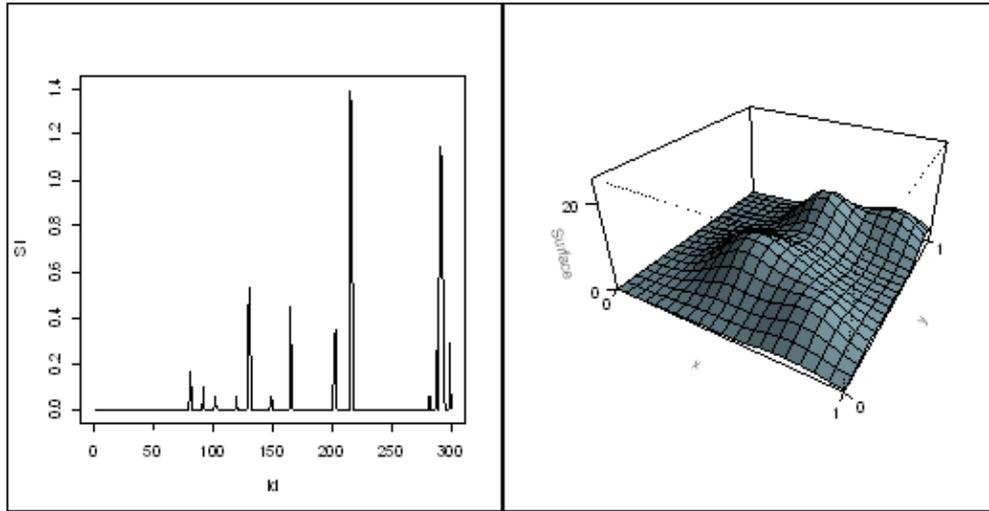


Figure 3: O gráfico da esquerda mostra a soma acumulada  $S_i$  no cenário 1, que não ultrapassa o limiar. O gráfico da direita mostra a superfície acumulada  $S_{300}(x, y)$  no mesmo cenário.

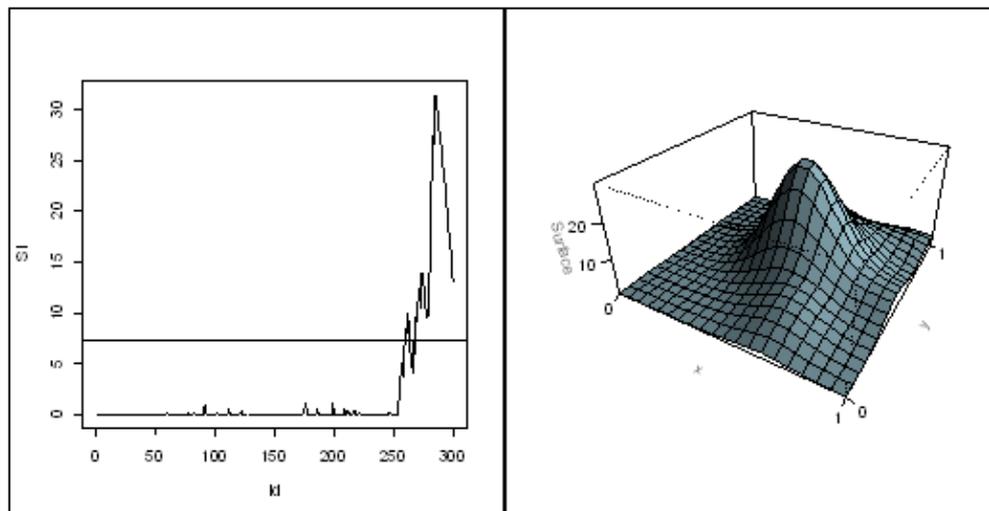


Figure 4: O gráfico da esquerda mostra a soma acumulada  $S_i$  no cenário 2, onde existe um cluster de 20 eventos entre as últimas observações. O gráfico da direita mostra a superfície acumulada  $S_{255}(x, y)$  no mesmo cenário.

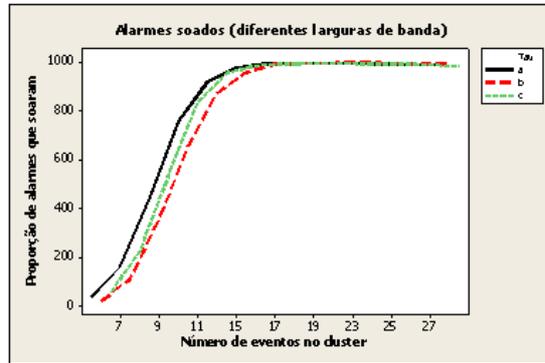


Figure 5: Gráficos de linhas comparando as proporções de alarmes que soaram em 1000 simulações para os três  $\tau$ 's, à medida que aumentamos o número de eventos no cluster.

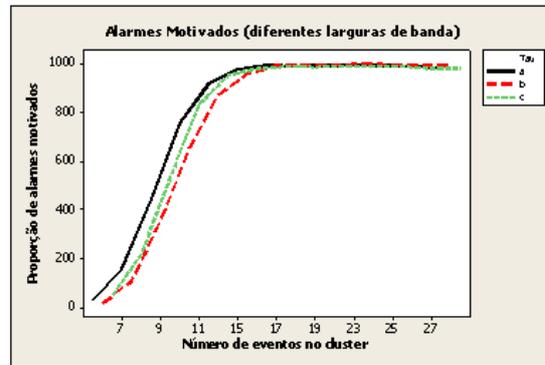


Figure 6: Gráficos de linhas comparando as proporções de alarmes motivados, dentre os alarmes que soaram em 1000 simulações, para os três  $\tau$ 's, à medida que aumentamos o número de eventos no cluster.

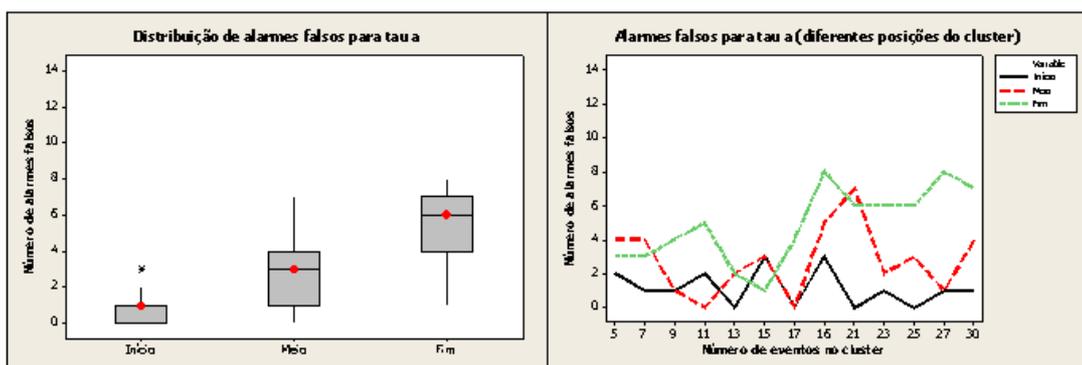


Figure 7: Box-plot com distribuição de alarmes falsos em  $\tau_a$  para as três posições do cluster, independente do tamanho do cluster. Gráfico de linhas da proporção de alarmes falsos para as diferentes posições do cluster, à medida que aumenta o tamanho do cluster.

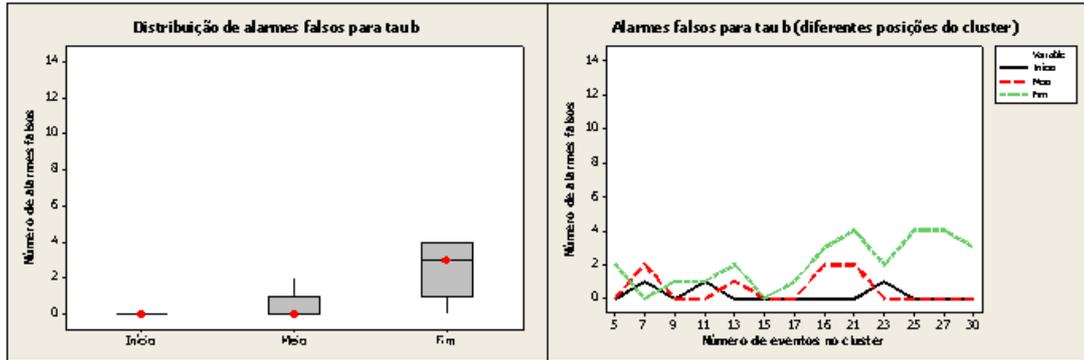


Figure 8: Box-plot com distribuição de alarmes falsos em  $\tau_b$  para as três posições do cluster, independente do tamanho do cluster. Gráfico de linhas da proporção de alarmes falsos para as diferentes posições do cluster, à medida que aumenta o tamanho do cluster.

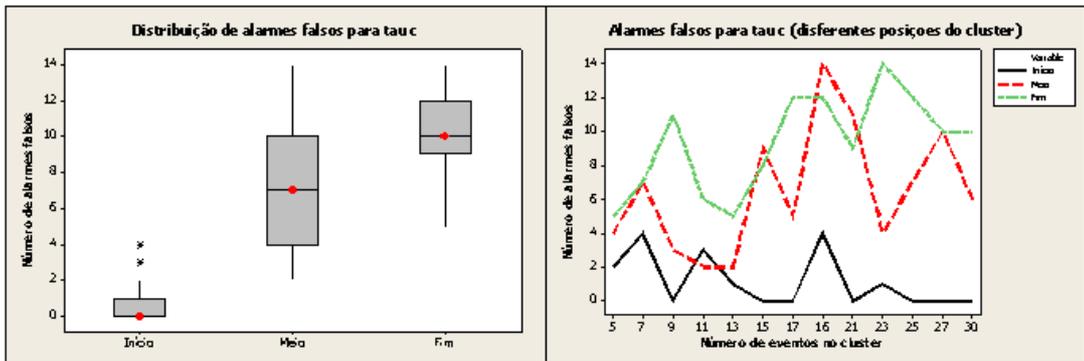


Figure 9: Box-plot com distribuição de alarmes falsos em  $\tau_c$  para as três posições do cluster, independente do tamanho do cluster. Gráfico de linhas da proporção de alarmes falsos para as diferentes posições do cluster, à medida que aumenta o tamanho do cluster.

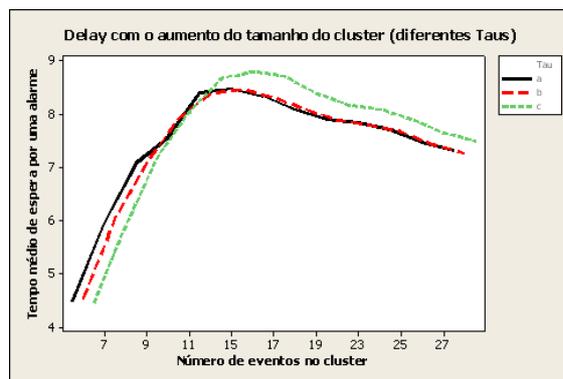


Figure 10: Gráficos de linhas comparando os tempos médios de espera por um alarme motivado (delay), para os três  $\tau$ 's, à medida que aumentamos o número de eventos no cluster.

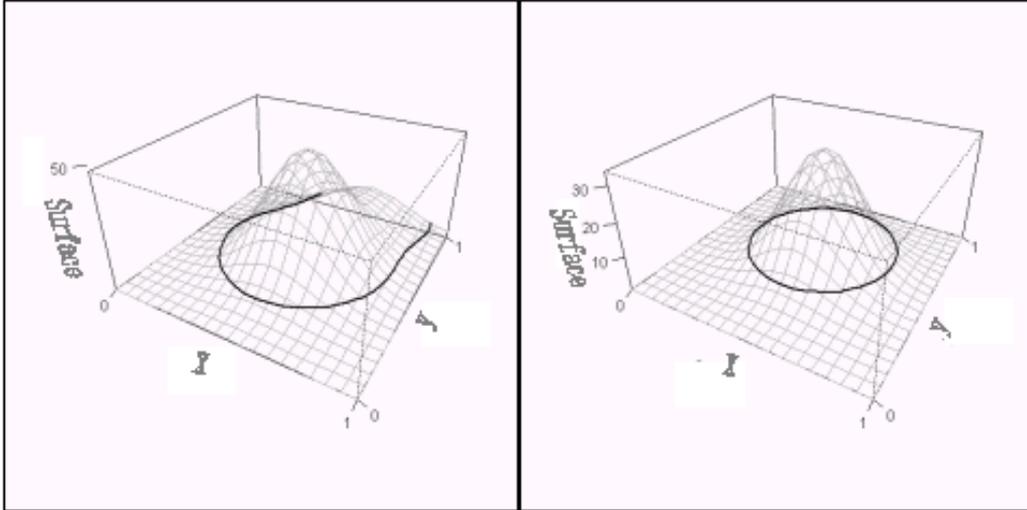


Figure 11: Figuras com superfície acumulada seccionada em diferentes patamares da superfície.

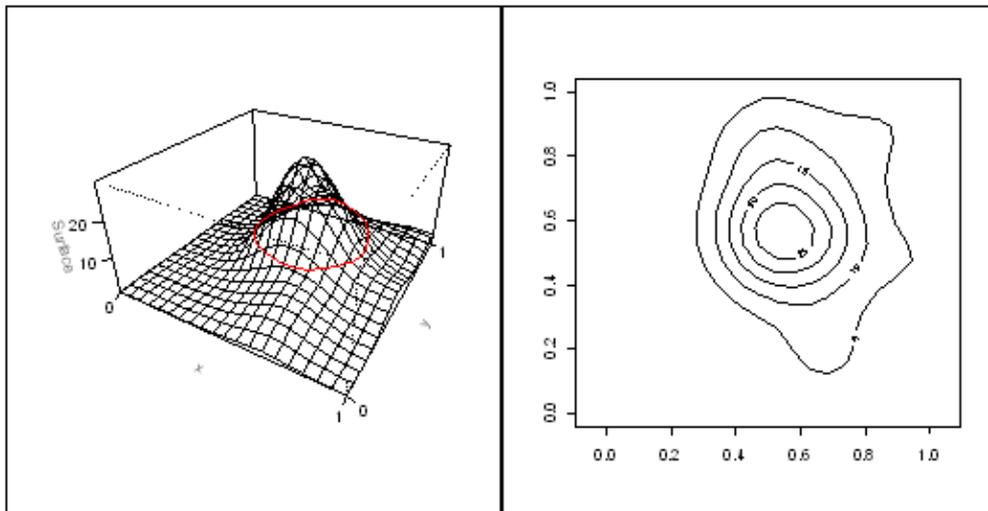


Figure 12: Superfície formada no momento em que o alarme soa, cortada pelo nível de contorno correspondente ao percentil 30%. Gráficos de níveis de contorno.

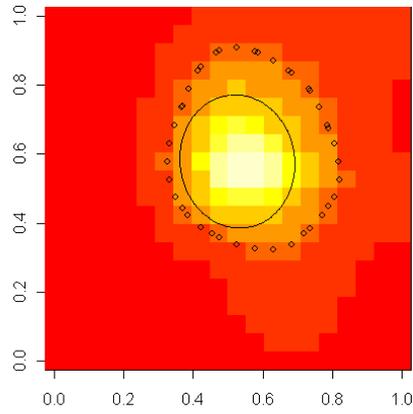


Figure 13: A figura mostra o gráfico de imagem que é formado no momento de ocorrência do 87<sup>o</sup> evento. Os pontos representam os eventos do cluster e a elipse as coordenadas estimadas da provável localização do cluster.

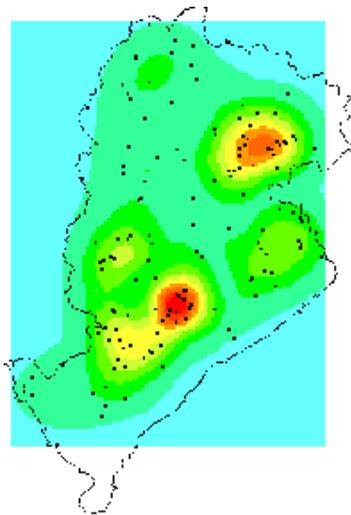


Figure 14: Mapa de kernel mostrando a distribuição da ocorrência de meningite meningocócica, na cidade de Belo Horizonte, entre os anos de 1998-2000.

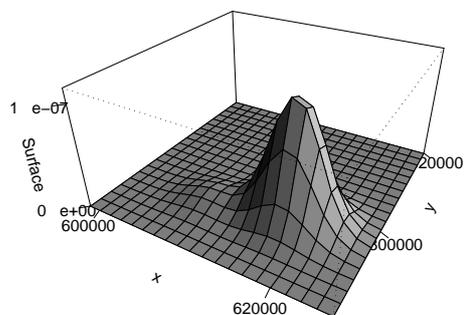


Figure 15: Superfície formada no 94º evento, para os casos de Meningite Meningocócica, BH 1998-2000.

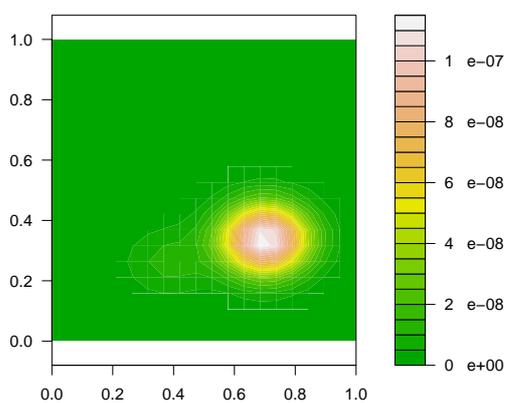


Figure 16: Gráfico de imagem da superfície no momento do 94º evento, destacando a região com ocorrência excessiva de casos de Meningite no passado recente.

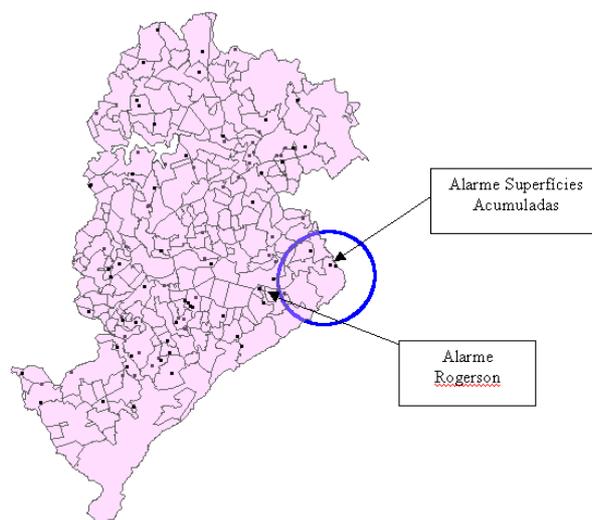


Figure 17: Distribuição de casos de Meningite, indentificando eventos que soaram o alarme no método Knox+CUSUM e no método das superfícies acumuladas, incluindo identificação do cluster (BH, 1999).

# Livros Grátis

( <http://www.livrosgratis.com.br> )

Milhares de Livros para Download:

[Baixar livros de Administração](#)

[Baixar livros de Agronomia](#)

[Baixar livros de Arquitetura](#)

[Baixar livros de Artes](#)

[Baixar livros de Astronomia](#)

[Baixar livros de Biologia Geral](#)

[Baixar livros de Ciência da Computação](#)

[Baixar livros de Ciência da Informação](#)

[Baixar livros de Ciência Política](#)

[Baixar livros de Ciências da Saúde](#)

[Baixar livros de Comunicação](#)

[Baixar livros do Conselho Nacional de Educação - CNE](#)

[Baixar livros de Defesa civil](#)

[Baixar livros de Direito](#)

[Baixar livros de Direitos humanos](#)

[Baixar livros de Economia](#)

[Baixar livros de Economia Doméstica](#)

[Baixar livros de Educação](#)

[Baixar livros de Educação - Trânsito](#)

[Baixar livros de Educação Física](#)

[Baixar livros de Engenharia Aeroespacial](#)

[Baixar livros de Farmácia](#)

[Baixar livros de Filosofia](#)

[Baixar livros de Física](#)

[Baixar livros de Geociências](#)

[Baixar livros de Geografia](#)

[Baixar livros de História](#)

[Baixar livros de Línguas](#)

[Baixar livros de Literatura](#)  
[Baixar livros de Literatura de Cordel](#)  
[Baixar livros de Literatura Infantil](#)  
[Baixar livros de Matemática](#)  
[Baixar livros de Medicina](#)  
[Baixar livros de Medicina Veterinária](#)  
[Baixar livros de Meio Ambiente](#)  
[Baixar livros de Meteorologia](#)  
[Baixar Monografias e TCC](#)  
[Baixar livros Multidisciplinar](#)  
[Baixar livros de Música](#)  
[Baixar livros de Psicologia](#)  
[Baixar livros de Química](#)  
[Baixar livros de Saúde Coletiva](#)  
[Baixar livros de Serviço Social](#)  
[Baixar livros de Sociologia](#)  
[Baixar livros de Teologia](#)  
[Baixar livros de Trabalho](#)  
[Baixar livros de Turismo](#)