**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

# Resumo

Sistemas de Vigilância têm importantes aplicações em diversas questões práticas. Freqüentemente, 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

# 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.

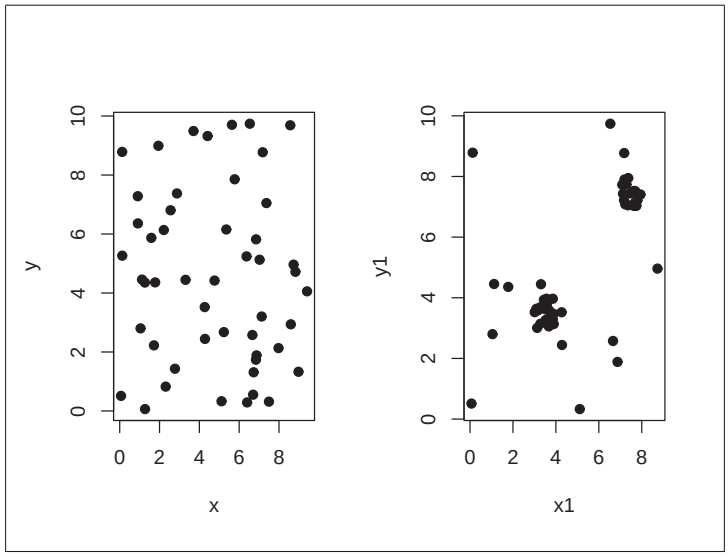


Figura 2.1: Exemplos de processos pontuais.

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 freqüências 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.

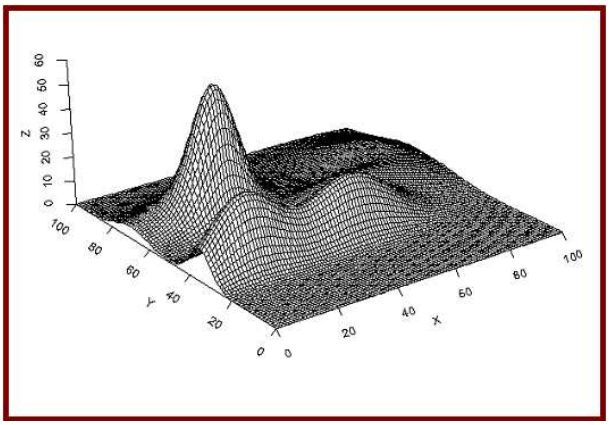


Figura 2.2: Visualização de densidade de kernel.

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.

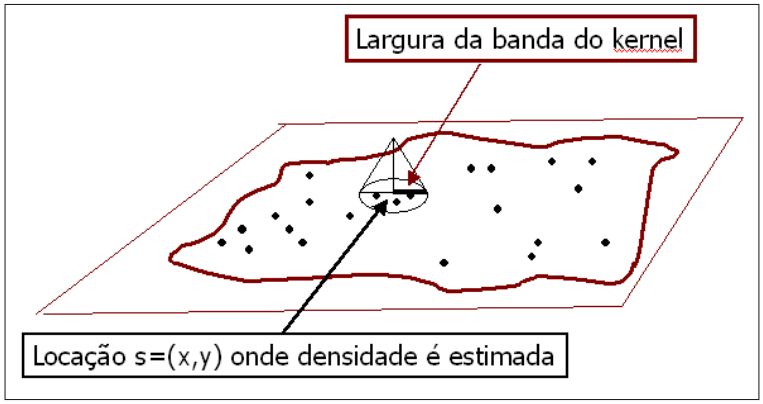


Figura 2.3: Estimador de intensidade de distribuição de pontos - kernel triangular.

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 $(u_1, u_2, \ldots, u_n)$ 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 $(u_i, \ldots, u_{i+m-1})$ 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.

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

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 $\tau \to 0$, 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 $\tau \to \infty$, 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:

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

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 decrescimento é gradual. A função triangular pesa pontos próximos mais do que pontos distantes dentro do círculo, mas o decrescimento 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:

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

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:

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

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

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:

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

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:

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

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

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 definí-los conforme as equações abaixo:

$$K = k\frac{\sigma}{\sqrt{n}} \tag{2.9}$$

$$H^+ = h\frac{\sigma}{\sqrt{n}} \tag{2.10}$$

$$H^- = -h\frac{\sigma}{\sqrt{n}} \tag{2.11}$$

onde $k$ e $h$ são constantes (freqüentemente $k = 0.5$ 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 obsevados 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 $ARL = 1/P(Alarme falso)$.

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:

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

## 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 $X = \{X(t);\ t = 1, 2, \ldots\}$ 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

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

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 ($ARL_0$).

$$ARL_0 = E[t_A \mid \lambda = \infty] \tag{2.14}$$

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:

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

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:

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

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 na 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.; Hjalmars,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 Scientifc 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 Worshop, 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*