UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Sistemas Prospectivos para Vigilância Espaço-tempo

Thais Rotsen Correa

Tese de doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Doutor em Estatística.

Orientador: Prof. Dr. Renato Martins Assunção

Belo Horizonte/MG, 09 de Novembro de 2011

Agradecimentos

Este trabalho não teria sido possível sem a ajuda de muitas pessoas, às quais agradeço imensamente.

Agradeço em primeiro lugar a Deus, por me iluminar, me dar força e saúde durante toda esta caminhada.

À minha família que eu amo tanto, e que é a minha base. Ao meu pai, que sempre vibrou com as minhas vitórias. À minha mãe, que esteve sempre ao meu lado me motivando e me ajudando. À minha irmã, pela amizade e confiança.

Ao Felipe, por todo amor, carinho e paciência, por me fazer feliz!

À Fátima e Bia, que sempre me apoiaram.

Aos amigos da UFOP, que sempre acreditaram no meu sucesso. Em especial, agradeço ao Anderson pelas várias ajudas, sem as quais teria sido tudo muito mais difícil. Ao Júlio e ao Flávio, pelos cafés providencias para esfriar a cabeça.

Ao Renato, por todos estes anos de orientação e ensinamentos e pelo exemplo de profissional que é e sempre foi para mim.

Ao pessoal do LESTE, sempre presente em partes importantes (e divertidas) desta caminhada.

Aos professores e funcionários do Departamento de Estatística da UFMG, em especial ao Marcelo e à Rogéria.

Aos membros da banca, professores Glaura Franco, Fábio Demarqui, Francisco Louzada Neto e Ronaldo Dias, pelas sugestões e considerações.

À todos aqueles, amigos e familiares, que torceram por mim.

Resumo

A demanda por sistemas capazes de detectar mudanças nos padrões espacial e temporal de ocorrência de eventos tem crescido em diversas áreas do conhecimento. Os avanços tecnológicos ocorridos nos últimos anos facilitaram a coleta e análise da informação geográfica, causando um aumento no interesse por sistemas para detecção de conglomerados espaço-tempo. O estudo deste tipo de sistema é recente e ainda existem poucas propostas.

Nesta tese nós estudamos sistemas prospectivos de vigilância espaço-temporal utilizando dados pontuais e dados de área. Nós propomos dois sistemas, uma para dados pontuais e outro para dados de área, ambos baseados na estatística de Shiyayaev-Roberts. A eficiência destes sistemas é avaliada via simulação. Dados reais são utilizados para ilustrar a aplicação destes sistemas. Nós propomos também outros três sistemas, com propriedades parecidas, para dados pontuais. Simulações também são utilizadas para verificar o desempenho destes sistemas. Finalmente, nós fazemos uma análise crítica quanto ao uso da estatística scan no contexto prospectivo.

Abstract

The demand for systems capable of detecting changes in spatial and temporal patterns of occurrence of events has grown in several areas of the knowledge. Technological advances in recent years have facilitated the collection and analysis of geographic information, causing an increase in interest in systems for detecting space-time clusters. The study of this type of system is still new and there are few proposals.

In this thesis we study systems of prospective space-time surveillance using point data and areal data. We propose two systems, one for point data and one for areal data, both based on the Shiyayaev-Roberts statistic. The efficiency of these systems is evaluated via simulation. Real datasets are used to illustrate the application of these systems. We propose three other systems, with similar properties, for point data. Simulations are also used to check the performance of these systems. Finally, we make a critical analysis about the use of the scan statistic in the prospective context.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Atualmente, os registros feitos pela maioria das agências de saúde pública trazem informações sobre o tempo e o local de ocorrência dos casos das principais doenças que atingem uma população. A informação geográfica é coletada e analisada regularmente, através do uso de softwares de geoprocessamento.

Uma vez que os registros são atualizados com uma frequência cada vez maior, as agências passaram a demandar métodos e softwares apropriados para a detecção precoce de mudanças nos padrões espacial e temporal de ocorrência destes eventos. Este tipo de método é útil não apenas no contexto de saúde pública. Métodos para detecção prospectiva de mudanças no padrão espacial ou temporal dos valores de um processo estocástico, de forma rápida e eficiente, são de grande interesse em diversas áreas do conhecimento, tais como vigilância de acidentes de trânsito, crimes em grandes cidades, dinâmica ecológica, etc.

Nesta tese, são estudados métodos prospectivos de vigilância espaço-temporal. Dois tipos de dados estatísticos são considerados. O primeiro deles é constituído pelos processos pontuais espaço-temporais. Neste tipo de dados, observamos eventos pontuais da forma (x_i, y_i, t_i), onde (x_i, y_i) são as coordenadas geográficas e t_i é a coordenada temporal do evento. Exemplos de situações dando origem a este tipo de dados são as ocorrências de uma doença dentro de uma região geográfica e o monitoramento de queimadas na Amazônia.

O segundo tipo de dados considerado são os dados de área. Considere uma região geográfica particionada em H áreas denotadas por $i = 1, \ldots, H$. A observação na área i no instante de tempo t será denotada por y_{it}. Neste trabalho, os dados de área terão sempre uma estrutura de tempo discreto com $t = 1, 2, \ldots$ Assim, os dados podem ser vistos como uma série temporal de mapas onde, em cada tempo t, temos um mapa composto pelas observações $y_{1t}, \ldots, y_{Ht}$.

Com estes dois tipos de dados, estaremos ocupados em estudar métodos de vigilância prospectiva. Para isto, é útil ver as duas estruturas de dados como processos estocásticos $\{X(t); t > 0\}$, onde $X(t)$ depende do tipo de dado. No caso de dados de processos pontuais a tempo contínuo, temos

$$X(t) = \{N(x, y, t)\},$$

onde $N(x, y, t)$ é o número de eventos ocorridos na posição (x, y) desde o início do processo em $t = 0$ até um instante arbitrário t. Seja Y_{it} a variável aleatória observada na área i no tempo t. No caso de dados de área, temos

$$X(t) = (y_{1t}, \ldots, y_{Ht}),$$

o mapa de valores de Y_{it} no instante t.

No caso de processos pontuais ditos *simples*, temos $\mathbb{P}(N(x,y,t) \geq 2) = 0$ para todo (x,y,t). Isto é, em cada posição (x,y), podemos ter no máximo um evento em qualquer instante de tempo. Isto significa que o processo pontual espaço-temporal a tempo contínuo pode ser equivalentemente visualizado como um conjunto de pontos ou de setas no espaço tridimensional. A Figura 1 mostra um exemplo de eventos de um processo pontual vistos desta forma. Cada evento é representado por uma seta. A coordenada temporal t é dada pela altura da seta.

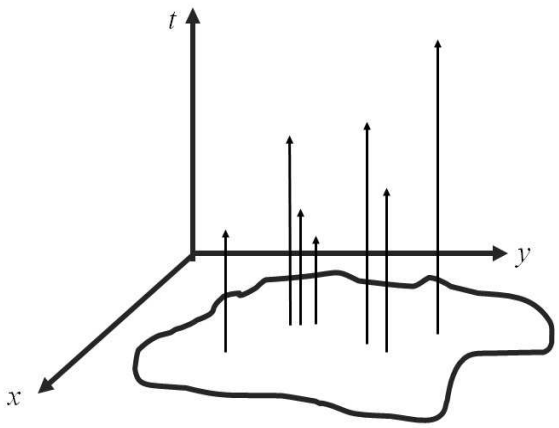

Figura 1: Exemplo de um processo pontual espaço-temporal a tempo contínuo visualizado como um conjunto de setas no espaço tridimensional, onde cada seta representa um evento. A altura da seta é igual à coordenada temporal.

Assuma que existe um modelo descrevendo a distribuição do processo estocástico $X(t)$. Isto é, existe uma família consistente de distribuições $\{\mathcal{F}_t, t > 0\}$ tal que $\mathcal{F}_t$ descreve a distribuição da família $\{X(s), 0 \leq s \leq t\}$. Seja τ o instante em que ocorre uma mudança no processo $X(t)$. O instante τ é desconhecido. Se a distribuição do processo não mudar, então $\tau = \infty$. Dizemos que o processo está *sob controle* em um tempo qualquer t se $\tau > t$. Quando $t > \tau$, dizemos que o processo está *fora de controle* em t.

Um *sistema de vigilância prospectivo* pode ser definido como um critério para decidir, em cada instante de tempo t, se o processo está sob controle ou não. Quando temos evidência suficiente de que o status do processo mudou do estado sob controle para o estado fora de controle, dizemos que temos um *alarme*. Podemos ter um alarme no instante $t < \tau$. Neste caso, temos um falso alarme. Alternativamente, podemos ter um alarme soando num instante $t > \tau$. Neste caso, temos um alarme motivado.

A situação ideal de um sistema de vigilância prospectivo é nunca soar um alarme enquanto o processo estiver sob controle e soar um alarme tão logo o processo saia de controle.

É claro que esta situação ideal dificilmente ocorre na prática. Assim, almejamos encontrar um sistema em que o alarme soe pouco quando o processo está sob controle, e que seja rápido para detectar mudanças no processo, quando elas acontecerem. A área do conhecimento denominada estatística já desenvolveu vários métodos de vigilância no caso puramente temporal. Esta área da estatística é conhecida como Controle de Qualidade (veja Ryan, 1989, Wetherill and Brown, 1991 e Montgomery, 1996).

1.1 Métodos de Vigilância em Controle de Qualidade

Suponha que o processo estocástico é simplesmente uma série temporal $Y_1, Y_2, \ldots$. A distribuição $\mathcal{F}_t$ corresponde à distribuição do vetor t-dimensional $(Y_1, \ldots, Y_t)$. Quando $\mathcal{F}_t$ muda de um modelo para outro, o status do processo muda de sob controle para fora de controle. No caso mais simples possível, suponha que $Y_1, Y_2, \ldots$ sejam independentes e identicamente distribuídas com distribuição acumulada $F_0(y)$. No instante τ, as variáveis aleatórias Y_i passam a seguir a distribuição acumulada $F_1(y)$. O interesse é soar um alarme o mais rapidamente possível após τ e, ao mesmo tempo, evitar que o alarme soe enquanto $t < \tau$.

Um dos métodos mais utilizados é a carta de controle de Shewart (veja, por exemplo, Ryan, 1989, Wetherill and Brown, 1991 e Montgomery, 1996). Este método assume que $Y_1, Y_2, \ldots$ são variáveis aleatórias independentes e identicamente distribuídas tal que $F_0(y)$ é a distribuição acumulada Normal com média μ e variância σ^2. A carta de contole de Shewart soa um alarme no tempo t se o valor de Y_t for inferior a um limite L_i ou superior a um limite L_s. Os limites L_i e L_s da carta de Shewart são dados por

$$L_i = \mu - c\sigma \;\text{ e }\; L_i = \mu + c\sigma\,,$$

onde c é a distância dos limites de controle L_i e L_s em relação à μ, expressa em unidades de desvio padrão.

Em geral, usa-se $c = 3$. Se $Y_t \sim \text{Normal}(\mu, \sigma^2)$, então $\mathbb{P}(\mu - 3\sigma \leq Y_t \leq \mu + 3\sigma) = 0,9973$. Logo, se $y_t \notin [\mu - 3\sigma, \mu + 3\sigma]$, temos um forte índicio de que a distribuição subjacente de Y_t não é mais $\text{Normal}(\mu, \sigma^2)$.

O método de Shewart é muito útil quando o processo sofre mudanças muito bruscas. Ele é um método bastante simples, porém insensível a mudanças de pequena magnitude ou mudanças graduais − quando o processo vai pouco a pouco passando do estado sob controle para um estado cada vez mais fora de controle. Este tipo de mudança gradual é melhor detectada pelo método de somas acumuladas (veja, por exemplo, Ryan, 1989, Wetherill and Brown,

1991 e Montgomery, 1996). A soma acumulada no tempo t é dada por

$$S_t = \max(0, S_{t-1} + y_t - k), t \geq 1, S_0 = 0\,,$$

onde k é uma constante que depende apenas dos parâmeros das distribuições acumuladas $F_0(y)$ e $F_1(y)$.

O método de somas acumuladas (CUSUM) soa um alarme no tempo t se S_t excede um limiar h. A escolha do limiar h está associada ao Average Run Length (*ARL*), definido como o tempo médio até que o alarme soe dado que o processo está sob controle. O *ARL* será definido formalmente na seção 4. Valores altos de h estão associados a valores altos para o *ARL* e valores baixos de h estão associados a valores baixos para o *ARL*. Nos casos em que o valor do *ARL* é alto, teremos poucos alarmes falsos, mas mudanças reais no status do processo não serão detectadas tão rapidamente. Da mesma forma, quando o valor do *ARL* é baixo, as mudanças reais geralmente serão detectadas mais rapidamente, mas alarmes falsos também serão mais freqüentes.

A constante k é escolhida de forma que, para um dado limiar h, k minimiza o tempo médio necessário para que o alarme soe motivadamente. Se $F_0(y)$ é a distribuição acumulada Poisson com média λ_0 e $F_1(y)$ é a distribuição acumulada Poisson com média λ_1, então

$$k = \frac{\lambda_1 + \lambda_0}{\log(\lambda_1) + \log(\lambda_0)}\,.$$

Se $F_0(y)$ é a distribuição acumulada Normal com média μ_0 e variância σ^2 e $F_1(y)$ é a distribuição acumulada Normal com média μ_1 e variância σ^2, então $k = (\mu_0 + \mu_1)/2$. Se $\mu_1 = \mu_0 + 2w\sigma$, então S_t acumulada acumula desvios da média μ_0 que excedem w desvios padrão

$$S_t = \max(0, S_{t-1} + y_t - \mu_0 - w\sigma), t \geq 1, S_0 = 0\,.$$

Tanto a carta de Shewhart quanto o CUSUM assumem que as observações são independentes no tempo, o que não é um suposição realista em várias aplicações. Kenett and Pollak (1996) propõem um sistema de monitoramento puramente temporal, baseado na estatística de Shiryayev-Roberts, que não requer independência entre as observações. Seja $f_k(Y_1, Y_2, \ldots, Y_n)$ a densidade conjunta das primeiras n variáveis quando $\tau = k$. Seja $f_\infty(Y_1, Y_2, \ldots, Y_n)$ a densidade conjunta das primeiras n variáveis quando $\tau = \infty$. A estatística de Shiryayev-Roberts é dada pela soma da razão de verossimilhança f_k / f_∞ sob todos os valores possíveis para o

ponto de mudança k. Quando esta estatística ultrapassa um certo limiar, o alarme soa. O sistema proposto por Kenett and Pollak (1996) é descrito na seção 5.

1.2 Métodos Espaço-temporais Prospectivos

Existe hoje um grande interesse em estender sistemas de vigilância puramente temporais para o caso espaço-tempo, devido à demanda originada pela recente facilidade na coleta da informação geográfica. No contexto de vigilância espaço-tempo, os métodos são recentes e ainda não há nenhum método amplamente aceito como sendo superior aos demais. Descrevemos a seguir algumas das recentes propostas de métodos espaço-temporais para vigilância prospectiva.

Kulldorff (2001) sugere o uso da estatística scan espaço-tempo, baseada na estatística scan espacial (Kulldorff, 1997), para monitoramento prospectivo de doenças em dados de área. A varredura no caso espaço-tempo é feita sob todos os cilindros vivos e busca-se aquele que maximiza a razão de verrossimilhança, baseada no modelo Poisson ou Bernoulli.

Rogerson (2001) propôs o uso de uma estatística de Knox local para monitoramento de conglomerados espaço-temporais. Marshall et al. (2007) demonstrou que, para um processo sob controle, o tempo de espera até o alarme do método proposto por Rogerson (2001) é influenciado pela densidade populacional e pela forma da região. Como consequência, o valor nominal das medidas de desempenho não são válidas. Marshall et al. (2007) também encontrou vários problemas com as aproximações usadas em Rogerson (2001).

Kulldorff et al. (2005) desenvolveram uma estatística scan espaço-tempo de permutação que não requer dados da população de risco. Uma das vantagens desta estatística é que ela pode ser aplicada a dados de processos pontuais.

Diggle et al. (2005) e Rodeiro and Lawson (2006) sugeriram métodos bayesianos para modelar a evolução espaço-temporal de taxas de incidência e monitorar mudanças.

Takahashi et al. (2008) propuseram uma estatística scan espaço-tempo flexível para detecção de conglomerados não circulares. Ao contrário da estatística scan espaço-tempo de Kulldorff (2001), que considera uma janela cilíndrica tridimensinal de base circular, a estatística scan flexível considera uma janela prismática tridimensinal cuja base tem formato arbitrário.

Tango et al. (2011) argumentam que a estatística scan espaço-tempo de Kulldorff (2001) compara o número observado de casos com o número esperado condicional e sugerem uma nova estatística scan espaço-tempo que compara o número observado de casos com o número

esperado não condicional.

Corberán-Vallet and Lawson (2011) aplicam a ordenada preditiva condicional ao contexto de vigilância para detectar pequenas áreas em que a incidência de uma doença é maior. A ordenada preditiva condicional é uma ferramenta bayesiana que permite detectar observações atípicas.

Frisén et al. (2011) consideram o caso de vigilância multivariada. Eles estudam processos em que as mudanças ocorrem simultaneamente ou em intervalos de tempo conhecidos. O princípio da suficiência é utilizado para esclarecer a estrutura de alguns problemas, encontrar métodos eficientes e determinar métricas de avaliação apropriadas.

O desenvolvimento de métodos espaço-temporais é portanto uma área de pesquisa importante e bastante ativa, ainda sem uma definição clara de métodos ótimos. Por isto, o foco desta tese é o desenvolvimento destes métodos, como explicamos a seguir.

2 Objetivos

2.1 Objetivos Gerais

Os dois principais objetivos desta tese são:

1. Desenvolver sistemas de vigilância espaço-tempo prospectivos para detecção de conglomerados emergentes.

2. Analisar alguns aspectos dos sistemas de vigilância baseados na estatística scan proposta por Kulldorff (2001).

2.2 Objetivos Específicos

O primeiro objetivo desta tese é desenvolver sistemas de vigilância para dados pontuais e dados de área. Buscamos métodos automáticos, com pouca interferência do usuário e que não exijam a especificação dos padrões marginais espacial e temporal. A taxa de alarme falso deve estar controlada e o método deve ser rápido para detectar mudanças reais. Procuramos por métodos simples, cujos resultados são de fácil interpretação para um usuário sem conhecimento estatístico.

O segundo objetivo é analisar, de forma crítica e detalhada, o sistema de vigilância proposto por Kulldorff (2001). Este sistema utiliza uma estatística scan espaço-tempo, que é também usada em vários outros sistemas de vigilância (por exemplo Takahashi et al., 2008, Tango et al., 2011 e Kulldorff et al., 2005). Esta estatística tem sido amplamente utilizada. Na nossa opinião, ela não é adequada para o contexto prospectivo. Existem alguns pontos importantes que devem ser analisados cuidadosamente quando esta estatística é utilizada neste contexto.

3 Organização

O corpo principal desta tese é formado pela coleção de quatro trabalhos que tratam de sistemas de vigilância prospectivos para a detecção de conglomerados espaço-temporais. O primeiro deles considera um sistema de vigilância para dados pontuais baseado em uma versão espacial da estatísitica de Shiryayev-Roberts (Kenett and Pollak, 1996). Este artigo, intitulado "Surveillance to detect emerging space-time clusters", foi publicado no volume 53 do periódico *Computational Statistics and Data Analysis*. Este artigo é apresentado na seção 6.

O segundo trabalho desta tese considera um sistema de vigilância para dados de área, também baseado na versão espacial da estatística de Shiryayev-Roberts (Kenett and Pollak, 1996). Este trabalho está condensado no segundo artigo, intitulado "Prospective space-time surveillance for areal data", que será submetido ao periódico *Statistics in Medicine*. Este artigo é apresentado na seção 7.

A análise do sistema de vigilância proposto por Kulldorff (2001) é o conteúdo do terceiro artigo desta coleção, intitulado "A close look on prospective surveillance using a scan statistic". Este artigo, que foi submetido ao periódico *Biometrics* em outubro de 2011, é apresentado na seção 8.

O quarto e último trabalho desta tese, apresentado na seção 9, considera outros três sistemas de vigilância para dados pontuais. Este trabalho ainda não gerou resultados relevantes o suficiente para constituir um artigo científico.

O restante deste capítulo está organizado da seguinte forma. Na seção 4 definimos duas medidas importantes no contexto de vigilância prospectiva: Average Run Length e Conditional Expected Delay. Na seção 5 apresentamos a estatística de Shiryayev-Roberts, conforme Kenett and Pollak (1996). Esta estatística é a base de dois dos trabalhos mencionados acima. As seções 6 a 9 trazem os trabalhos mencionados acima.

4 Tempo Médio de Espera pelo Alarme

No contexto de vigilância prospectiva, a utilização das tradicionais probabilidades dos erros tipos I e II não fazem sentido, pois não temos um tamanho de amostra fixo. A cada novo evento, temos que decidir se o processo está sob controle ou não. Como os métodos prospectivos são aplicados de forma sequencial, muitas vezes as probabilidades dos erros tipo I e II são iguais a 1 e 0, respectivamente. Considere, por exemplo, a carta de controle de Shewart, um método de vigilância prospectiva bastante simples, para uma sequência de variavéis aleatórias $Y_1, Y_2, \ldots$ independentes e identicamente distribuídas. Suponha o caso simples $Y_i \sim N(0,1)$ quando o processo está sob controle e $Y_i \sim N(1,1)$ quando o processo está fora de controle. A carta de Shewart para detecção de um aumento na média indica que o processo está fora de controle quando Y_i supera um limiar c pela primeira vez. Se o procedimento é aplicado indefinidamente para um processo sob controle, temos que $\mathbb{P}(\min_i\{Y_i > c,\ i = 1,2,...\} < \infty | \tau = \infty) = 1$. Ou seja, a probabilidade do erro tipo I é igual a 1. De forma análoga, para um processo fora de controle, a probabilidade do erro tipo II é igual a 0, pois $\mathbb{P}(\bigcap_{i=1}^{\infty}\{Y_i \leq c\} | \tau < \infty) = 0$. Por este motivo, as medidas tradicionalmente utilizadas para caracterizar o comportamento do processo sob controle e fora de controle no contexto prospectivo são, respectivamente, o Average Run Lenght (ARL) e o Conditional Expected Delay (CED).

Considere o processo estocástico $\{X(t); t > 0\}$. Um tempo de parada com respeito a $X(t)$ é um tempo aleatório T_A tal que, para cada $t > 0$, o evento $\{T_A = t\}$ é completamente determinado pela informação total conhecida até o tempo t, ou seja, pelos eventos contidos em $\{X_0, \ldots, X_t\}$. Se o processo estocástico é dado pela série temporal $Y_1, Y_2, \ldots$, então um tempo de parada com respeito a sequência de variáveis aleatórias $Y_1, Y_2, \ldots$ é uma variável aleatória T_A com a seguinte propriedade: para cada tempo t, a ocorrência ou não ocorrência do evento $\{T_A = t\}$ depende somente dos valores de $Y_1, Y_2, \ldots, Y_t$.

Um sistema de vigilância consiste em um tempo de parada, T_A, no qual considera-se que há evidência suficiente para acreditar que o processo está fora de controle. No contexto prospectivo, o tempo T_A é conhecido como o tempo em que o alarme soa, ou seja, um alarme soa quando há evidência empírica de que o processo está fora de controle.

O Average Run Length (ARL) é definido como o tempo médio de espera até que o alarme soe, dado que o processo está sob controle:

$$ARL = E\left[T_A | \tau = \infty\right].$$

O Conditional Expected Delay (CED) é definido como o tempo médio até que o alarme soe motivadamente. Isto é, o CED é o tempo médio de espera pelo alarme após o processo passar ao estado fora de controle:

$$CED(t) = E\left[T_A - \tau | T_A \geq \tau = t\right].$$

5 Sistema de Vigilância Shiryayev-Roberts

O sistema de vigilância Shiryayev-Roberts (Kenett and Pollak, 1996) é um sistema puramente temporal que se destaca por apresentar algumas vantagens em relação ao CUSUM e à carta de controle de Shewhart. Usaremos *SR* para denotar o sistema de vigilância Shiryayev-Roberts.

5.1 Descrição do Sistema *SR*

Suponha uma sequência de variáveis aleatórias $Y_1, Y_2, \ldots$, não necessariamente independentes. Seja $f_k(Y_1, Y_2, \ldots, Y_n)$ a densidade conjunta das primeiras n variáveis dado $\tau = k$. Seja $f_\infty(Y_1, Y_2, \ldots, Y_n)$ a densidade conjunta das primeiras n variáveis dado $\tau = \infty$.

Seja T_A o tempo de parada, ou seja, a primeira vez que o alarme soa, indicando que o processo está fora de controle. Seja $E_k(\cdot)$ a esperança com respeito a f_k. Então $E_\infty(T_A)$ é o *ARL*. Seja B o mínimo aceitável para o *ARL*, ou seja, $ARL = E_\infty(T_A) \geq B$, onde B é uma constante conhecida.

A estatística de Shiryayev-Roberts, R_n^{SR}, é dada pela soma da razão de verossimilhança f_k / f_∞ sob todos os valores possíveis para o ponto de mudança k:

$$R_n^{SR} = \sum_{k=1}^{n} \frac{f_k(Y_1, X_2, \ldots, Y_n)}{f_\infty(Y_1, X_2, \ldots, Y_n)}.$$

O alarme soa quando a estatística R_n^{SR} supera um limiar A. O tempo de parada T_A é dado por

$$T_A = \min\left\{ n \,|\, R_n^{SR} \geq A \right\}.$$

A questão se reduz então a encontrar o limiar A tal que $ARL = E_\infty(T_A) \geq B$. Supondo $\tau = \infty$, a sequência

$$\Lambda_{k,n} = \frac{f_k(Y_1, Y_2, \ldots, Y_n)}{f_\infty(Y_1, Y_2, \ldots, Y_n)}$$

é martingala com esperança unitária, mesmo com observações dependentes. Então,

$$R_n^{SR} - n = \sum_{k=1}^{n} (\Lambda_{k,n} - 1)$$

é martingala com esperança nula. Pelo Teorema da Amostragem Opcional temos:

$$E_\infty(R^{SR}_{T_A} - T_A) = 0\,,$$

e portanto

$$E_\infty\,(T_A) = E_\infty\,(R^{SR}_{T_A})\,.$$

Por definição, $R^{SR}_{T_A} \geq A$ e então $E_\infty\,(T_A) \geq A$. Logo, tomando-se $A = B$ a condição $E_\infty\,(T_A) \geq B$ é satisfeita.

Uma vez que T_A é o primeiro tempo em que a estatística R^{SR}_n excede o limiar A, o excesso é tipicamente pequeno, de forma que considerar $A = B$ gera um procedimento moderadamente conservador que satisfaz à equação $E_\infty\,(T_A) \geq B$. Então, o sistema *SR* soa um alarme quando $R^{SR}_n \geq A$ pela primeira vez, onde A é o valor desejado de $E_\infty\,(T_A) = ARL$.

5.2 Vantagens do Sistema *SR*

Pollak (1985) provou que, quando as observações são independentes e identicamente distribuídas, o sistema *SR* é assintoticamente ($B \to \infty$) ótimo no sentido de minimizar

$$\sup_{k\geq 1} E_k(T_A - k \mid T_A \geq k)$$

dentre todos os tempos de parada T_A tal que $E_\infty\,(T_A) \geq B$. Pollak and Tartakovsky (2009) mostraram, também para observações independentes e identicamente distribuídas, que o procedimento *SR* é estritamente ótimo no sentido de minimizar

$$\sum_{k=1}^{\infty} E_k(T_A - k; T_A \geq k)$$

para todo $B > 1$ na classe de procedimentos em que $E_\infty(T_A) \geq B$.

Os sistemas *SR* e CUSUM são similares, em termos do tempo até o alarme soar motivadamente (Shiryaev, 1963, Roberts, 1966, Mevorach and Pollak, 1991, Pollak and Siegmund, 1991 e Pollak and Siegmund, 1985).

A principal vantagem do sistema *SR* em relação ao CUSUM é a relativa facilidade de sua aplicação sob suposições mínimas. Ao contrário do CUSUM, *SR* não requer independência entre as observações. Além disso, o método *SR* geralmente detecta uma mudança tão rápido quanto o CUSUM e, em problemas com uma estrutura de parâmetros complicada, ele é

freqüentemente mais fácil de ser aplicado (Kenett and Pollak, 1996).

6 Vigilância espaço-tempo para detecção de conglomerado emergentes

Esta seção traz o artigo “Surveillance to detect emerging space-time clusters”, publicado no periódico *Computational Statistics and Data Analysis*.

6.1 Abstract

The interest is on monitoring incoming space-time events to detect an emergent space-time cluster as early as possible. Assume that point process events are continuously recorded in space and time. In a certain unknown moment, a small localized cluster of increased intensity starts to emerge. Its location is also unknown. The aim is to let an alarm to go off as soon as possible after its emergence but avoiding that it goes off unnecessarily. The alarm system should also provide an estimate of the cluster location. In addition to that, the alarm system should take into account the purely spatial and the purely temporal heterogeneity, which are not specified by the user. A space-time surveillance system with these characteristics using a martingale approach to derive the surveillance system properties is proposed. The average run length for the situation when there are clusters present in the data is appropriately defined and the method is illustrated in practice. The algorithm is implemented in a freely available stand-alone software and it is also a feature in a freely available GIS system.

6.2 Introduction

The ongoing and systematic collection of data, as well as its analysis, became essential long ago to the planning, implementation, and evaluation of public health practice. However, as more and more health data are stored in electronic form in a timely way, it is increasing the need for methods which can detect quickly anomalies in a continuously updated database with minimum input requirements from users. Currently, early disease outbreak detection systems are object of intense demand by government agencies, specially public health departments. One reason for this heightened interest on the subject is the threat of bioterrorism (Buehler et al., 2004, Henderson, 1999), but these systems have a much larger and much older application scope. In fact, early outbreak detection methods have always been a matter of concern to public health (Hardy, 2001). In the new context of spatially referenced data, these methods face important analytical challenges that include dealing with the adjustment for natural temporal and spatial variation, the unknown time, place and size of an emergent

cluster, detecting an outbreak as early as possible, and the lack of suitable population-at-risk data.

Most statistical methods in use for the early detection of disease outbreaks are purely temporal in nature (Sonesson and Bock, 2003, Höhle, 2007, Höhle and Paul, 2008). Hence, they are usually applied to monitor data from large area regions without concern to their geographical location within the monitored regions. They lack power to detect outbreaks that start locally since the affected areas are submerged into large regions with usual incidence rates. One possible solution is to partition the large region into small areas and to apply the purely temporal methods in each small area separately and in parallel. However, this procedure leads to a severe problem of multiple testing, generating many more false signals than the nominal statistical significance level indicate. As a consequence, these purely temporal methods are not appropriate when the data are collected with space and time information. In addition to these problems, there is the expectation that using the now readily available spatial information can facilitate the detection and localization of emergent clusters (Buckeridge et al., 2005).

The primary purpose of this paper is to suggest a method for the quick detection of space-time emergent clusters in a set of point process events. The requirements to establish a surveillance system, either accounting for spatial structure or not, are generally structured around a basic trade-off: the need for quickly detecting possible outbreaks must be balanced against the need to avoid a high rate of false alarm signals. Our method allows the user to control these trade-off elements in a simple way.

We introduce a stochastic model to describe eventually emerging spatial clusters with minimum requirement of user-defined parameters. When there are no clusters, we assume that the events' density is separable meaning that it is the product of arbitrary spatial and temporal functions. More importantly, our method does not require the functional specification of these purely spatial or the purely temporal functions. As an alternative to this model, we assume that somewhere, at some moment, one or more space-time high intensity clusters start to emerge. We develop a likelihood model for this pair of hypotheses and monitor the incoming events with a spatial version of the Shyriaev-Roberts statistic. The Shiryaev-Roberts statistic is well known on industrial statistics applications but it is not so common in biometrical work. We use the martingale structure of the Shiryaev-Roberts statistic to derive the values for the tuning parameters of our method.

The next Section contains a brief review of the prospective space-time methods available and introduces the main definitions and notation used in the paper. In Section 6.4, we present

our proposal using Shiryayev-Roberts control chart method based on martingales. Section 6.5 presents an analysis of the impact of tuning parameters. In Section 6.6 we show the results of a Monte Carlo study of the method performance. For this, we need to define appropriately what is the expected time until detection when there are clusters present in the data. We are specially interested in the effects of the tuning parameters in our method performance. Section 6.7 illustrates the use of the method to detect and to identify the particular events that are associated with the space-time clusters. We use the classic Burkitt's lymphoma dataset (Williams et al., 1978) and a Brazilian dataset of Meningitis cases in three years of observation. We analyzed the data using a freely available stand-alone software where our method is implemented. Finally, we close in Section 6.8 with a discussion and a summary of the main conclusions.

6.3 Prospective space-time surveillance for localized clusters

The traditional methods for space-time cluster detection are retrospective in nature. That is, they search in a database of past events for evidence of presence of space-time clusters. In contrast, our interest is on prospective methods for geographically restricted: an events' database is updated regularly and then an algorithm should run to help deciding on the emergence of localized space-time clusters. Hence, the clusters must be alive, in the sense that at least some of the most recent events belong to the eventually detected clusters. The regularly updating nature of the database brings two difficult problems. In the first place, the possibility of using too many significance tests as, for example, if one statistical test is carried out every time the database is updated. This induces a severe multiple testing problem with too many false alarms for clusters. As a consequence, such a method would be soon discredited as unreliable. In the second place, reducing in some way the false alarm rate could imply in a long delay to signal a truly emerging space-time cluster. The trade-off between these two problems must be explicitly recognized in any methodology.

A thorough literature review can be found in the book edited by Lawson and Kleinman (2005), in Sonesson and Bock (2003), and in Waller and Gotway (2004). We give here a brief overview of the main proposals. There are non-spatial, purely temporal methods derived from quality control ideas concerned with the monitoring of a stochastic process in time. The Shewart Chart Control is a very simple and popular method but it is not sensitive to small changes in the process. The Cumulative Sum (CUSUM) method accumulates the recent evidence to the previous data until a certain threshold is crossed. It is better than

Shewart to detect small changes in the purely temporal process and it has been shown that it has optimal properties in very simple scenarios (see Frisén, 2003). Exponentially weighted moving average also accumulates evidence, as the CUSUM method, but it discounts observations as they get old (Frisén, 2003). All these methods assume that data are independent in time, which is not a realistic assumption in many applications. Kenett and Pollak (1996) uses a Shiryaev-Roberts statistics to allow for dependent data. We review this work later in this section.

There are few space-time oriented proposals. One recent promising method has been suggested by Kulldorff (2001) who used a space-time scan statistic for area data. The main difficulty is the control of overall significance level for a sequence of periodic tests, although each individual test has error type I adjusted for all previous analysis at each time moment. We discuss this issue in more detail in the last section.

Rogerson (2001) suggested a statistic based on local Knox statistic that requires only cases data in the form of a space-time point process. Marshall et al. (2007) found severe problems with the probability approximations used in this methods, suggesting that it should not be used. Marshall et al. (2007) demonstrate that the ARL performance of the Rogerson method is highly influenced by some required threshold values, by the population density, and by the region shape. As a consequence, the nominal performance measures associated with Rogerson method are not valid and this makes impossible the tuning of the method without computer simulation.

Kulldorff et al. (2005) have developed a space-time permutation scan statistic for the early detection of disease outbreaks, which is currently in use by the New York City Department of Health for syndromic surveillance. They use a Poisson based likelihood ratio test statistic scanning over all possible cylinders as clusters candidates. This method does not control overall error type I level for the sequence of periodic analysis. Diggle et al. (2005) and Rodeiro and Lawson (2006) proposed a Bayesian method to model the space-time evolution of the incidence rate and to monitor for changes.

We base our proposal in the Shiryaev-Roberts (SR) surveillance method that was developed only for temporal processes (Shiryaev, 1963, Roberts, 1966, Kenett and Pollak, 1996). Suppose that a sequence of possibly dependent random variables $X_1, X_2, \ldots$ is observed. Two possible models are considered. In one, a sudden change in the stochastic process occurs at the unknown moment k and $f_k(x_1, x_2, \ldots x_t)$ is the joint density distribution of the first t random variables. In the second model, no change ever occurs. In this case, we assume that $k = \infty$ and write $f_\infty(x_1, x_2, \ldots x_t)$ for the joint density. Any surveillance method implies a

stopping time T, the first moment when the alarm goes off. Let $E_k(\cdot)$ be the expectation with respect to f_k. The mean $E_\infty(T)$ is called the *Average Run Length* and it is denoted by ARL^0. Clearly, it is desirable to make ARL^0 large. Typically, the user establishes an acceptable minimum threshold B for this parameter. That is, we want $ARL^0 = E_\infty(T) > B$, where B is known.

One approach would be to maximize the likelihood ratio over all possible values of the unknown parameter k defining the statistic

$$\max_{1\le k\le t} \frac{f_k(X_1,\ldots,X_t)}{f_\infty(X_1,\ldots,X_t)}$$

Rather than adopting this approach, the Shiryaev-Roberts statistic R_t uses the sum of likelihood ratios f_k/f_∞ for all possible change-point moments k:

$$R_t = \sum_{k=1}^{t} \frac{f_{(k)}(X_1,X_2,\ldots,X_t)}{f_\infty(X_1,X_2,\ldots,X_t)}\,.$$

The alarm goes off if R_t is too large, that is, if $R_t \ge A$. The stopping time T_A is defined as

$$T_A = \min\{t\,|R_t \ge A\}\,.$$

It remains to find A such that $ARL^0 = E_\infty(T_A) > B$.

Following the notation of Kenett and Pollak (1996), under P_∞, the sequence

$$\Lambda_{k,t} = \frac{f_k(X_1,X_2,\ldots,X_t)}{f_\infty(X_1,X_2,\ldots,X_t)}$$

is a martingale with expected value equal to 1, even with dependent observations. Therefore,

$$R_t - t = \sum_{k=1}^{t} (\Lambda_{k,t} - 1)$$

is a zero mean martingale. By the Optional Sampling Theorem, we have

$$E_\infty(R_{T_A} - T_A) = 0\,,$$

and therefore

$$E_\infty(T_A) = E_\infty(R_{T_A})\,.$$

By definition, $R_{T_A} \geq A$ and hence $E_\infty(T_A) \geq A$. Therefore, taking $A = B$ satisfies the condition $E_\infty(N_B) \geq B$.

There are several advantages associated with the Shiryaev-Roberts method in the time series context. First, it can be shown that it exhibits some optimal properties in some simple scenarios (Pollak, 1985). Pollak (1985) proved that the Shiryaev-Roberts procedure is asymptotically (as $B \to \infty$) optimal in the sense of minimizing the supremum average delay to detection

$$\sup_{k \geq 1} E_k(T - k \mid T \geq k) .$$

over all stopping times T that satisfy $E_\infty(T) \geq B$. Yakir (1997) found that this procedure is strictly optimal for the problem of minimizing the average run length to detection over all stopping times T that satisfy $E_\infty(T) \geq B$ when $X_1, X_2, \ldots, X_{k-1}$ are iid random variables. Furthermore, in terms of the delay time for the alarm going off after the purely temporal clusters starts to emerge, the Shiryaev-Roberts and the usual CUSUM method are similar (Shiryaev, 1963, Roberts, 1966, Pollak and Siegmund, 1985, Mevorach and Pollak, 1991). The Shiryaev-Roberts method does not require independence between observations. It can also be shown that Shiryaev-Roberts is at least as efficient as some optimal classical procedures (Kenett and Pollak, 1996).

One difficulty to use the Shiryaev-Roberts method is that it depends on the complete specification of the joint distribution of $X_1, \ldots, X_n$ after a change occurs at k. This is not simple to be done in the purely temporal context and the difficulty increases in the space-time situation. However, we suggest a solution, as we explain next.

6.4 Detection of emerging space-time clusters

6.4.1 A model for emerging clusters

Let N be a Poisson process in $\mathbb{R}^3$ partially observed in the three-dimensional region $\mathcal{A} \times (0, \mathcal{T}]$.

The events $(\mathbf{s}_i, t_i) = (x_i, y_i, t_i)$ are indexed by $i = 1, 2, \ldots$, and we assume that $t_1 < t_2 < \ldots$. Let $N(C)$ be the number of events in the set $C \subset \mathcal{A} \times (0, \mathcal{T}]$. We have $N(C)$ distributed as a Poisson random variable with mean $\mu(C)$ given by the integral of the intensity function $\lambda(x, y, t) \geq 0$ over C:

$$\mu(C) = \int_C \lambda(x, y, t)\, dx\, dy\, dt .$$

A special type of set C is a cylinder given by $C = B(\mathbf{s}, \rho) \times (t_a, t_b]$ where $B(\mathbf{s}, \rho)$ is the disc

centered at $\mathbf{s} = (x,y) \in \mathcal{A}$ with radius ρ, and $t_a < t_b$.

Let $\mu = \mu(\mathcal{A} \times (0,\mathcal{T}])$ be the expected number of events in the observation region and define the marginal spatial and temporal densities by

$$\lambda_S(x,y) = \mu^{-1} \int_{(0,\mathcal{T}]} \lambda(x,y,t)\, dt\,,$$

and

$$\lambda_{\mathcal{T}}(t) = \mu^{-1} \int_{\mathcal{A}} \lambda(x,y,t)\, dx\, dy\,,$$

respectively. Note that

$$\int_{\mathcal{A}} \lambda_S(x,y)\, dx\, dy = \int_{(0,\mathcal{T}]} \lambda_{\mathcal{T}}(t)\, dt = 1\,.$$

Given that an event $(\mathbf{s},t)$ occurred in $\mathcal{A} \times (0,\mathcal{T}]$, the functions $\lambda_S(x,y)$ and $\lambda_{\mathcal{T}}(t)$ represent the probability density of $\mathbf{s}$ and t, respectively.

We define now the pair of situations we will consider. The first one is that without space-time clusters. In this case, we have a separable intensity $\lambda(x,y,t) = \mu\lambda_S(x,y)\,\lambda_{\mathcal{T}}(t)$ where $\lambda_S(x,y)$ and $\lambda_{\mathcal{T}}(t)$ are arbitrary and unspecified. That is, they are nuisance parameters.

The alternative situation assumes that there exists a time τ, a constant $\varepsilon > 0$, and a cylinder $C = B(\mathbf{s},\rho) \times (t_a,t_b]$ (yet to be defined) such that

$$\lambda(x,y,t) = \mu\lambda_S(x,y)\,\lambda_{\mathcal{T}}(t)\,(1 + \varepsilon I_C(x,y,t))\,.$$

This intensity function can not be written as a product of two functions, one depending only in space and the other only in time. The parameter ε is the relative change on the events intensity within the cluster and it must be specified by the user. Assunção and Maia (2007) and Assunção et al. (2007) used a similar pair of alternative models in the context of space-time point process data.

To define a useful class of cylinders C, we start considering that, if a higher incidence cluster emerges, we must be able to detect it through the observed events. That is, non-events (or void spaces) do not bring information about an emerging cluster. Hence, we decided to constrain t_a to be equal to one of the observed events times t_k. Additionally, the cylinders should be centered around its corresponding spatial location $\mathbf{s}_k$. Since the interest is only on alive clusters, the endpoint t_b is equal to the time t_n of the current last observed event. That is, we consider cylinders of the form $B(\mathbf{s}_k,\rho) \times (t_k,t_n]$, with $k < n$, where $(\mathbf{s}_k,t_k)$ is one

previously observed event while $(\mathbf{s}_n, t_n)$ is the last observed event at a given moment. The disc $B(\mathbf{s}_k, \rho)$ has a radius ρ specified by the user. To simplify notation, we denote by $C_{k,n}$ the cylinder $B(\mathbf{s}_k, \rho) \times (t_k, t_n]$ with $k < n$. We extend the notation to include the case $k = n$, writing $C_{n,n}$ to represent the null set.

6.4.2 A sequential procedure to detect emerging clusters

To define a statistic, we consider the likelihood of the space-time Poisson processes when n events have been observed. If no cluster is emerging, we have

$$L_\infty = \left(\prod_{i=1}^{n} \lambda(x_i, y_i, t_i) \right) \exp\left(- \int_{R^3} \lambda(x,y,t)\, dx\, dy\, dt \right) .$$

Under the alternative scenario that a cluster started emerging at time $t_k < t_n$, we have

$$\begin{aligned} L_k &= \left(\prod_{i=1}^{n} \lambda(x_i, y_i, t_i)\, (1 + \varepsilon I_{C_{k,n}}(x_i, y_i, t_i)) \right) \\ &\quad \exp\left(- \int_{R^3} \lambda(x,y,t)\, dx\, dy\, dt \right) \exp\left(-\varepsilon \int_{C_{k,n}} \lambda(x,y,t)\, dx\, dy\, dt \right) , \end{aligned}$$

where $\lambda(x,y,t) = \mu \lambda_S(x,y)\, \lambda_{\mathcal{T}}(t)$ and $C_{k,n}$ is the putative cluster cylinder.

Therefore, a space-time version of the SR test statistic R_n becomes

$$\begin{aligned} R_n &= \sum_{k=1}^{n} \frac{L_k}{L_\infty} \\ &= \sum_{k=1}^{n} \left\{ \left[\prod_{i=1}^{n} (1 + \varepsilon I_{C_{k,n}}(x_i, y_i, t_i) \right] \exp\left(-\varepsilon \int_{C_{k,n}} \lambda(x,y,t)\, dx\, dy\, dt \right) \right\} \\ &= \sum_{k=1}^{n} (1+\varepsilon)^{N(C_{k,n})} \exp(-\varepsilon \mu(C_{k,n})) \qquad (1) \\ &= \sum_{k=1}^{n} \Lambda_{k,n} . \qquad (2) \end{aligned}$$

The expression $\Lambda_{k,n}$ can be seen as a contrast between the observed number $N(C_{k,n})$ of events in $C_{k,n}$ and its expected value under the no-cluster situation. In fact, if ε is small,

$$\Lambda_{k,n} \approx (1+\varepsilon)^{N(C_{k,n})}\, (1-\varepsilon)^{\mu(C_{k,n})} \approx 1 + \varepsilon \left(N(C_{k,n}) - \mu(C_{k,n}) \right) .$$

The parameter $\varepsilon > 0$ is known (user-specified) and measures the anticipated relative change in the events' density. Note that

$$\Lambda_{n,n} = (1+\varepsilon)^{N(C_{n,n,})} \exp(-\varepsilon\mu(C_{n,n})) = (1+\varepsilon)^0 \exp(0) = 1\,.$$

Our surveillance method calculates R_n as the n-th event arrives, substituting the unknown $\mu(C_{k,n})$ by an estimate $\hat{\mu}(C_{k,n})$. The estimation of $\mu(C_{k,n})$ is discussed in Section 6.4.3. The alarm goes off when $R_n \geq A$ for the first time.

In summary, the algorithm associated with our proposal requires as input:

- a set of n case events, specified by their spatial coordinates x, y and time t;
- the value of three user-specified tuning parameters:
  - the anticipated relative change ε in the density within the cluster;
  - the anticipated radius ρ for the cluster;
  - the threshold A, which should be approximately equal to the desired ARL^0.

Iteratively in n, we calculate R_n. The output is a sequence of values R_n where n is the number of events. If $R_n > A$ for any n, the alarm goes off.

If the alarm goes off, one important practical issue is the space-time cluster location. Suppose that the alarm goes off at the n-th event. That is, $R_t \geq A$ for the first time at $t = n$. Since

$$R_n = \sum_{k=1}^{n} \Lambda_{k,n} = \Lambda_{1,n} + \Lambda_{2,n} + \ldots \Lambda_{n,n}\,.$$

The large values of $\Lambda_{k,n}$ are those contributing to the alarm triggering. Let $\Lambda_{k*,n} = \max\{\Lambda_{k,n}, 1 \leq k \leq n\}$. A cluster estimate is built by taking the spatial coordinates (x_k*, y_k*) of the $k*$-th event as the center of the cylinder basis, and by taking its height equal to $[t_k*, t_n]$. That is, $C_{k^*,n}$ is the space-time cluster estimate.

6.4.3 Estimation of $\mu(C_{k,n})$

Typically, the user will not be able to specify the purely spatial $\lambda_S(x,y)$ and the purely temporal $\lambda_T(t)$ functions. Therefore, it is relevant in practice to alleviate him from such requirements. Rather than using the mean $\mu(C_{k,n})$ in (1), we use the data themselves to estimate it. From the non-homogeneous Poisson process properties, under the null hypothesis, for

$k < n$ we have:

$$\mu(C_{k,n}) = \int_{C_{k,n}} \lambda(x,y,t)\,dx\,dy\,dt = \mu \int_{B(\mathbf{s}_k,\rho)} \lambda_S(x,y)\,dx\,dy \int_{(t_k,t_n]} \lambda_T(t)\,dt\,.$$

Therefore, an estimate of $\mu(C_{k,n})$ under the hypothesis that there are no clusters is given by

$$\hat{\mu}(C_{k,n}) = \frac{N(B(\mathbf{s}_k,\rho)\times(0,t_n])\,N(\mathcal{A}\times(t_k,t_n])}{n}\,,$$

where $N(B(\mathbf{s}_k,\rho)\times(0,t_n])$ is the number of events within the disc $B(\mathbf{s}_k,\rho)$ irrespective of occurrence time, $N(\mathcal{A}\times(t_k,t_n])$ is the number of events between times t_k and t_n, irrespective of their spatial location, and n is the total number of events at that moment (see Figure 2).

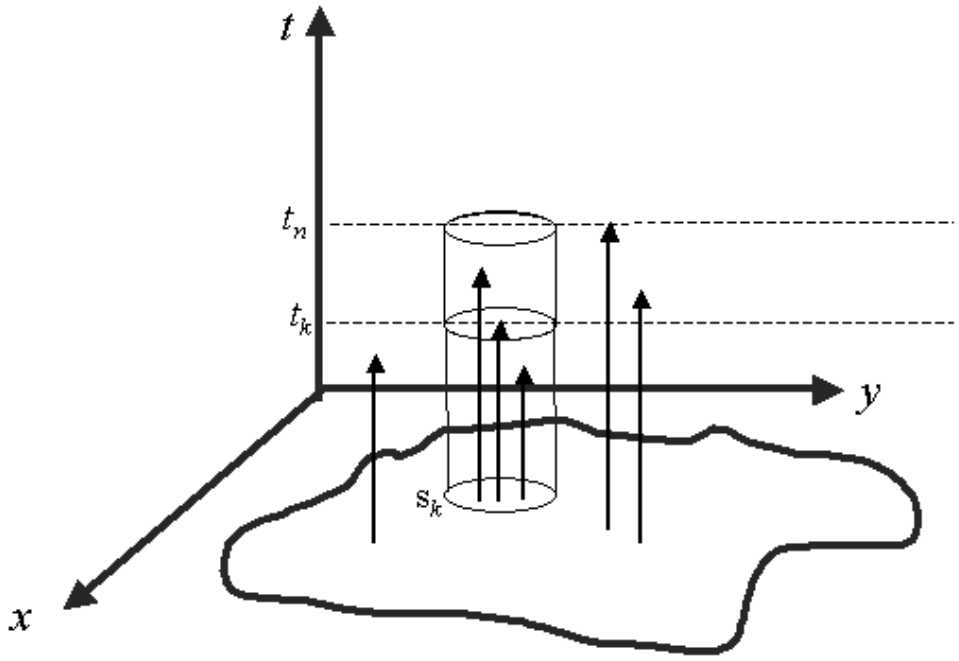

Figura 2: The estimate $\hat{\mu}(C_{k,n})$

6.4.4 Iterative calculation of R_n

Every new event requires the recalculation of all n terms in (2). We can decrease substantially the amount of numerical calculations by means of an iterative procedure. For $k < n$, let $I_{k,n} = I(\|\mathbf{s}_n - \mathbf{s}_k\| \leq \rho)$. We have

$$N(C_{k,n+1}) = N(C_{k,n}) + I_{k,n+1}$$

and

$$\mu(C_{k,n+1}) = \mu(C_{k,n}) + \mu\left(B(\mathbf{s}_k, \rho) \times (t_n, t_{n+1}]\right) .$$

Therefore, since $\Lambda_{1,1} = 1$, we can write recursively for $k < n+1$,

$$\Lambda_{k,n+1} = \Lambda_{k,n}\,(1+\varepsilon)^{I_{k,n+1}} \exp\left(-\varepsilon\mu\left(B(\mathbf{s}_k, \rho) \times (t_n, t_{n+1}]\right)\right) .$$

By definition, $\Lambda_{n+1,n+1} = 1$ and this completes the recursion.

In practice, to run the procedure, we need $\hat{\mu}(C_{k,n+1})$, rather than $\mu(C_{k,n+1})$. However, this estimated term can also be calculated iteratively for $k < n+1$:

$$\begin{aligned}
\hat{\mu}(C_{k,n+1}) &= \frac{1}{n+1} N(B(\mathbf{s}_k, \rho) \times (0, t_{n+1}])\, N(\mathcal{A} \times (t_k, t_{n+1}]) \\
&= \frac{n+1-k}{n+1} \left(N(B(\mathbf{s}_k, \rho) \times (0, t_n]) + I_{k,n+1}\right) \\
&= \frac{n}{n+1} \frac{n-k+1}{n-k}\, \hat{\mu}(C_{k,n}) + \frac{n+1-k}{n+1}\, I_{k,n+1} \\
&= \hat{\mu}(C_{k,n}) + \frac{k}{(n+1)(n-k)} \hat{\mu}(C_{k,n}) + \frac{n+1-k}{n+1}\, I_{k,n+1} .
\end{aligned}$$

For $k < n+1$, we can write

$$\hat{\Lambda}_{k,n+1} = (1+\varepsilon)^{N(C_{k,n})+I_{k,n+1}} \exp\left(-\varepsilon J_{n,k}\right) + \hat{\Lambda}_{n+1,n+1} ,$$

where

$$J_{n,k} = \frac{n(n-k+1)}{(n+1)(n-k)} \hat{\mu}(C_{k,n}) + \frac{n+1-k}{n+1} I_{k,n+1} .$$

Therefore,

$$\begin{aligned} R_{n+1} &= \sum_{k=1}^{n+1} \hat{\Lambda}_{k,n+1} \\ &= 1 + \sum_{k=1}^{n} \hat{\Lambda}_{k,n} \exp\left(\frac{-\varepsilon k}{(n+1)(n-k)} \hat{\mu}(C_{k,n}) \right) L_{\varepsilon,n,k}\,, \end{aligned}$$

where

$$L_{\varepsilon,n,k} = \left((1+\varepsilon) \exp\left(-\varepsilon \, \frac{n+1-k}{n+1} \right) \right)^{I_{k,n+1}}.$$

We let $1 = \Lambda_{n+1,n+1} = \hat{\Lambda}_{n+1,n+1}$. As a consequence, we obtain R_{n+1} simply by updating the values of $\hat{\Lambda}_{k,n}$ with a few numerical calculations.

6.5 Choice of tuning parameters

The variance of the test statistic $R_n = \sum_k \Lambda_{k,n}$ increases with ε when we have no clusters. More importantly, the distribution of R_n is quite asymmetric. Indeed, for one side a large negative deviate of $N(C_{k,n})$ from its mean $\mu(C_{k,n})$ push $\Lambda_{k,n}$ towards its lower bound, equal to zero. For the other side, increasing a large positive deviate drives $\Lambda_{k,n}$ towards infinity. As a consequence, the false alarm rate increases with ε. The simulations in Section 6.6 will show these effects of changing ε on the surveillance system performance. Although analytical results are difficult to obtain, simple approximations provide some insight into this trade-off.

When there is no cluster emerging, we have $E(\Lambda_{k,n}) = 1$ for all ε because

$$\begin{aligned} E(\Lambda_{k,n}|H_0) &= \exp(-\varepsilon\mu(C_{k,n}))\, E\left[(1+\varepsilon)^{N(C_{k,n})}\right] \\ &= \exp(-\varepsilon\mu(C_{k,n}))\, \exp(\mu(C_{k,n})(1+\varepsilon))\, \exp(-\mu(C_{k,n})) \\ &= 1\,. \end{aligned}$$

Then, $E(R_n) = n$ for all ε, as we could expect from the martingale approach of Section 6.3.

Suppose now that a cluster emerges and that $N(C_{k,n}) \sim \text{Poisson}(\mu(C_{k,n})\ (1+\varepsilon^*))$. At this point, we distinguish the true relative change ε^* from the one specified by the method (ε).

They do not need to coincide. In this alternative situation we have

$$\begin{aligned} E(\Lambda_{k,n}) &= \exp(-\varepsilon\,\mu(C_{k,n}))\,E[(1+\varepsilon)^{N(C_{k,n})}] \\ &= \exp(-\varepsilon\,\mu(C_{k,n}))\,\exp[\mu(C_{k,n})(1+\varepsilon^*)(1+\varepsilon-1)] \\ &= \exp[\mu(C_{k,n})\varepsilon\varepsilon^*] > 1\,. \end{aligned}$$

That is, $E(\Lambda_{k,n})$ increases with ε (and with ε^*).

Therefore, apparently, the choice of ε does not affect R_n when there is no cluster. When a cluster emerges, choosing a large ε^* will increase R_n and speed up the threshold crossing, as we wish in this case. Hence, it seems that taking $\varepsilon \to \infty$ is the a good strategy.

However, there is a penalty for choosing ε^* too large in that the $\mathrm{Var}(\Lambda_{k,n})$ increases with ε^* when there is no cluster. In fact,

$$\mathrm{Var}(\Lambda_{k,n}) = \mathrm{Var}[\exp(-\varepsilon^*\,\mu(C_{k,n}))\,(1+\varepsilon^*)^{N(C_{k,n})}]$$

which is equal to

$$\exp(-2\varepsilon^*\,\mu(C_{k,n}))\,\{\exp(\mu(C_{k,n})((1+\varepsilon^*)^2-1)) - \exp(2\mu(C_{k,n})(1+\varepsilon^*-1))\}\,.$$

This reduces to

$$\exp(\mu(C_{k,n})\varepsilon^{*2}) - 1 \to \infty\,,$$

as $\varepsilon^* \to \infty$.

Ultimately, this will increase the false alarm rate. Under no cluster, increasing ε^* too much implies in a larger variability of R_n. This is so because the pairs of terms $\Lambda_{k,n}$ are either uncorrelated (if the corresponding cylinders are non-intersecting) or positively correlated (with correlation proportional of the intersecting volume). As a consequence, the variance of the stopping time T_A increases as well as the probability that R_{T_A} crosses the threshold A at a very early moment, as well as much later than the expected $E(T_A) = A$. Hence, there is a larger probability that the threshold will be crossed before any cluster emergence. These effects will be clear in the simulations of Section 6.6.

As shown in the simulations, changing the radius produces small effects on the surveillance system performance.

6.6 Method performance

We evaluated the performance of our method with Monte Carlo simulation using three types of scenarios. In the first type, there were no clusters and the purpose is to evaluate if the approximation $A = B \approx ARL^0$ is appropriate. The geographical region was the rectangle $[0,10] \times [0,10]$ and the spatial location of an event was obtained by independently and uniformly generating coordinates on the square. Times between events were modelled by independent exponential random variables with mean equal to 1. Times and locations were independently generated and hence $\lambda(x,y,t) = 0.01$. The events were generated sequentially until the alarm was triggered.

In the second type of scenario, in addition to the events generated as in the first scenario, we also simulated events within a cylindrical cluster that emerged at some moment. The purpose of this second type of scenario is to evaluate the detection performance and the effects of the user-specified tuning parameters of our method. The cluster had the square basis $[4,5] \times [4,5]$. It was kept alive from its outbreak until the alarm went off. We selected three different times for the cluster outbreak. In one case, the cluster starts at the beginning of the time period (when $t = 50$) and we label this as *case B*. In the second case, the cluster starts in the middle of the time period, when $t = 150$, and this is labelled *case M*. Finally, in the third case, the cylinder cluster starts late, at $t = 300$, and we labelled this as *case L*.

In each cylinder cluster, we have two types of events: those generated in the larger square and that happened to fall within the cluster, and those events generated within the cluster itself. The intensity at a location (x,y,t) within the cluster is slightly larger than 1.2 times the intensity outside the cluster. This implies that the correct value of the tuning parameter is approximately $\varepsilon = 0.2$.

We will call this second type of scenario as the homogeneous scenario because there is no spatial variation in the events' intensity except that due to the cluster emergence. The third type of scenario was generated in the same way as the second type except by the use of a spatially heterogeneous density of the events locations. In this third scenario type, the spatial coordinates were generated using a mixture of four bivariate normal distributions. Therefore, at any time, some regions were more likely to observe an event than other regions.

We generated 1000 independent replications of each scenario. To run the surveillance procedure, we used several values for ε: $0.1, 0.2, 0.4, 0.5, 2.5, 5, 10, 20$. Note that, for the scenario without clusters, the true value of this parameter is zero, while in the scenarios with clusters, it is equal to 0.2.

No comparison with the scan statistic method from Kulldorff (2001) has been attempted

because it takes too long to run. In each scenario, for a single simulated space time dataset with 1000 events, Kulldorff's method takes 2 hours and 26 minutes in a machine with 1.66GHz and 1 Gb RAM when we ask for 999 Monte Carlo replications. To repeat this procedure for all the simulated datasets over the many scenarios is unfeasible.

6.6.1 Scenario without clusters

We used the values 50, 100, 200, 300, 400, and 500 for the threshold limit A and $\rho = 1$ for the spatial radius of the presumed cluster. If there are no clusters, any surveillance signal is a false signal. If the time period goes to infinity, the statistic R_n will cross the threshold A with probability one, irrespective of the presence of a real cluster. It is expected that, in a fixed time period, the number of events until the alarm goes off increases with the increase of the threshold A. If the approximation $A = B$ is reasonable, we should have $A \approx ARL^0$.

The graph in Figure 3 shows in the vertical axis the estimated ARL^0, the average number of events observed before a false signal is issued, versus the threshold limit A in the horizontal axis. The dashed line represents the line $ARL^0 = A$. The other lines represent the test results for different values of ε.

As we expect, ARL^0 increases with the threshold A. For $\varepsilon = 0.1$ and 0.2, the approximation $ARL^0 = A$ is very good. For other values of ε this approximation deteriorates with the increase of A. Initially, the departure from the line $ARL^0 = A$ is larger the greater is ε. For example, when $A = 500$ and $\varepsilon = 0.5$, the estimated ARL^0 is 33% larger than the nominal value of 500. However, the difference between the nominal value A and the true ARL^0 does not increase monotonically with ε. With $\varepsilon \geq 10$, this difference is almost null. Setting $\varepsilon \geq 10$ is an extreme choice since it means an emerging cluster with intensity 10 times larger than the baseline intensity. It is unlikely that our method is envisioned for such type of anticipated changes.

This result shows that, when no cluster is present and within practical bounds for the expected relative change in intensity, selecting larger values for the user-specified ε parameter leads to conservative ARL^0. That is, if the user is unsure about ε, selecting a larger value leads to longer ARL^0 times than the nominal value A.

However, there is a trade-off in selecting ε too large in that the false alarm rate can increase, as we show in the next section. The reason for this behavior is the increase on the standard deviation of the stopping time T_A with the increase of ε. Table 1 shows this standard deviation for different choices of ε and threshold A. The larger variability will increase the chances of observing much earlier and much later T_A than its expected value. This increases

the chance that an unmotivated alarm sounds off before any cluster emerges. This behavior can be fully understood after we present the results of simulations with clusters.

Threshold	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.4$	$\varepsilon = 0.5$	$\varepsilon = 2.5$	$\varepsilon = 5$	$\varepsilon = 10$	$\varepsilon = 20$
50	0.498	0.518	0.928	1.143	7.115	12.610	17.473	23.009
100	0.561	0.996	1.933	2.483	26.050	39.971	40.410	40.037
200	1.051	2.088	4.746	6.630	95.582	103.240	96.583	97.584
300	1.565	3.291	8.851	13.663	180.976	178.143	142.034	163.956
400	2.137	4.647	15.484	26.544	254.284	233.516	196.761	210.668
500	2.982	6.915	25.488	48.344	294.398	283.244	239.812	254.411

Tabela 1: Standard deviation of the stopping time T_A for different choices of threshold A and ε. In all cases we used $\rho = 1$.

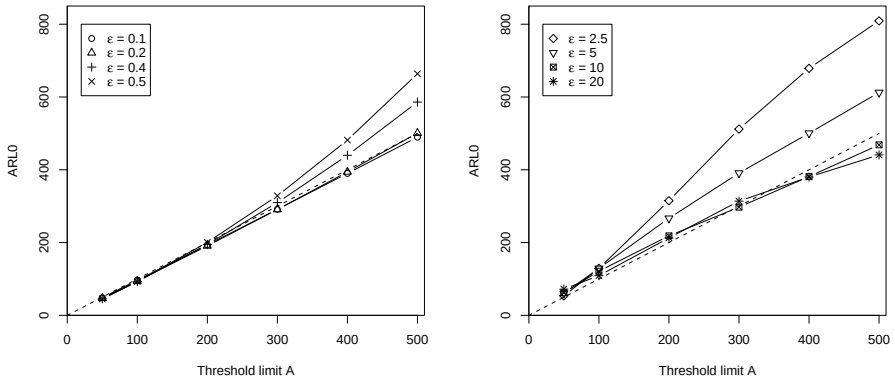

Figura 3: Scenario without cluster. The plots show the estimated $ARL^0 = E(T_A)$, the average number of events observed before a false signal is issued, versus the threshold limit A. Each curve corresponds to a value of ε. In all cases, we used $\rho = 1$.

6.6.2 Scenarios with clusters

One of the usual performance measures of temporal surveillance systems is the CED, the expected delay to trigger a signal after the emergence of a cluster. Assume that the cluster emerges at τ. Given that the stopping time T_A is greater or equal to τ, the usual definition of CED is the expected number of observations one needs to wait until the alarm signal. That is, $CED(\tau) = E[T_A - \tau | T_A \geq \tau]$. One problem with this definition in the space-time situation is that all events between τ and T_A contribute to CED, either they belong to the cluster or not.

We think that a more appropriate CED definition is the average number of events within the space-time cluster until the alarm goes off. We denote this measure by CED^* to distinguish it from the more usual temporal definition.

We tested with four different values for the spatial radius ρ: $0.25, 0.5, 1$, and 2. It is worth remembering that the cluster basis is an unit square. We considered a more refined grid for values of the threshold A than in the without cluster scenario. Namely, we set A equal to $50, 100, 150, 200, 250, 300, 350, 400, 450$, and 500.

The plots in Figure 4 show the estimated $CED^*(\tau)$ against the values of the threshold A. The first row of plots corresponds to the homogeneous scenarios (Hom) while the second row corresponds to the inhomogeneous scenarios (Inh). The different columns correspond to different values of ε. There is one plot for each value of the anticipated relative intensity increase ε. In each plot, we have three lines depending on the value of τ, which can be $\tau = 50$ (case B), $\tau = 150$ (case M), and $\tau = 300$ (case L). It only makes sense to analyze the data for threshold values larger than the emerging cluster time τ. Accordingly, for case B, we do not show $CED^*(\tau)$ when $A = 50$ because almost always the alarm is falsely motivated. The same occurs in case M with threshold $A \leq 150$, and in case L with threshold $A \leq 300$. The estimated $CED^*(\tau)$ in these plots are obtained in simulations where the false alarm percentage is zero.

We analyze initially only the the Middle and End scenarios. In these cases, the plots in Figure 4 lead to the preliminary conclusion that the larger the ε, the better the performance of the surveillance method. Indeed, the $CED^*(\tau)$ decreases with the increase of ε. One should rather set $\varepsilon = 0.5$ than $\varepsilon = 0.2$, which is the true value used in the simulation.

However, increasing ε too much leads to very large false alarm rates. Figure 5 shows the false alarm rates for ε equal to $2.5, 5.0, 10$, and 20 in the homogeneous scenarios. The left hand size plot corresponds to the Middle and the right hand side to the End scenario. Then, it is clear that increasing ε without bounds renders the method useless because the false alarm rate is beyond tolerable standards. This behavior is virtually identical in the inhomogeneous scenarios.

Returning to the more practically oriented values of ε shown in Figure 4, we see that the delay for the alarm to go off is smaller if the cluster is at the end of the observation period. Basically, this means that the alarm system learns what is the intensity for a long time under the no cluster situation. When a cluster finally starts emerging, it is quickly detected. Unless the cluster is at the beginning of the observation period, the $CED^*(\tau)$ curves are almost parallel lines. Hence, the effect of the emerging time τ in the $CED^*(\tau)$ is approximately linear, unless τ is close to zero.

There is a heavy penalty for clusters located at the beginning of the observation period and with ε larger than the true relative change in intensity. In this situation, the expected delay increases very quickly. The alarm system takes a very long time to tell apart what is the baseline intensity from the emerging cluster higher intensity.

Contrasting the homogeneous with the inhomogeneous case, we can see from Figure 4 that, in the inhomogeneous case, the $CED^*(\tau)$ is greater than in the homogeneous case. This effect is especially dramatic if the cluster is located in the beginning of the observation period and ε is larger than the true relative change.

In Figure 6, we show the effects of the radius ρ on the surveillance system performance for the homogeneous case. We ran simulations of the homogeneous scenario with spatial radius $\rho = 0.25, 0.5, 1$, and 2. The inhomogeneous case has virtually identical conclusions and it is not shown. The relative change parameter ε had the values $0.1, 0.2, 0.4$, and 0.5. With respect to the observation period, the clusters started early ($\tau = 50$), at the middle ($\tau = 150$), or late ($\tau = 300$). Most often, the procedure is insensitive to the choice of the radius ρ. Except for the cluster at the beginning and large ε, there is very little difference on the estimated $CED^*(\tau)$. The exceptional behavior occurs only in a special conjunction of factors: when the parameter ρ is larger than the true cluster radius, when the clusters starts emerging early, and when ε is much larger than its true value.

6.7 Illustrative examples

We illustrate our method using two real datasets: the Burkitt's lymphoma cases in Uganda, the same dataset used in Rogerson (2001), and the *Meningitis* cases occurred in Belo Horizonte, Brazil, between 2001 and 2005.

6.7.1 Burkitt's lymphoma cases in Uganda

A classical example of retrospective detection of space-time clustering is that based on the Burkitt's lymphoma in Uganda (Williams et al., 1978). The data consist of the place of residence and onset time for all 188 cases of Burkitt's lymphoma between 1961 and 1975 in the West Nile district in Uganda (see Figure 7). Rogerson (2001) found evidence of space-time clusters using local Knox tests and adopting a probability of false alarm of 0.1. Notwithstanding the problems found in Rogerson's method by Marshall et al. (2007), we compare our results with his.

The tuning parameters in our surveillance method were: $\varepsilon = 0.1, 0.2, 0.4$, and 0.5; $\rho =$

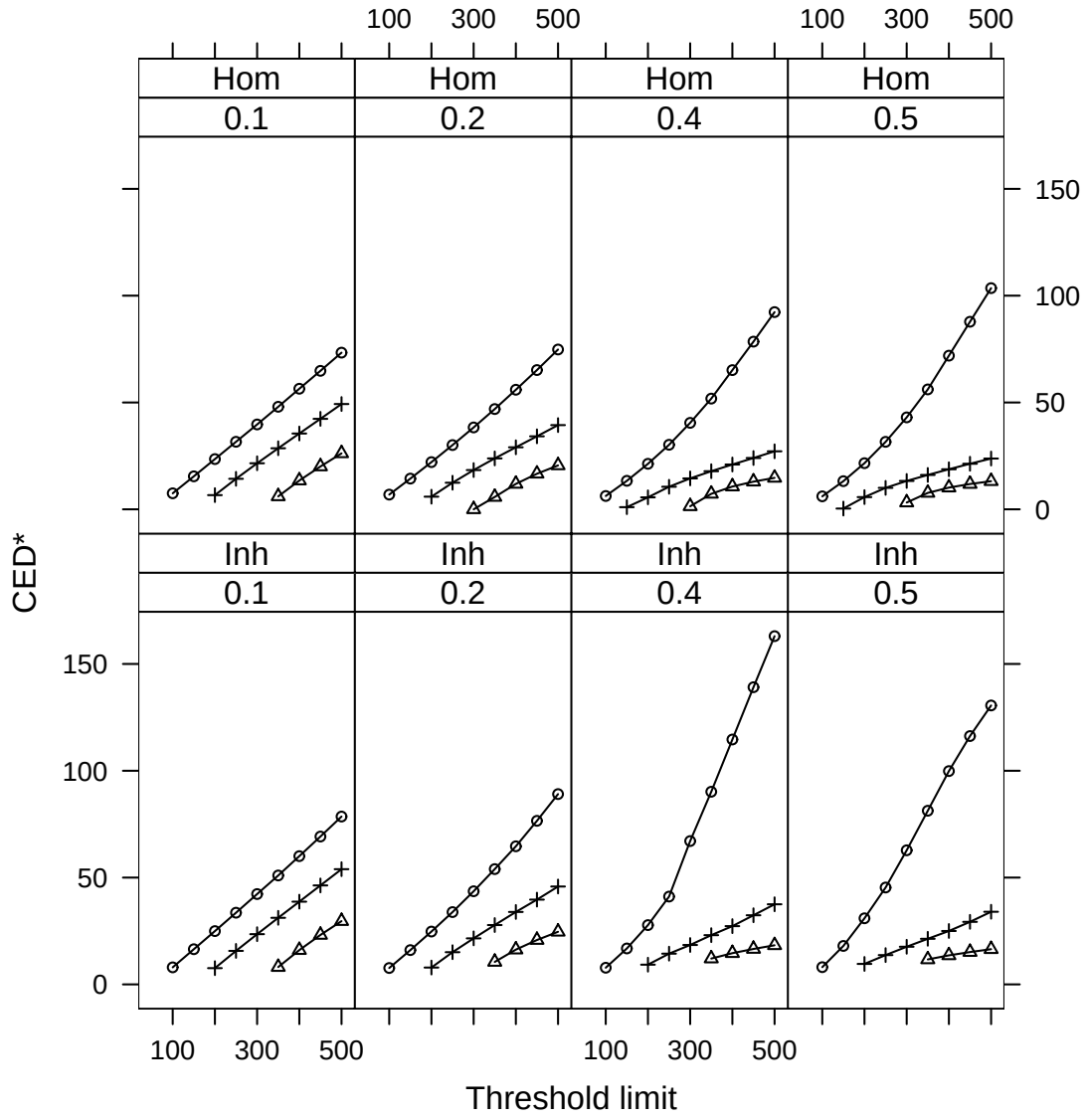

Figura 4: Estimated $CED^*(\tau)$ against the values of the threshold A. The first row of plots corresponds to the homogeneous scenarios (Hom) while the second row corresponds to the inhomogeneous scenarios (Inh). The different columns correspond to different values of ε. Each plot has three lines. The circles correspond to case B, when the cluster emerges soon in the observation period ($\tau = 50$). The crosses correspond to case M ($\tau = 150$), and the triangles correspond to case L ($\tau = 300$).

$2.5, 5, 10$, and 20 km; alarm threshold $A = 161$. Hence, in average, we expect 161 events before the alarm goes off falsely. In all cases, the alarm went off and the triggering event varied from 142 to 158. The difference is $158 - 142 = 16$, which corresponds to 8.5% of

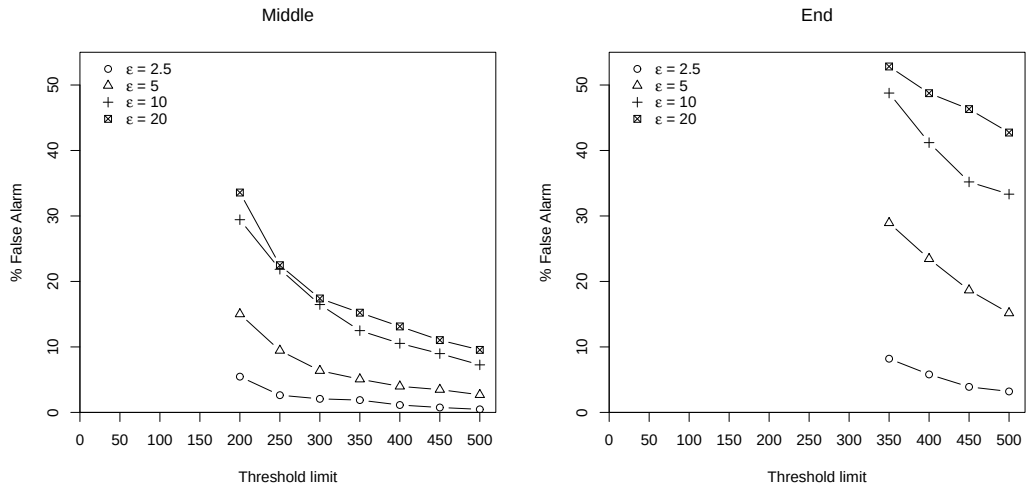

Figura 5: Left hand side: False alarm rates versus threshold limit *A* for the scenarios when the cluster emerges on the middle of the observation period with $\varepsilon = 2.5, 5, 10, 20$. Right hand side: Idem for cluster emerging at the end of the observation period.

the total number of observations. With respect to the emerging time τ, the estimates varied from 103 to 138. The largest number corresponds to the smallest radius (2.5). Except for this rather extreme radius, all the estimated starting times τ varied from 103 to 107, a very short range. More relevant than this, the estimated clusters are located approximately in the same region in all cases. Hence, the results are relatively insensitive to the tuning parameters values.

Figure 7 shows the graphs of R_n versus n for ρ equal to 10 and 20 km. For $\rho = 20$ km and $\varepsilon = 0.5$, the alarm goes off at event number 148 (February, 1973) and the method estimates that the clusters started on event 107 (November, 1970). This cluster is represented in the map on the right hand of Figure 7. Note that, among the 40 events occurring all over the map between November, 1970 and February, 1973, only 20 belong to the cluster.

It is notable that the cases in the emerging cluster shown in Figure 7 coincide with one of the clusters identified by Williams *et al.* (1978) through the pairs of cases whose disease onset was in the period 1972-1973, within 10 km apart and 180 days of each other.

Rogerson (2001) ran his procedure with several different tuning parameters and his results are sensitive to these choices. Many of his detected clusters coincide with ours. However, one should keep in mind the criticisms to Rogerson procedure presented in Marshall et al. (2007).

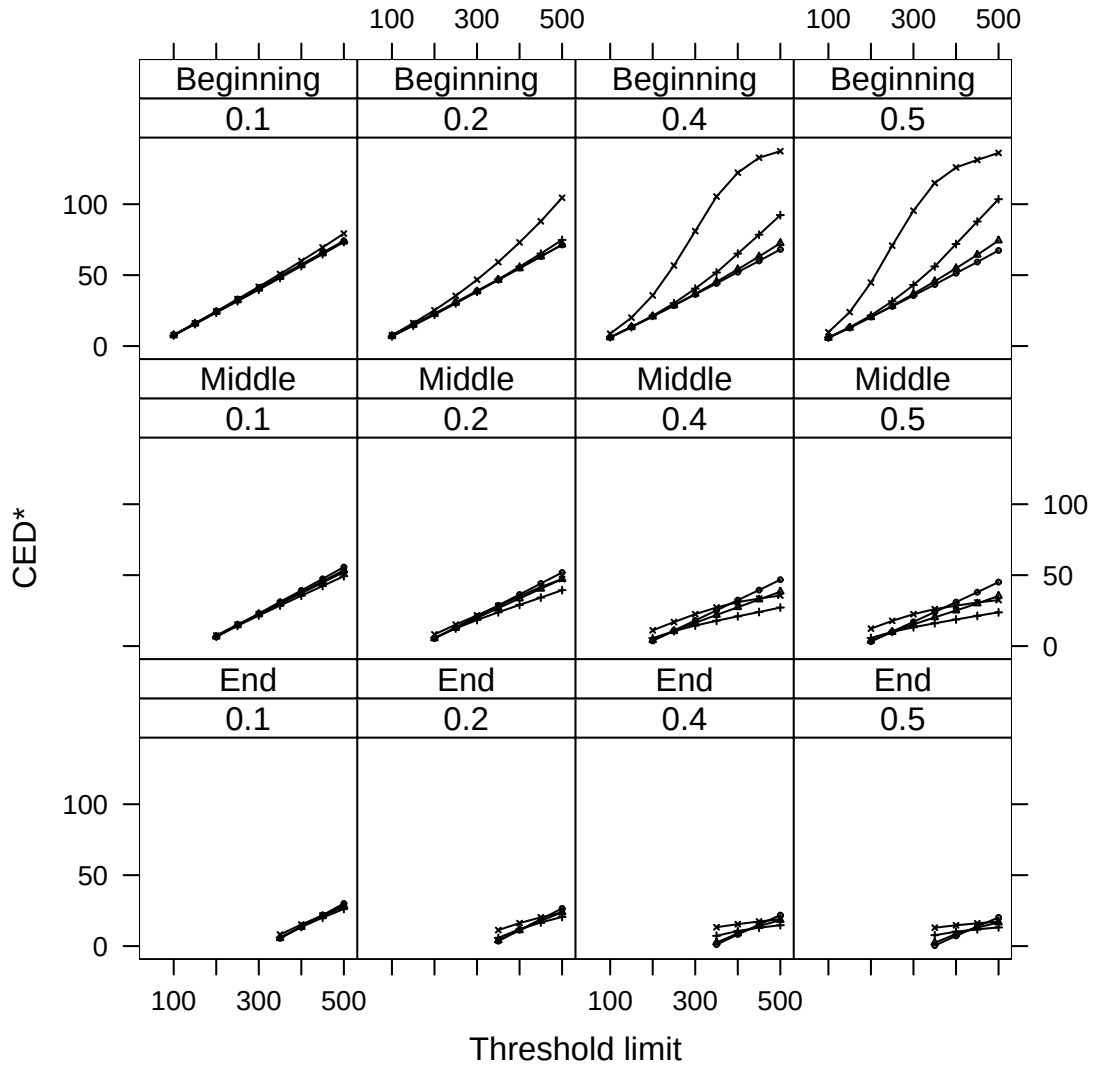

Figura 6: Effect of changing ρ. The rows correspond to the three cluster emerging time $\tau = 50, 150, 300$, and the columns correspond to different values of ε. Only the homogeneous case was considered. The curves with circles correspond to $\rho = 0.25$, the curves with triangle to $\rho = 0.5$, the curves with crosses to $\rho = 1.0$, and the curves with axes to $\rho = 2.0$. The true value of ρ is 0.5.

We ran the space-time prospective scan statistic method proposed by Kulldorff (2001) and implemented in the software SaTScan (Kulldorff, 2003). SaTScan is a freely available software and it was developed under the joint auspices of Martin Kulldorff, the National Cancer Institute, and Farzad Mostashari of the New York City Department of Health and Mental Hygiene.

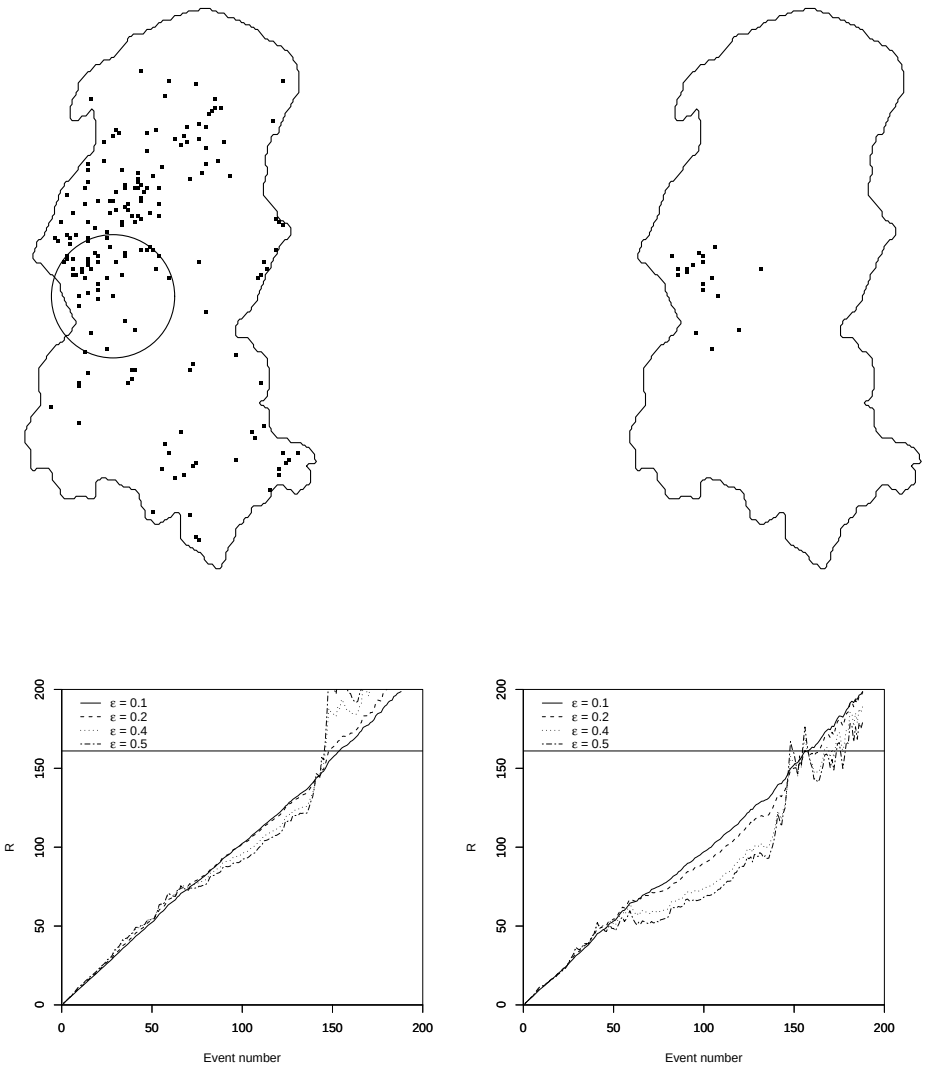

Figura 7: Burkitt's lymphoma cases in West Nile district of Uganda from 1961 to 1975 (study region is approximately 80 km $\times$ 170 km). The left hand side map shows all the events in the period while the right hand side map shows the events identified in the emerging cluster by our method. Each one of the plots shows R_n versus n for four different choices of ε: $0.1, 0.2, 0.4$, and 0.5. The left hand side plot uses $\rho = 10$ km and the right hand side plot uses $\rho = 20$ km.

There is a difficulty to compare our method with that of Kulldorff (2001). His method, as implemented in the software, receives a dataset and scans for an alive cluster. That is, it searches for a cylinder-shaped cluster whose height ends at the last available observation time. Running his method with all 188 events will not allow for clusters that could have been found before the last observation. To avoid running his method manually repeatedly

by adding a single observation each time, we ran it initially with all 188 observations. The scan statistic finds a non-significant cluster centered at the spatial coordinates $(273, 332)$, with a radius equal to 13.42 km, and ranging from 9/2/1972 to 10/24/1975, when the scan alarm sounded off. Based on 999 simulations, the Monte Carlo p-value associated with this space-time cluster is equal to 0.129.

We then used this same 13.42 km radius found by the scan statistic in our own method. The alarm sounded off four times, in 6/15/1973, 4/24/1973, and 2/1/1973, for ε equal to 0.1, 0.2, 0.4, and 0.5, respectively. The clusters found were all centered at the spatial coordinates $(273, 332)$, the same non-significant cluster found by the scan statistic based on all data points. Our method estimated the cluster emergence at 31/3/1971.

Finally, we ran Kulldorff's method again but using only the events that happened until the dates our method sounded off. Using the events until the endpoint 6/15/1973 or until the endpoint 4/24/1973, we have essentially the same results as our own method. A 5% significant cluster is found with radius equal to 11.40 km centered at $(273, 332)$, and ranging from 2/9/1972 until the corresponding endpoint. Running Kulldorff's procedure with the events until 2/1/1973, we find a 5% significant cluster with radius equal to 5.0 km centered at the spatial coordinates $(263, 333)$, and emerging at 9/2/1972, the same date as the previous cases. Therefore, in this example, the two methods give similar results.

6.7.2 Meningitis cases in Belo Horizonte

We apply our method using data with place and onset date for 1001 Meningitis cases that occurred between 2001 and 2005 in Belo Horizonte, Brazil (see Figure 8). In 60% of the days no cases were recorded and, in 72% of the remaining days, only one case was recorded. The maximum number of cases in a single day was equal to 6. From the time series plots in Figure 8, no discernible trend or periodicity is present.

The tested the following parameters: $\varepsilon = 0.1, 0.2, 0.4$, and 0.5; $\rho = 1, 2, 3$, and 4 km; alarm threshold $A = 500$. Threshold $A = 500$ means that we expect 500 cases before the alarm goes off without need. Since we have around 200 cases per year, we are expecting two unmotivated alarms in a period of 5 years.

The alarm went off in all situations, irrespective of the parameters values. The triggering event varied from 490 to 949 and it increased if either of the tuning parameters, ε and ρ, increased. However, there was a positive interaction between these two tuning parameters: the waiting time increased faster with ρ if ε was larger.

Figure 8 shows the R_n statistics (2) versus n for $\varepsilon = 0.1, 0.2, 0.4, 0.5$ and $\rho = 1.0, 2.0$ km.

These plots illustrate the effects of changing ρ and ε.

City health officials were not suspicious of any emerging cluster during this period before our analysis. Since the triggering event occur around the expected time under the no cluster situation, our method supports this opinion. There is no convincing evidence for the emergence of meningitis clusters in Belo Horizonte.

This lack of evidence is reinforced by Kulldorff's prospective scan statistic method. Running his method with all the events, we did not find a significant cluster. Its most likely cluster, with p-value equal to 0.129, had radius zero, including only two events with identical spatial coordinates.

6.8 Conclusions

Compared to the main space-time surveillance methods in the literature, our method has some advantages and disadvantages. Kulldorff (2001) does not requires tuning parameters and he does not use the concepts of average run length and conditional expected delay preferring the hypothesis testing concepts of error type I and power. One problem with his method is that it does not control the error type I over repeated and periodic surveillance. It adjusts for all previous analysis in each time moment the scan is performed and a correct α error type I probability is achieved at that moment. However, considering the simultaneous inference for all points in time, the correct level is not α. This is clearly seen if one considers the situation in which no cluster ever emerges. Running a α level test at each moment in sequence, even when controlling for the past analysis at each moment, we are bound to provide a significant result eventually. Therefore, the true significance level would be 1 for the infinite sequence of tests.

We avoid this problem by adopting the quality control ideas of average run length but our method requires the specification of tuning parameters. Although we think that users should be able to propose reasonable values for these parameters, one needs more studies to understand fully the impact of them in practice. This need to set tuning parameters is also a requirement in Rogerson (2001) but, as Marshall et al. (2007) shows, his method has several shortcomings.

Our surveillance method assumes a spatially circular shaped cluster. A completely arbitrary shaped cluster is unfeasible computationally and circular shaped clusters provide a good trade-off between computational cost and meaningful and practical solutions. However, there are situations when a truly irregularly shaped cluster could be of concern such as a narrow

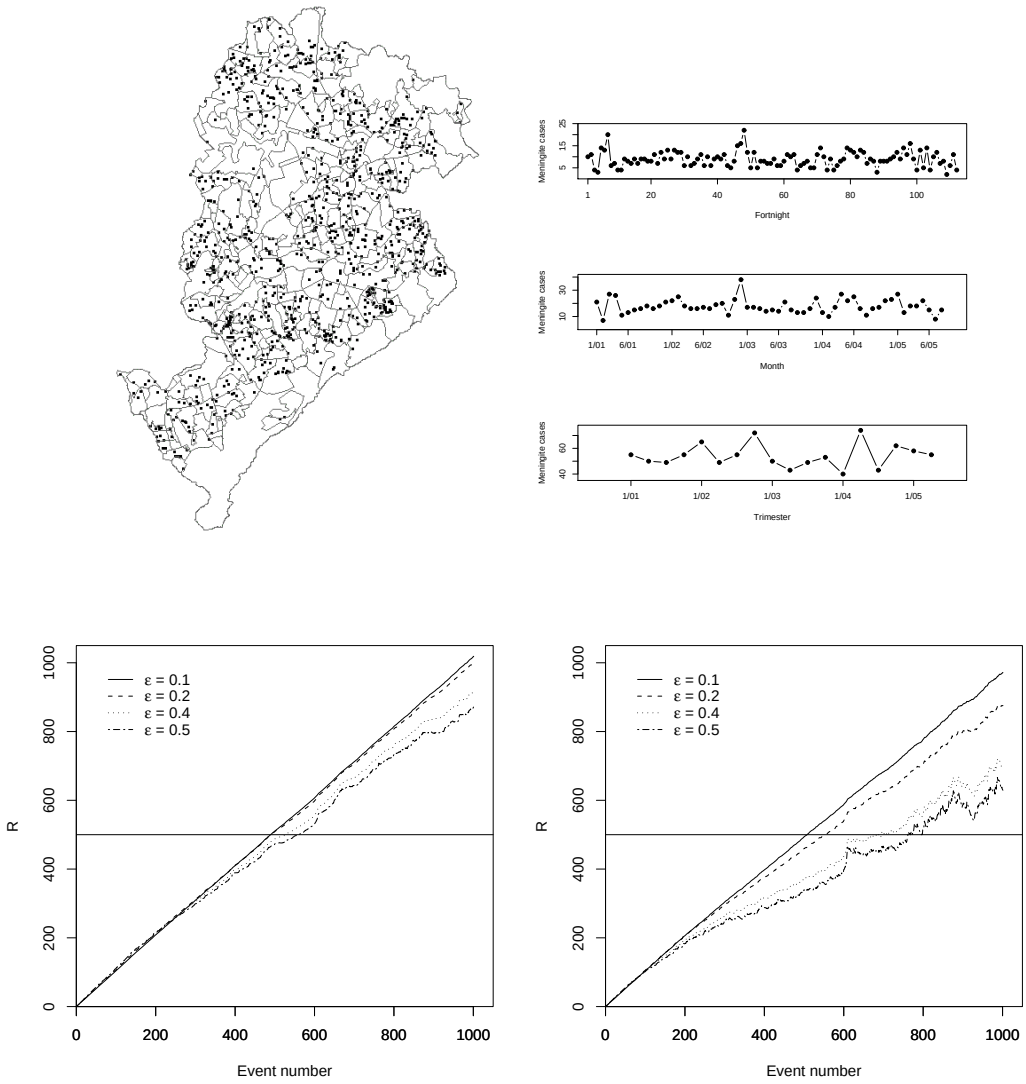

Figura 8: Map of Belo Horizonte divided into neighborhoods and the location of 1001 Meningitis cases that occurred between 2001 and 2005. The study region has approximately 31 km $\times$ 16 km. The time series plots show the number of events in different time units: fortnightly, monthly, and quarterly. Each one of the plots shows R_n versus n for four different choices of ε: $0.1, 0.2, 0.4$, and 0.5. The threshold is $A = 500$. The left hand side plot uses $\rho = 1$ km and the right hand side plot uses $\rho = 2$ km.

zone along a river or an avenue. To solve this problem in the purely spatial cluster detection context, several authors have proposed spatial scan statistics using an irregular shaped scanning window (Duczmal and Assunção, 2004, Patil and Taillie, 2003, Duczmal et al., 2006, Tango and K., 2005, Assunção et al., 2006). Such scanning windows could also be adopted

for our space-time statistic with some cost in additional computing time.

Our method has many desirable features. First, it does not require information about the population at risk, only cases are necessary. Second, it adjusts for purely spatial and purely temporal clustering, and it provides statistical inference for the emerging cluster detected. Third, it does not require many input parameters and the ones it does have a clear practical interpretation. This interpretation should help the user to establish reasonable values for them. We think it will be of great use in many practical applications.

Our method has been implemented in a stand-alone C++ software as well as in TERRAVIEW, a free GIS software based on the open-source TERRALIB library. A suite of R functions has also been developed and they are available upon request.

7 Vigilância espaço-tempo prospectiva para dados de área

Esta seção traz o artigo “Prospective space-time surveillance for areal data”, que será submetido ao periódico *Statistics in Medicine*.

7.1 Abstract

Nowadays, government agencies, specially public health departments, are monitoring disease cases looking for the early detection of space-time high intensity clusters. We propose a method for prospective space-time disease surveillance using counting data at fixed spatial areas with discrete time. The method sounds off an alarm when there is empirical evidence that a spatially localized cluster starts to emerge, keeping the rate of false alarms at a desired level. The number of disease cases is assumed to be Poisson distributed. The method adjusts for purely spatial and temporal variability over the map and it looks for cylindrical clusters with circular base. It does not depend on the knowledge of the population at risk. When the monitoring statistic, based on the Shiryayev-Roberts statistic, exceeds a threshold, the alarm sounds off. A martingale structure is used to derive this threshold. We use Monte Carlo simulation to study the impact of the input parameters in the performance of the proposed method. An example of meningococal cases in Germany is present to illustrate the application of the method.

7.2 Introduction

In Assunção and Correa (2009), we proposed a method for the detection of space-time emergent clusters using data from a point process at continuous time and space. This method is available in the R package Surveillance (Höhle, 2007). However, the most common is to record the number of cases at fixed spatial areas (eg. neighborhood, city) and time intervals (eg. week, month, year). The reasons for that may be linked to confidentiality concerns. In this case, there is one time series of counts for each spatial area.

Consider the problem of purely temporal prospective surveillance using a single time series of counts. The Farrington method (Farrington et al., 1996), the Shewhart control chart for counts and the Poisson cumulative sum (see Wetherill and Brown, 1991, Montgomery, 1996 and Ryan, 1989 for Shewhart control charts and cumulative sum methods) can be used to deal with this problem. In this prospective case, at each time a new count is available, the surveillance method should test if the series remains in the same baseline state. The test is

based on the series values up to the current time. Therefore, the tests are conducted sequentially in time. The Farrington method did not take this inherent multiple testing structure of the problem into account, since it is based on the repeated use of confidence intervals over time. The Shewhart control charts have this same difficulty. The cumulative sum methods overcomes this difficulty by controlling the expected waiting time for a false signal, instead of adopting the error type I probability used in the statistical hypothesis tests framework. Of course it is possible to evaluate the expected waiting time for a false signal in the Shewhart control charts. However, the control limits used in these charts are calculated so that, at each time, the error type I probability is equal to a fixed level α. Correa et al. (2011) discuss the problem of using the error type I probability in the prospective surveillance context.

Recently, as geographic information is collected more easily, many methods have been proposed for on line space-time surveillance using one time series of counts for each spatial area. The aim is to sound off an alarm as soon as a spatially localized cluster starts to emerge. The rate of false alarms should be kept at a required level. Kulldorff (2001) for example, suggested to use the space-time scan statistic for this purpose. Takahashi et al. (2008) proposed the use of a flexibly shaped space-time scan statistic. Unlike the space-time scan statistic suggested by Kulldorff (2001), which considers a cylindrical window with circular base, the flexible scan statistic consider a prismatic window whose base has an arbitrary shape. Tango et al. (2011) argue that the space-time scan statistic proposed by Kulldorff (2001) compares the observed number of cases with the conditional expected number of cases. They suggest the use of a new space-time scan statistic that compares the observed number of cases with the unconditional expected number of cases. Correa et al. (2011) highlighted the problems of using a method based on a space-time scan statistic for prospective surveillance.

Neil and Cooper (2010) proposed the multivariate Bayesian scan statistic (MBSS) for event detection and characterization in multivariate spatial time series of counts. MBSS integrates prior information and observations from multiple data streams, calculating the posterior probability of each type of event in each space-time region. Corberán-Vallet and Lawson (2011) apply the conditional predictive ordinate to the surveillance context to detect small areas of increased disease incidence. They incorporate a common probability that each small area signals an alarm when there is no change in the risk pattern of disease to address the problem of multiple comparisons.

Höhle et al. (2009) illustrate the use and the potential of prospective and retrospective statistical surveillance in veterinary epidemiology. Höhle (2009) proposed an extension of the

stochastic susceptible-infectious-recovered (SIR) model to support a regression framework for modeling infectious disease data. A multivariate counting process specified by conditional intensities is the base of the proposal. The conditional intensities contain an additive epidemic component and a multiplicative endemic component. Diggle et al. (2005) consider a point process model in which the spatial-temporal intensity has three multiplicative components. The two components describing the purely spatial and the purely temporal variation in the normal disease incidence pattern are both deterministic. The third component represents spatially and temporally localized departures from the normal pattern and it is an unobserved stochastic component.

In this paper we propose a method for space-time surveillance using counting data at pre-defined spatial areas with discrete time. Our method does not depend on the knowledge of the population at risk and it adjusts for purely spatial and temporal variability over the map. The number of cases at area i and time t is assumed to be Poisson distributed. We assume that somewhere, at some moment, one or more space-time high intensity clusters start to emerge. The Poisson mean is greater inside the cluster. We consider cylindrical clusters with circular base. The monitoring statistic is based on the Shiryayev-Roberts statistic (Assunção and Correa, 2009). The alarm goes off when the monitoring statistic exceeds a threshold. We adopted the martingale structure of the Shiryayev-Roberts statistic to derive the value of the threshold.

In Section 7.3, we present our proposal. In Section 7.4 we analyze the impact of the input parameters in the method performance using a simulation study. We also analyze the impact of the true spatial size of the cluster. Section 7.5 illustrates the method using a real dataset with meningococal cases in Germany. We close in Section 7.6 with a summary of the main conclusions and final considerations.

7.3 Statistical formulation

7.3.1 Monitoring statistic

Consider a map divided into H spatial areas. The map is observed at discrete times $t = 1, 2, \ldots$. Let Y_{it} be the number of disease cases in area i at time t. Let P_{it} be the population at risk in area i at time t. Assume λ_{it}, $0 < \lambda_{it} < 1$ for all i, t, to be the per capita rate of occurrence of cases. The number of cases Y_{it} are independent random variables such that

$$Y_{it} \sim \begin{cases} \text{Poisson}(\mu_{it}), & \text{if} \quad i \notin S_j \text{ or } t < \tau \\ \text{Poisson}(\mu_{it}(1+\varepsilon)), & \text{if} \quad i \in S_j \text{ and } t \geq \tau \end{cases}$$

where $\mu_{it} = \lambda_{it} P_{it}$, $\varepsilon > 0$ is a constant, τ is an unknown change time, and S_j is a spatial circle. The intensity λ_{it} may be a function of known covariates such as social characteristics of the areas and seasonal effects, as in Höhle et al. (2009):

$$\log(\lambda_{it}) = \alpha_i + \beta t + \sum_{s=1}^{S} \left[\gamma_s \sin\left(\frac{2\pi}{r} s t\right) + \delta_s \cos\left(\frac{2\pi}{r} s t\right) \right],$$

where r is a known period (e.g. 12 for monthly data). We assume that the parameters in λ_{it} have been previously estimated with training data and are known. For simplicity, we will take λ_{it} non-parametrically, not imposing any functional form on them. The center of S_j coincides with the centroid of a certain area j and the radius of S_j is equal to ρ. The process is under control at time t if $t < \tau$ and the process is out of control at time t if $t \geq \tau$. We let $\tau = \infty$ denote an under control process.

Our method sounds off an alarm at time m when there is empirical evidence that the process is out of control at time m. At each new time $(m+1)$ in which the map of counts is observed, the method should decide if the process is under or out of control at $(m+1)$. When the alarm sounds off, the method also provides an estimate for the spatial location of the emerging cluster.

Let $L_{\infty,m}$ be the likelihood function for m observed maps when $\tau = \infty$:

$$L_{\infty,m} = \prod_{1 \leq i \leq H, 1 \leq t \leq m} \frac{(\mu_{it})^{y_{it}}}{y_{it}!} \exp(-\mu_{it}).$$

Let $C_{j,k}$ be a cylinder with basis S_j that begins at time $t = k$ and reaches the current time $t = m$. We consider only those cylinders that reach the current time m, since we are interested in detecting live clusters. That is, we are interested in detecting clusters that are still in activity at the actual time m.

Let $L_{j,k,m}$ be the likelihood function for m observed maps when $\tau = k$ and a cluster around area j starts emerging. We have:

$$L_{j,k,m} = \prod_{1\leq i\leq H, 1\leq t\leq m} \frac{(\mu_{it}(1+\varepsilon I_{it}))^{y_{it}}}{y_{it}!} \exp(-\mu_{it}(1+\varepsilon I_{it}))$$

where

$$I_{it} = \begin{cases} 1, & \text{if} \quad i \in S_j \text{ and } t \geq k \\ 0, & \text{otherwise} \end{cases}$$

The likelihood ratio is given by

$$\begin{aligned} \Lambda_{j,k,m} &= \frac{L_{j,k,m}}{L_{\infty,m}} \\ &= \prod_{i\in S_j, k\leq t\leq m} (1+\varepsilon)^{y_{it}} \exp(-\varepsilon\mu_{it}) \\ &= (1+\varepsilon)^{\sum_{i\in S_j}\sum_{k\leq t\leq m} y_{it}} \exp\left(-\varepsilon \sum_{i\in S_j}\sum_{k\leq t\leq m} \mu_{it}\right) \\ &= (1+\varepsilon)^{y_{C_{j,k}}} \exp\left(-\varepsilon\mu_{C_{j,k}}\right), \end{aligned}$$

where $\sum_{i\in S_j}\sum_{k\leq t\leq m} y_{it} = y_{C_{j,k}}$ and $\sum_{i\in S_j}\sum_{k\leq t\leq m} \mu_{it} = \mu_{C_{j,k}}$.

One possibility is to define $\max_{j,k}(\Lambda_{j,k,m})$ as the monitoring statistic. This would lead to difficulties since we do not know the distribution of this statistic. Instead of this, we follow the idea of Kenett and Pollak (1996) and define the monitoring statistic R_m as the sum of the likelihood ratio $\Lambda_{j,k,m}$ for all possible change time k and all spatial location j. That is,

$$R_m = \sum_{k=1}^{m}\sum_{j=1}^{H} (1+\varepsilon)^{y_{C_{j,k}}} \exp\left(-\varepsilon\mu_{C_{j,k}}\right) y_{jk}\,.$$

The monitoring statistic has a factor y_{jk} multiplying the likelihood ratio $\Lambda_{j,k,m}$. There are two reasons for this. The first one is related to the possibilities for the spatial location j of the emerging cluster. The monitoring statistic considers only those areas that have at least one case at time k. There is no sense in considering an area that has no cases at time k as the center of a cluster that begins exactly at time k. The second reason is related to the

possibilities for the change time τ. Suppose a cylinder shaped cluster centered at area j and that starts emerging at time k. The y_{jk} cases observed in time k at area j did not occurred exactly at the same time. Then, based on these y_{jk} cases, there are y_{jk} possibilities for the change time τ, since the time of occurrence of each one of these y_{jk} cases is a possible value for τ.

Our surveillance method calculates R_m as the m-th map is observed, substituting the unknown $\mu_{C_{k,n}}$ by an estimate $\widehat{\mu}_{C_{k,n}}$. In Section 7.3.4 we discuss the estimation of $\mu_{C_{k,n}}$ and provide a non-parametric estimate that does not require the knowledge of the population sizes, only the number of cases in each area. The alarm goes off when $R_m \geq A$ for the first time, where A is a threshold specified by the user. The specification of A is discussed in Section 7.3.3. The user also has to specify values for ε and ρ.

Once we have an alarm at time m, we need to provide an estimate for the cluster spatial location. This estimate is given by the area with the greatest contribution to the monitoring statistic at the moment the alarm was triggered. That is, the cluster is centered at the j-th area that maximizes

$$\sum_{k=1}^{m} (1+\varepsilon)^{y_{C_{j,k}}} \exp\left(-\varepsilon \mu_{C_{j,k}}\right) y_{jk}\,.$$

It is unlikely, but possible, that two or more areas have the same contribution. If this is the case, the method gives more than one cylinder as an estimate for the cluster. These cylinders can have some intersection or not. It is also possible to detect the area with the second greatest contribution, the third greatest contribution, and so on.

7.3.2 Expected value for the monitoring statistic

We can obtain the expected value of the monitoring statistic R_m when $\tau = \infty$:

$$\begin{aligned}
E_{\tau=\infty}\{(1+\varepsilon)^{Y_{C_{j,k}}} \exp\left(-\varepsilon \mu_{C_{j,k}}\right) Y_{jk}\} &= E_{\tau=\infty}\{(1+\varepsilon)^{Y_{C'_{j,k}}} \exp(-\varepsilon(\mu_{C_{j,k}} - \mu_{jk})) \\
&\quad (1+\varepsilon)^{Y_{jk}} \exp(-\varepsilon \mu_{jk}) Y_{jk}\} \\
&= E_{\tau=\infty}\{(1+\varepsilon)^{Y_{C'_{j,k}}} \exp(-\varepsilon(\mu_{C_{j,k}} - \mu_{jk}))\} \\
&\quad E_{\tau=\infty}\{(1+\varepsilon)^{Y_{jk}} \exp(-\varepsilon \mu_{jk}) Y_{jk}\} \\
&= 1 \times (1+\varepsilon)\mu_{jk} \\
&= (1+\varepsilon)\mu_{jk}\,,
\end{aligned}$$

where $C'_{j,k}$ is the three-dimensional region formed by the cylinder $C_{j,k}$ excluding the area j at time k.

Therefore

$$E_{\tau=\infty}(R_m) = \sum_{k=1}^{m}\sum_{j=1}^{H} \mu_{jk}(1+\varepsilon) = (1+\varepsilon)\sum_{k=1}^{m}\sum_{j=1}^{H} \mu_{jk} = (1+\varepsilon)\mu_m\,, \tag{3}$$

where μ_m is the expected number of cases, under $\tau = \infty$, in the m observed maps.

For the out of control situation, when $\tau < \infty$, we have:

$$\begin{aligned} E_\tau\{(1+\varepsilon)^{Y_{C_{j,k}}} \exp\left(-\varepsilon\,\mu_{C_{j,k}}\right) Y_{jk}\} &= \exp(\mu_{C'_{j,k}}\varepsilon\varepsilon^*)\exp(\mu_{jk}\varepsilon\varepsilon^*)\mu_{jk}(1+\varepsilon)(1+\varepsilon^*) \\ &= \exp(\varepsilon\varepsilon^*(\mu_{C'_{j,k}}+\mu_{jk}))\mu_{jk}(1+\varepsilon)(1+\varepsilon^*) \\ &= \exp(\varepsilon\varepsilon^*\mu_{C_{j,k}})\mu_{jk}(1+\varepsilon)(1+\varepsilon^*)\,, \end{aligned}$$

where ε^* is the true value of the relative change in the expected value. Therefore,

$$\begin{aligned} E_\tau(R_m) &= \sum_{k=1}^{m}\sum_{j=1}^{H} \exp(\varepsilon\varepsilon^*\mu_{C_{j,k}})\mu_{jk}(1+\varepsilon)(1+\varepsilon^*) \\ &= (1+\varepsilon^*)(1+\varepsilon)\sum_{k=1}^{m}\sum_{j=1}^{H} \exp(\varepsilon\varepsilon^*\mu_{C_{j,k}})\mu_{jk} \\ &> (1+\varepsilon)\mu_m\,. \end{aligned}$$

The monitoring statistic R_m increases faster under $\tau < \infty$ than under $\tau = \infty$ by a global multiplicative factor $(1+\varepsilon^*)$. It is also larger due to local factors equal to $\exp(\varepsilon\varepsilon^*\mu_{C_{j,k}})$.

7.3.3 Specification of the threshold A

Under $\tau = \infty$,

$$\Lambda_{j,k,m} = \frac{L_{j,k,m}}{L_{\infty,m}} = \prod_{i\in S_j}\prod_{t=k}^{m}(1+\varepsilon)^{y_{it}}\exp(-\varepsilon\,\mu_{it})$$

is a martingale with respect to L_∞ (see the Appendix for details).

Let M be a random stopping time with respect to $[X(t) = (y_{1t},\ldots,y_{Ht}), t > 0]$. That is, M is a random time such that, for each t, the occurrence or non-occurrence of the event $\{M = t\}$

depends only on the values of $\{X(0), \ldots, X(t)\}$. Suppose

$$E_{\tau=\infty}(M) \geq B\,. \tag{4}$$

The stopping time M is discrete, since it must be equal to one of the discrete times the map is observed. The expectation $E_{\tau=\infty}(M)$ is the expected waiting time for a false alarm, called the Average Run Length (*ARL*). The constant B reflects the user acceptable *ARL*. We declare that we have an alarm at the first time M_A that R_m crosses a threshold A:

$$M_A = \min[m | R_m \geq A]\,.$$

Under $\tau = \infty$, R_m is a martingale (since R_m is a sum of martingales) with expectation equal to $(1+\varepsilon)\mu_m$. Therefore, $R_m - (1+\varepsilon)\mu_m$ is a martingale with zero expectation. By the Optional Sampling Theorem, $E_{\tau=\infty}(R_{M_A} - (1+\varepsilon)\mu_{M_A}) = 0$. Hence, $E_{\tau=\infty}(R_{M_A}) = E_{\tau=\infty}[(1+\varepsilon)\mu_{M_A}]$. By definition, $R_{M_A} \geq A$. Then, $E_{\tau=\infty}(R_{M_A}) = E_{\tau=\infty}[(1+\varepsilon)\mu_{M_A}] = (1+\varepsilon)\,\mu\, E_{\tau=\infty}(M_A)$, where μ is the expected number of cases in a single map. Since M_A is the first time R_m exceeds A, we assume that the excess is typically not great. Therefore, setting $A = B(1+\varepsilon)\,\mu$, where B is the desired *ARL*, yields a procedure that satisfies (4).

7.3.4 Estimation of $\mu_{C_{k,n}}$

In practice, $\mu_{C_{j,k}}$ is unknown. Under $\tau = \infty$, we assume that space and time are separable. Then, at the current time m, we can estimate $\mu_{C_{j,k}}$ using the following non-parametric estimator:

$$\widehat{\mu}_{C_{j,k}} = \frac{\left(\sum\limits_{i\in S_j}\sum\limits_{t=1}^{m} y_{it}\right)\left(\sum\limits_{i=1}^{H}\sum\limits_{t=k}^{m} y_{it}\right)}{\sum\limits_{i=1}^{H}\sum\limits_{t=1}^{m} y_{it}}\,. \tag{5}$$

Using the above estimate, we have the following monitoring statistic:

$$R_m = \sum_{k=1}^{m}\sum_{j=1}^{H}(1+\varepsilon)^{y_{C_{j,k}}}\exp\left(-\varepsilon\widehat{\mu}_{C_{j,k}}\right) y_{jk}\,.$$

This monitoring statistic does not require the knowledge of the population sizes, only the number of cases in each area.

One can also use a parametric or a semi-parametric estimator for $\mu_{C_{j,k}}$. The rate λ_{it} can be

estimated depending on covariates, as in Höhle (2009). For disease surveillance, the rate λ_{it} can be time varying due to seasonality of the disease. Factors as vegetation, control measures, or the existence of disease vectors, can lead to spatial heterogeneity. Höhle (2009) uses a time-dependent baseline risk $\exp(\lambda_{0t})$ and possible time-dependent $q \times 1$ covariate vector z_{it} to model the rate λ_{it}:

$$\lambda_{it} = \exp(\lambda_{0t} + {z_{it}}^T \beta),$$

where β is a $q \times 1$ vector of coefficients. All spatial heterogeneity is expressed through covariates and the baseline depends on time only.

7.4 Simulation study

We evaluated the performance of the proposed method in different scenarios using Monte Carlo simulation. The geographical region we considered was a regular grid of size 12×12 with $H = 144$ areas. Each area was represented by a square with side 1. The map is observed at times $t = 1, \ldots, 100$. We used four different values for the spatial radius ρ: $0.0, 1.0, 1.5, 2.0$. The value $\rho = 0.0$ means that the circle S_j includes only the area j. Setting $\rho = 1.0$ means that S_j contains five areas: area j and the four areas closest to area j. If $\rho = 1.5$, S_j contains nine areas: area j and the eight areas closest to area j. Finally, $\rho = 2.0$ means that S_j includes twelve areas: area j and the eleven areas closest to area j.

In Section 7.4.1, we evaluate the performance of the method for an under control process in two different situations: when $\mu_{C_{j,k}}$ is known and when it is estimated according to (5). In Section 7.4.2 we analyze the impact of estimating $\mu_{C_{j,k}}$ for both under and out of control processes. In Section 7.4.3 we evaluate the impact of the spatial size of the cluster for an out of control process. In this evaluation, we used $\widehat{\mu}_{C_{j,k}}$ since $\mu_{C_{j,k}}$ is unknown, in practice.

7.4.1 Simulation results for an under control process

The number of cases Y_{it} was generated independently according to a Poisson distribution with mean equal to $P_{it}\,\lambda_{it}$, where $\lambda_{it} = 0.005$ for all i, t. Then, the expected number of cases at a single map is $\mu = 0.005 \sum_{1 \leq i \leq H} P_i \approx 0.005 \times 5500 \times 144 = 3960$. The expected number of cases in the first m maps is equal to $\mu_m = \mu \times m \approx 3960 \times m$.

Figure 9 shows the behavior of the monitoring statistic R_m. In this Figure, the first, second, third and fourth rows of plots correspond to $\rho = 0.0, 1.0, 1.5, 2.0$, respectively. The first, second and third columns of plots correspond to $\varepsilon = 0.01, 0.03, 0.05$, respectively. In all plots,

the horizontal axis represents the time the map was observed. The vertical axis represents the trimmed average value of the monitoring statistic R_m, taken over 300 simulations, divided by μ. This trimmed average excludes the highest and lowest 1% of the data points. The dashed and dotted lines correspond to R_m calculated with known $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$, respectively. The solid line represents $E_{\tau=\infty}(R_m)$, given by (3), divided by μ.

According to Figure 9, the impact of estimating $\mu_{C_{j,k}}$ is a more conservative method. When estimating $\mu_{C_{j,k}}$ one have to wait longer for a false alarm, compared to the waiting time when $\mu_{C_{j,k}}$ is known. The waiting time for a false alarm with $\widehat{\mu}_{C_{j,k}}$ is also longer than the desired *ARL*. The dashed line, that corresponds to R_m calculated with known $\mu_{C_{j,k}}$, exceeds the threshold (solid line) earlier than the dotted line. The dotted line corresponds to R_m calculated with $\widehat{\mu}_{C_{j,k}}$. Even when $\mu_{C_{j,k}}$ is known, the waiting time for a false alarm is longer than the desired *ARL*, except for some especial combinations of ε and ρ. Here, as the process is under control, the true value for the parameter ε is equal to zero, and there is no circle S_j where the per capita rate λ is higher. Then, as the input values of ρ and ε increases, we are more distant from this true situation. When $\mu_{C_{j,k}}$ is known, the especial combinations of ε and ρ for which the waiting time for a false alarm is no longer than the desired *ARL*, are those where the values of ε and ρ are closest to the true situation. Both R_m, calculated with known $\mu_{C_{j,k}}$ and with $\widehat{\mu}_{C_{j,k}}$, move away down from its expected value as ε and ρ move away from their true values. That is, the waiting time for a false alarm increases with ρ and ε, for both $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$. For $\varepsilon = 0.01$, R_m calculated with the known $\mu_{C_{j,k}}$ is very close to its expected value for all values of ρ. The monitoring statistic R_m, calculated with $\widehat{\mu}_{C_{j,k}}$, moves away down from the expected value of R_m as ρ increases. For $\varepsilon = 0.03$, R_m calculated with the known $\mu_{C_{j,k}}$ is very close from its expected value for $\rho = 0.0, 1.0$. R_m calculated with $\widehat{\mu}_{C_{j,k}}$ also moves away down from the expected value of R_m as ρ increases, but faster than when $\varepsilon = 0.01$. For $\varepsilon = 0.05$, this deviation is even faster. In this case, R_m calculated with known $\mu_{C_{j,k}}$ is close from its expected value only for $\rho = 0.0$.

7.4.2 $\widehat{\mu}_{C_{j,k}}$ *versus* $\mu_{C_{j,k}}$ for under and out of control processes

In this section, we analyze the waiting time until the alarm sounds off, for under and out of control process, using both $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$. For the under control process, the number of cases Y_{it} was generated as in Section 7.4.1: $Y_{it} \sim \text{Poisson}(P_{it}\lambda_{it})$, where $\lambda_{it} = 0.005$ for all i,t. For the out of control process, the artificial cluster has spatial radius $\rho^* = 1.00$, $\varepsilon^* = 0.25$ and the change time is $\tau = 15$.

Figures 10 and 11 show summary statistics for the number of time periods until the

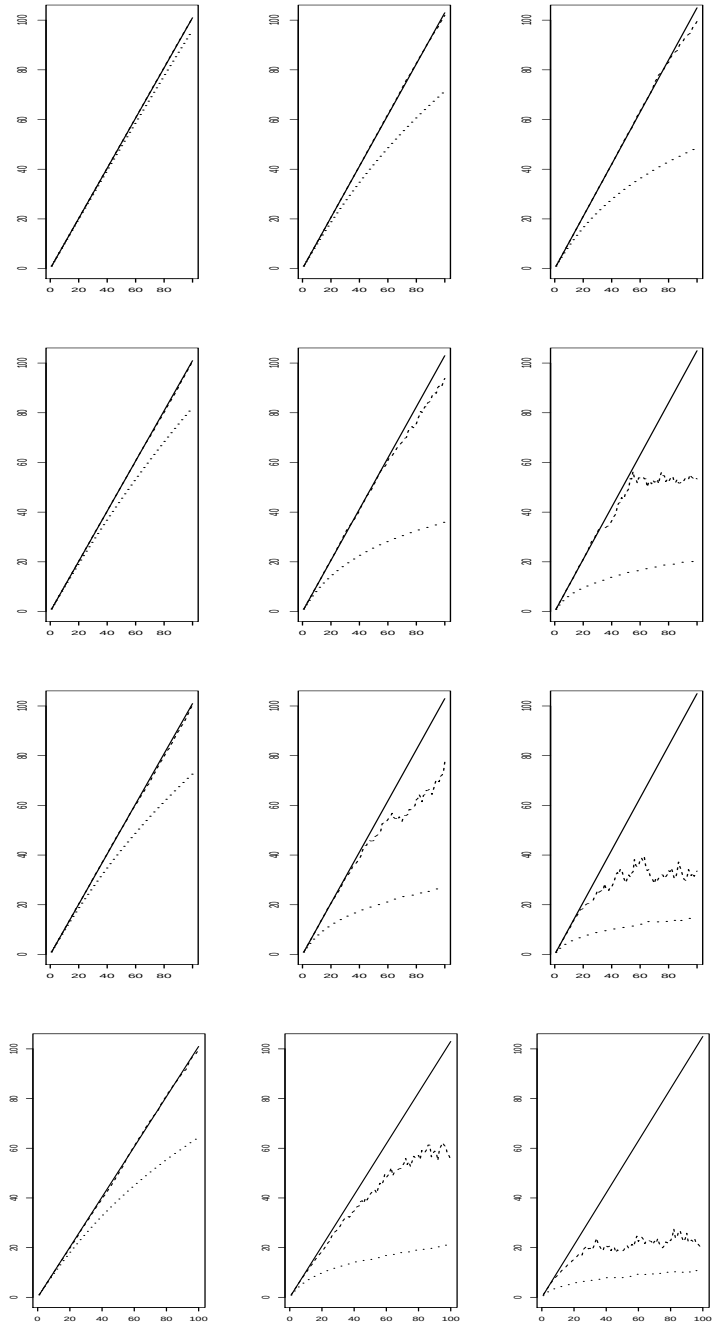

Figura 9: Time m (horizontal axis) versus the trimmed average value of R_m, taken over 300 simulations, divided by μ (vertical axis). The trimmed average value excludes the highest and lowest 1% of the data points. Rows $1,2,3,4$ correspond to $\rho = 0.0, 1.0, 1.5, 2.0$, respectively. Columns $1,2,3$ correspond to $\varepsilon = 0.01, 0.03, 0.05$, respectively. The dashed and dotted lines correspond to R_m calculated with $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$, respectively. The solid line represents $E_{\tau=\infty}(R_m)$ divided by μ.

alarm sounds off in 300 simulations in under control and out of control cases, respectively. In these two Figures, the first, second, third and fourth rows of plots correspond to $\rho = 0.0, 1.0, 1.5, 2.0$, respectively. The first, second and third columns of plots correspond to $\varepsilon = 0.20, 0.25, 0.30$, respectively. In all plots, the horizontal axis represents the threshold limit $A = ARL(1+\varepsilon)\mu$ for $ARL = 20, 30, 40$ time periods. The vertical axis represents the number of time periods until the alarm sounds off. The circles and the triangles represent the average value using $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$, respectively. The segments in each symbol represent one standard deviation above and below the mean. The numbers at the bottom correspond to the proportion of alarms when using $\mu_{C_{j,k}}$ and the numbers at the top correspond to this proportion when using $\widehat{\mu}_{C_{j,k}}$. For the out of control situation, shown in Figure 11, all the statistics (mean, standard deviation and proportion of alarms) are related to motivated alarms. In Figures 10 and 11, the mean and the standard deviation of the number of time periods until the alarm sounds off (or until a motivated alarm, in the out of control case corresponding to Figure 11) are censored measures, since the proportion of alarms is always smaller than one.

Figure 10 shows that, for an under control process, the proportion of simulations in which we had alarms is always smaller when using $\widehat{\mu}_{C_{j,k}}$. This is an expected behavior, since according to Figure 9 we have to wait longer for a false alarm when using $\widehat{\mu}_{C_{j,k}}$. In both cases, using $\mu_{C_{j,k}}$ or $\widehat{\mu}_{C_{j,k}}$, the proportion of alarms decreases as ε increases, for fixed ρ. For fixed ε, the proportion of alarms decreases as ρ increases. These two behaviors were also expected. Figure 9 shows that R_m, calculated with $\mu_{C_{j,k}}$ or $\widehat{\mu}_{C_{j,k}}$, decreases as ε and ρ increases. This decrease of the monitoring statistic R_m leads to a decrease in the proportion of alarms. Here, as we already mentioned, as ε and ρ increase we are more distant from the true situation ($\varepsilon^* = 0.0$ and no circle S_j). The difference between the proportion of alarms with $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$ decreases as we move away from this true situation.

Figure 11 shows that the best performance of the method (the higher proportion of motivated alarms and the lowest standard deviation of the number of time periods until a motivated alarm), occurs when ρ is equal to its true value ($\rho = \rho^* = 1.0$). In this case, the parameter ε seems to have a small impact in all statistics. For all value of ε, the proportion of motivated alarms is very similar (around 0.9) when using $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$. The mean of the number of time periods until a motivated alarm is also similar, but the standard deviation is larger for $\widehat{\mu}_{C_{j,k}}$. This larger standard deviation for $\widehat{\mu}_{C_{j,k}}$ can be explain by the fact that we introduce another source of variation in the monitoring statistic when estimating $\mu_{C_{j,k}}$. When $\mu_{C_{j,k}}$ is known, the proportion of motivated alarms varies little with both ρ and ε. When using $\widehat{\mu}_{C_{j,k}}$, this variation is larger, and the parameter ρ seems to impact it more. In the out of control situation shown

in this figure, the cost of estimating $\mu_{C_{j,k}}$ is a longer waiting time for a motivated alarm. If ρ is set equal to its true value ρ^*, then this cost is quite small. Except for $\rho = \rho^*$, this cost increases with ρ and ε.

7.4.3 Impact of the spatial size of the cluster

We used out of control processes, with $\widehat{\mu}_{C_{j,k}}$ in the monitoring statistic, to evaluate the impact of the spatial size of the cluster. The input value for ρ was always set equal to the true value ρ^*, taken as $\rho^* = 0.0, 1.0, 1.5, 2.0$. The number of cases Y_{it} was generated as in Section 7.4.1: $Y_{it} \sim \text{Poisson}(P_{it}\,\lambda_{it})$, where $\lambda_{it} = 0.005$ for all i,t. We used $\varepsilon^* = 0.25$ and $\tau = 15$.

Figure 12 shows summary statistics for the observed number of time periods until a motivated alarm in 300 simulations. Each plot corresponds to one value for the spatial radius $\rho = \rho^*$. In all plots, the horizontal axis represents the threshold limit $A = ARL(1+\varepsilon)\mu$ for $ARL = 20, 30, 40$ time periods. The vertical axis represents the observed number of time periods until the motivated alarm. The traces, crosses, and losanges represent the average value for $\varepsilon = 0.20, 0.25, 0.30$, respectively. The segments in each symbol represent one standard deviation above and below the mean. The numbers at the top correspond to the proportion of motivated alarms. For each threshold limit, the first, second and third numbers are related to $\varepsilon = 0.20, 0.25, 0.30$, respectively.

Figure 12 shows that, as expected, the proportion of motivated alarms increases as the spatial size of the cluster (ρ^*) increases. The standard deviation of the number of time periods until a motivated alarm decreases as ρ^* increases. It seems that there is no significant impact of the parameter ε in the proportion of motivated alarms. For fixed ρ, the mean of the number of time periods until a motivated alarm is almost the same for all values of ε, but the standard deviation increases with this parameter. Except for a cluster composed by only one area ($\rho^* = 0.0$), the method seems to be efficient. This efficiency is in the sense that the proportion of motivated alarms is high and the mean waiting time for the alarm, after the change time τ, is small. Here $\tau = 15$ time periods, and the mean number of time periods until the motivated alarm is around 20. Then, the waiting time for the alarm, after the change time $\tau = 15$, is around 5 time periods only.

7.5 Illustrative example

Infection with meningococci in human generates public concern because of its often lethal outcome, and its occasional appearance in clusters. We analyzed a meningococcal disease

database provided by the National Reference Center for Meningococci, German Ministry of Health. The database contains 636 meningococcal cases that occurred in Germany from 2002 to 2008.

Figure 13 shows all the available 636 meningococcal cases from 2002 to 2008 on a map of Germany, divided in federal states. There are two areas in the map where the number of cases appears to be higher: Berlim and the Ruhr district. The Ruhr district is an urban area in North Nordrhein-Westfalen. In fact, a higher number of cases in these two areas is expected since they are areas with large population. Berlin is the largest German city with a population of more than 3.4 million inhabitants. The Ruhr district is the largest urban agglomeration in Germany, with a population density of 1195.04 inhabitants per km^2.

The time series plot in Figure 14 shows the number of cases per year/quarter. In all the years, the number of cases is larger in the first quarter. It is due to the seasonality of the meningococcal disease, with a larger incidence in the winter. The cold causes people to agglomerate, facilitating the transmission of the disease.

We applied our method using the number of cases in each state at each quarter. We used three different values for the parameter ε: $0.20, 0.25, 0.30$. The spatial radius ρ was set as 20% of the maximum distance between the states. Figure 15 shows the value of the monitoring statistic R at each quarter. The solid, dashed and dotted lines correspond to $\varepsilon = 0.20, 0.25,$ and 0.30, respectively.

We adopted the threshold $A = ARL(1+\varepsilon)\mu$, where μ was set as the total number of cases (636) divided by the number of quarters (28). We tried seven different values for the ARL: $4, 8, 12, 16, 20, 24, 28$ quarters. Table 2 shows, for each ε and each ARL, the number of quarters until the alarm sounds off and the estimate for the spatial location of the cluster. The first value in each cell is the number of quarters until the alarm sounds off, and the second one is the estimate for the spatial location (identification number of the state). NA means that the alarm did not sound off. The alarms around quarter number 20 (last quarter of year 2006) are related to the peak in the monitoring statistic around this quarter, observed in the three curves of Figure 15. The spatial estimate for the cluster in these cases is the state number 2 (Bayern). For the considered value of the spatial radius ρ, the circle S_j, $j = 2$ (Bayern), includes only the Bayern state. For the alarms observed when using $ARL = 4$ and 8 quarters, the spatial estimate for the cluster are states numbers 4 (Brandenburg) and 10 (Nordrhein-Westfalen). For the considered value ρ, the circle S_j, $j = 4$ (Brandenburg), includes three states: Brandenburg, Sachsen and Berlin. The circle S_j, $j = 10$ (Nordrhein-Westfalen), includes only the Nordrhein-Westfalen state.

Tabela 2: In each cell, the first value is the number of quarters until the alarm sounds off, and the second one is the estimate for the spatial location of the cluster (identification number of the state). NA means that the alarm did not sound off.

ε	*ARL* **(quarters)**						
	ARL=4	**ARL=8**	**ARL=12**	**ARL=16**	**ARL=20**	**ARL=24**	**ARL=28**
$\varepsilon = 0.20$	5, 10	11, 10	17, 2	19, 2	25, 2	NA	NA
$\varepsilon = 0.25$	6, 4	12, 10	17, 2	19, 2	20, 2	NA	NA
$\varepsilon = 0.30$	6, 4	12, 10	17, 2	19, 2	19, 2	20, 2	NA

7.6 Final considerations

We analyzed the characteristics of the proposed method using under an out of control simulated processes. We believe that the results we observed in the particular cases of our simulation represent the general characteristics of the method. Future studies could be done using different scenarios to check the performance of the method.

The results we present in Section 7.4 show that, the cost of estimating $\mu_{C_{j,k}}$ is a more conservative method. When estimating $\mu_{C_{j,k}}$ the waiting time for an alarm, under $\tau = \infty$, is longer than the desired *ARL*, and also longer than when $\mu_{C_{j,k}}$ is known. Furthermore, the waiting time for a false alarm increases with ρ and ε, for both $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$.

Under $\tau < \infty$, the impact of estimating $\mu_{C_{j,k}}$ is a longer waiting time for a motivated alarm. However, if ρ is equal to its true value ρ^*, this impact is very small. In this situation ($\rho = \rho^*$), we observed the higher proportion of motivated alarms and the lowest standard deviation of the number of time periods until a motivated alarm. For known $\mu_{C_{j,k}}$, we observed a little variation in the proportion of motivated alarms with both ρ and ε. When estimating $\mu_{C_{j,k}}$, this variation is larger, and the parameter ρ seems to impact it more than the parameter ε.

As expected, the waiting time for a motivated alarm decreases as the spatial size of the cluster increases, when using $\widehat{\mu}_{C_{j,k}}$. In this case, if the input value of ρ is equal to its true value ρ^*, the proportion of motivated alarms was observed to be high, and the mean waiting time for a motivated alarm was observed to be small, except for a cluster formed by only one area ($\rho = \rho^* = 0.0$).

One should be aware of the cost of estimating $\mu_{C_{j,k}}$ and the cost of setting the spatial radius ρ too different from its true value (unknown, in practice) when applying the method. We believe that the method can used with different values for ρ and ε as a guide for on line disease control.

8 Um olhar cuidadoso sobre vigilância prospectiva usando uma estatística scan

Esta seção traz o artigo “A Close Look on Prospective Surveillance Using a Scan Statistic’, submetido ao periódico *Biometrics*.

8.1 Abstract

The scan statistic is undoubtedly a great success. It was extended to the prospective case and it has a lot of successful applications. However, we feel that there are some difficulties of interpretation related to the prospective scan statistic that require the attention of the statistical community. In this paper we used simulation results from the *SaTScan*™ software to raise these difficulties systematically. Our aim is that it can be proposed statistical solutions. We also analyzed some characteristics of the scan statistic through simulation results. For the prospective context, we evaluated the average run length and compared it with the recurrence interval.

8.2 Introduction

Epidemiologists typically perform geographical surveillance of diseases to detect statistically significant temporal, spatial or space-time disease clusters. Looking for hints about unknown risk factors, they also want to test whether a disease is randomly distributed over space, over time or over space and time. This procedure can help on the evaluation of the statistical significance of disease cluster alarms. The scan statistic method proposed by Kulldorff (1997) is a superb method for these purposes. For retrospective surveillance and cluster detection, it is one the most popular methods among public health officials and researchers and the *SaTScan*™ software (see Kulldorff, 2003) is a widely used free implementation of the method with users from many countries.

The main reason for the widespread popularity of the scan statistic method for retrospective surveillance and cluster detection is its control of error type I probability over multiple tests. When searching for temporal or geographical disease clusters, there are a huge number of potential candidates due to the overwhelming number of combinations of areas or times to form a cluster. Carrying out a statistical significance test for each potential cluster candidate leads to a large number of false positives, an undesirable situation that Kulldorff (1997) solved in a simple way. The scan statistic is based on the attained maximum likelihood ratio

for a simple comparison of disease occurrence between two groups, within and outside the candidate cluster. This maximum is obtained by scanning all possible candidate clusters and hence the multiple tests situation is reduced to a single test situation. Its statistical significance is evaluated by means of Monte Carlo replications. This false positive control feature, coupled with its good power performance in simulation studies, transformed the scan statistic into a standard test for retrospective surveillance and cluster detection problems.

The scan statistic methodology has also been proposed to prospective surveillance, when we perform repeated time periodic disease surveillance for early detection of disease outbreaks. This is an important issue for public health agencies and, unfortunately, in contrast with retrospective surveillance and cluster detection, prospective space-time surveillance models are rare. Among the few options available, we have the methods proposed by Kulldorff (2001), Kulldorff et al. (2005), Takahashi et al. (2008), Tango et al. (2011). All these proposals are based on the repeated application of the scan statistic as time accrues. Recent reviews covering the space-time situation include Woodall et al. (2008) and Unkel et al. (2011).

While the scan statistic method proposed by Kulldorff (1997) is a standard method for cluster detection in a retrospective setting, we feel that it has many difficulties to deal with the prospective situation. The essence of this difficulty is that the scan statistic is based on a statistical hypothesis testing framework that is not adequate to the prospective situation. Woodall et al. (2008) pointed this out by criticizing the use of p-values and recurrence intervals, since these measures do not reflect appropriately the statistical performance of procedures repeated indefinitely.

Statistical hypothesis testing uses error type I and II probabilities as performance measures and tuning parameters. These concepts are not meaningful when a statistical test is applied sequentially with no pre-defined number of repeated applications. This is different from the retrospective situation, when we scan over a number of potential candidates generating a large but pre-established number of statistics. When the number of test statistics is undefined, the error type I and II probabilities may be equal to 1 and 0, respectively, rendering these performance measures worthless in the prospective case. For example, consider the most simple prospective surveillance method for a sequence of iid random variables $Z_1, Z_2, \ldots$ in which $Z_i \sim N(0,1)$ in the *in control* state and $Z_i \sim N(1,1)$ in the *out of control* state. A Shewart chart procedure declares that the system is out of control when the first Z_i is larger than a threshold c. If the system is run indefinitely under the in control state, it is clear that $\mathbb{P}(\min_i\{Z_i > c,\ i = 1,2,...\} < \infty) = 1$. Therefore, any meaningful error type I probability definition will be equal to 1 in this case. In the same way, under the out of control state, the

error type II probability is equal to 0 for any c.

Instead of error type I and II probabilities, one should use more appropriate measures for the prospective context. We say that an alarm goes off at time t when we have some empirical evidence that the system under surveillance has changed from the in control state to the out of control state. In this situation, it is more common to use the Run Length (*RL*), the waiting time until the alarm goes off. The Average Run Length (*ARL*) is the expected *RL* when the process is under control. One tries to fix a target *ARL* and aims for a procedure that minimizes the expected waiting time for a true alarm when the process is out of control. This is called the conditional expected delay in the quality control literature.

One of the main advantages of the scan test statistic, its error type I probability control, is not guaranteed over the multiple sequential tests in the prospective situation. Kulldorff (2001) was aware of this difficulty and he proposed an adjustment by establishing critical thresholds for the test statistic at each time based on its previous values. Unfortunately, this adjustment is not enough to overcome the problem, as we will show. Both, Kulldorff (2001) and Kulldorff et al. (2005), are highly cited papers but it seems that this prospective issue is not sufficiently clear in the application of the prospective scan. For example, Tango et al. (2011) proposed recently a modification of Kulldorff (2001) that is perfectly fine in the retrospective setting but that retains part of the major problems we see in the scan method in the prospective setting. Since their method is proposed for both situations, we decided to write this note where the prospective difficulties are highlighted.

In this paper, we aim to show through a simulation study using the software *SaTScan*™ that, when applied in a prospective way to detect emerging clusters, the scan statistic methods from Kulldorff (2001), Kulldorff et al. (2005), Takahashi et al. (2008) and Tango et al. (2011) are difficult to interpret. More specifically, the scan statistic do not adjust for the sequential and repeated tests carried out during the surveillance. We insist on the specificity of our comments: they are directed towards the use of the scan testing methodology in the prospective scenario. We think that the scan statistic is an excellent technique for cluster detection in the retrospective situation.

In section 8.3, we review the prospective methods based on the scan statistic. In section 8.4, we present simulation results that highlight the problems with the prospective scan statistic. We analyzed some important aspects of the scan statistic in retrospective and prospective cases. For the prospective situation, we evaluated the behavior of the p-values and the proportion of alarms when the scan statistic is applied sequentially as data become available. We considered different possibilities, with and without adjustments for earlier analysis, showing

that these adjustments do not work as expected. We also verified the quality of the recurrence interval, as considered by Kleinman (2005) and Kleinman et al. (2004) in the spatial-temporal aggregated case, to estimate the average run length of the scan prospective method. We close with final considerations in section 8.5.

8.3 The Scan Statistic for Emerging Outbreaks

In this section we review the scan-based methods of Kulldorff (2001) and Tango et al. (2011).

8.3.1 Kulldorff (2001)

The purely spatial scan statistic uses a circular window that moves on the map, including different sets of neighboring areas. The radius of each circle increases so that the circle includes at most 50% of the total population at risk. The number of events can be considered either Poisson or Bernoulli distributed. The spatial scan statistic S is the maximum likelihood ratio over all possible circles Z, conditioning on the observed total number of cases N,

$$S = \frac{\max_Z [L(Z)]}{L_0} = \max_Z \left(\frac{L(Z)}{L_0} \right) \tag{6}$$

where L_0 is the likelihood function under the null hypothesis of a purely random Poisson process and $L(Z)$ is the maximum likelihood for circle Z.

Define n_Z as the number of cases inside circle Z. Assuming the Poisson model, $\mu(Z)$ is the expected number of cases under the null hypothesis. Kulldorff (1997) shows that

$$\frac{L(Z)}{L_0} = \left(\frac{n_Z}{\mu(Z)} \right)^{n_Z} \left(\frac{N - n_Z}{N - \mu(Z)} \right)^{N - n_Z}$$

if $n_Z > \mu(Z)$ and $L(Z)/L_0 = 1$ otherwise.

Circles containing more than half the population at risk would be more suitably interpreted as a negative cluster of lower risk. The choice of 50% of the population at risk as the maximum circle is intuitive, since it ignores these type of clusters.

The space-time scan statistic uses a cylindrical window in three dimensions, instead of a circular window in two dimensions. The base of the cylinder represents the space, and the third dimension is the time. Let s and t be the start and end dates of a cylinder, respectively. Let $[Y_1, Y_2]$ be the time interval for which data exists. The prospective space-time scan statistic

considers all cylinders for which $Y_1 \leq s \leq t = Y_2$. That is, in the prospective context, the space-time scan statistic considers only alive clusters - clusters that reach the actual time.

To evaluate the statistical significance for a cluster, the distribution of the statistic given in (6) under the null hypothesis is constructed by Monte Carlo replications. For the random data sets, cases are generated so that space and time are independent. To solve the problem of multiple testing, the likelihood for the random data sets is maximized over all cylinders used in the previous analysis in addition to the current cylinders, i.e, those cylinders for which $Y_1 \leq s \leq t \leq Y_2$ and $t \geq Y_m$, where Y_m is the time in which the surveillance began. At a given moment, if the probability of having detected a cluster with higher likelihood during any of the previous analysis or the present analysis is at most α, then the observed cluster is statistically significant at the α level. That means, using the random data sets, it is possible to find the critical value for the α level of significance.

The method might be too conservative to adjust for all previous analysis if it is in place for a long period. It is possible to handle this problem by including only those cylinders analyzed during he preceding v times for the random data sets, i.e, those cylinders for which $Y_1 \leq s \leq t$ and $Y_2 - v < t \leq Y_2$.

8.3.2 Tango et al. (2011)

Tango et al. (2011) criticized the conditional expected number of cases used in Kulldorff (2001). They propose an outbreak model based on a new space-time scan statistic which compares the observed number of cases with the unconditional expected number of cases.

Under the null hypothesis of no outbreaks, N_{it}, the number of cases in region i at time t is assumed to be independent negative binomial distributed

$$H_0 : N_{it} \sim NB(\mu_{it}, \phi_{it}),$$

where $E(N_{it}) = \mu_{it}$ and $Var(N_{it}) = \mu_{it} + (\mu_{it})^2/\phi_{it} = \mu_{it} w_{it}$.

The temporal overdispersion is given by

$$w_{it} = 1 + \frac{\mu_{it}}{\phi_{it}}.$$

The parameter ϕ_{it} regulates the overdispersion. If there is no overdispersion in region i, then $w_{it} = 1$ or $\phi_{it} = \infty$.

The parameters μ_{it} and ϕ_{it} are estimated using data from a predefined baseline period and applying a negative binomial regression model:

$$\log E(Y_{it}) = \sum_{j=1} x_{it\,j}\beta_j + b_i,\, Y_{it} \sim NB(\mu_{it}^{(Y)}, \phi_i^{(Y)}),$$

where Y_{it} is a random variable for the observed count data in the predefined baseline period, b_i is a random regional effect independently normally distributed with mean 0, $x_{it\,j}$ is the value of covariate j, and ϕ_{it} is usually assumed to be constant over time, i.e., $\phi_{it} = \phi_i$. Then, $\mu_{it} = \hat{\mu}_{it}^{(Y)}$ and $\phi_{it} = \hat{\phi}_i^{(Y)}$.

If the surveillance system starts in the recent past, a moving average method is used to estimate μ_{it} and ϕ_{it}, instead of the negative binomial regression model. In this case, μ_{it} is assumed to be constant during the baseline period and the parameters μ_{it} and ϕ_{it} can be estimated using baseline mean and variance respectively.

The alternative is given by

$$H_1 : N_{it} \sim NB(\theta_{it}\mu_{it}, \phi_{it}),$$

where θ_{it} is the relative risk in region i at time t and (μ_{it}, ϕ_{it}) are known.

The outbreak model is given by

$$\theta_{it} = \begin{cases} h(\tau + \beta_W(t - t_p + u)), & \text{if } (i,t) \in W = Z \times I_u \\ 1, & \text{otherwise}, \end{cases}$$

where $h(\tau)$ is the relative risk at $t = t_p - u$. The initial slope of emerging disease outbreak starts just after the time point $t_p - u$ within the domain W. It is given by

$$\left[\frac{\partial \theta_{it}}{\partial t}\right]_{t=t_p-u} = \beta_W h'(\tau)$$

where $h(.)$ is any monotonically increasing function so that $h(\tau) = 1$ and the differentials $h'(.)$ and $h''(.)$ are both finite. The above hypothesis testing is then reduced to the following hypothesis testing over all possible sets $W = Z \times I_u$:

$$H_0 : \beta_W = 0, \;\; H_1 : \beta_W > 0\,.$$

The likelihood function under the outbreak model is know, but the likelihood ratio test for $H_0 : \beta_W = 0$ against $H_1 : \beta_W > 0$ requires the maximum likelihood estimator for β_W and the functional form of $h(.)$.

Let

$$S_1 = \sup_{1\leq u\leq T, Z\in\mathcal{Z}} \frac{\sum_{i\in Z}\sum_{t\in I_u}(n_{it}-\mu_{it})(t-t_p+u)/w_{it}}{\sqrt{\sum_{i\in Z}\sum_{t\in I_u}\mu_{it}(t-t_p+u)^2/w_{it}}} \sim N(0,1)\,.$$

S_1 is a score test statistic for $H_0 : \beta_W = 0$. The score test based on S_1 does not depend on neither the maximum likelihood estimator for β_W nor the functional form of $h(.)$ and it is asymptotically equivalent to the likelihood ratio test.

One should use $w_{it} = 1$ for all regions if the Poisson distribution is adopted instead of the negative binomial distribution. The most likely outbreak is identified by the domain $W^* = Z^* \times I_{u^*}$ that maximizes the score statistic. Monte Carlo simulated p-value is defined as in Kulldorff (2001).

If a hotspot cluster model is assumed

$$\theta_{it} = \begin{cases} \tau_W(>1), & \text{if } (i,t) \in W = Z \times I_u \\ 1, & \text{otherwise}\,, \end{cases}$$

then the score test statistic for $H_0 : \tau_W = 1$ is

$$S_2 = \sup_{1\leq u\leq T, Z\in\mathcal{Z}} \frac{\sum_{i\in Z}\sum_{t\in I_u}(n_{it}-\mu_{it})/w_{it}}{\sqrt{\sum_{i\in Z}\sum_{t\in I_u}\mu_{it}/w_{it}}} \sim N(0,1)\,.$$

If the Poisson distribution is used, the score statistic is reduced to

$$S_3 = \sup_{W\in\mathcal{W}} \frac{n(W)-\mu(W)}{\sqrt{\mu(W)}} \sim N(0,1)$$

which is asymptotically equivalent to the unconditional likelihood ratio test, since the expected number of cases $\mu(W)$ is calculated unconditionally from the baseline data.

8.4 Simulation Results

In this section, we present the results of the prospective and retrospective scan statistics for purely temporal analysis. The issues we want to discuss are presented in both, space-time and purely temporal contexts. Addressing the purely temporal case is simpler to understand and simulate, and it is enough to show the problems implied by the prospective situation. We generated 1000 time series of iid random variables $Y_1, \ldots, Y_{400}$ with Poisson distribution

with mean equal to 3. For each time series, a sequential analysis is carried out. We begin with a time series of length $n = 100$ and increased it sequentially until the length reached $n = 400$. That is, the first n observations for a time series of length $n+1$ are the same as the n observations for the series of length n. For each time series and each length n, we used the *SaTScan*$^{\text{TM}}$ software selecting always 999 Monte Carlo replications in each analysis to obtain the critical value for $\alpha = 0.05$ significance level and the p-value. In all analysis, we used the maximum temporal cluster size as 50% of the study period, as suggested by Kulldorff (2001). We considered four different situations: retrospective scan, prospective scan with no adjustment, prospective scan adjusting for all previous analysis, and prospective scan adjusting for a fixed number of analysis.

Figure 16 shows the results for the retrospective scan (here named situation (a)). The horizontal axis in all three graphs represents the length n of the time series. In this retrospective situation, the scan statistic analysis is carried out at each n without any concerns with the analysis carried out in different time series lengths. That is, the analysis at length n is run as a single analysis, as if no other analysis would be carried out in later moments. For each time series length n, we used the *SaTScan*$^{\text{TM}}$ software to calculate the scan test statistic given in (6). The set of candidate clusters were all the time intervals up to time n, including the non-alive ones. We denote this test statistic by S_n. We also took the critical value associated with the 5% significance level and the p-value from *SaTScan*$^{\text{TM}}$ software. This p-value is correctly obtained by evaluating the test statistic S_n in many Monte Carlo replications in each time series and each length n. That is, for each n and each realized time series Y_t, we replicated 999 time series with the same mean as the observed series, evaluated the test statistic S_n considering all possible clusters, alive at n or not, and obtained the p-value by ranking the observed S_n with respect to these replicated S_n values. We repeated this procedure 1000 times. Then, for each time series length n, we can calculate summary statistics for the critical value and p-value.

The first graph shows in solid line the average critical value for the scan statistic taken over the 1000 simulations together with the pointwise 95% confidence bands $L(n)$ and $U(n)$ in dashed lines. That is, 2.5% of the 1000 simulations were below $L(n)$ at each time series length n, and 2.5% of them were above $U(n)$ at each length n. There is a continuous, approximately linear, increase in the critical value with increasing time series length n. This is due to the larger number of temporal clusters scanned and the implied larger variance of the scan statistic. The second graph shows in solid line the p-value averaged over the 1000 simulations together with the 95% confidence bands for the observed p-value at each time series length

n (in dashed lines). As predicted by the theory, the p-value at each length n should have a uniform distribution in $(0,1)$ and this plot is in agreement with this result. The third graph shows the proportion of simulations in which the p-value is at most α. It shows, for each time series length n, an estimate of the error type I probability which is known to be 0.05. The clear serial correlation in this plot is due to the high autocorrelation in the p-value series for each individual time series realization. That is, the p-value for a realized time series of length $n+1$ is typically very close to the p-value of that same series taken only up to length n.

Figures 17 to 19 are identical to Figure 16, except by the type of analysis. They correspond to different types of prospective scan statistic analysis. That is, the candidate clusters for a time series of length n must include the current last observation at time n. The scan test statistic given in (6) is calculated using only alive clusters and to differentiate it from the situation (a) we denote this test statistic by S_n^a, where the index a stands for alive. In these prospective time periodic analysis, we will say that we have an alarm at length n when its corresponding p-value is smaller or equal than $\alpha = 0.05$. The calculation of the p-value uses a different reference distribution in each figure, as we explain next.

Figure 17 shows the results for the prospective scan with no adjustment for earlier analysis (named situation (b)). There are two differences between this situation and situation (a), both connected to the alive status of the candidate clusters. One of them is the test statistic (6), which is equal to S_n in situation (a), and equal to S_n^a in situation (b). The other difference is the reference distribution to calculate the p-value. In situation (a), we replicated 999 time series and obtained S_n in each one of them to generate a Monte Carlo p-value. In situation (b), we obtained S_n^a in each replicated time series and used these values to calculate the Monte Carlo p-value.

Figure 18 corresponds to the prospective scan adjusting for all previous analysis (named situation (c)). The only difference between this situation and situation (b) is the calculation of the p-value. At length n, we find the most likely alive cluster and its associated scan test statistic S_n^a. The p-value is calculated empirically by considering the rank of the observed S_n^a with respect to the distribution of $S_{n,n}^a = \max_m\{S_m^a \text{ for } m \leq n\}$. This distribution is found by Monte Carlo replications in each particular time series Y_t. We replicated 999 time series and obtained S_n^a in each one of them, as well as S_m^a for $m < n$. Note that we ignore the time series values after m when calculating these S_m^a. We then calculated the $S_{n,n}^a = \max\{S_n^a, S_{n-1}^a, S_{n-2}^a, \ldots\}$ in each replication. These maxima compose the reference distribution. It is important to observe that the reference distribution for the test statistic S_n^a uses a different statistic, the maximum $S_{n,n}^a$. With respect to situation (a), the situation (c) has two differences: the test

statistic, which is S_n in (a), and the reference distribution, which always considers series of fixed length at each n in (a).

Finally, Figure 19 corresponds to the prospective scan adjusting for the last 100 analysis (situation (d)). In this case, the only difference with respect to the situation (c) is that the reference distribution is that of $S^a_{n,99} = \max_m\{S^a_m \text{ for } n-99 \leq m \leq n\}$.

Considering Figure 17, we can see in the first plot that the range of the average threshold is smaller than in situation (a). This is due to the smaller set of candidate clusters. In situation (a), all possible clusters, alive or not, are under consideration while, in situation (b), only those alive at the given length are considered. As a consequence, the range of the scan statistic (6) and its threshold are smaller. There is a strong increasing trend in this average threshold but, in contrast with Figure 16, we can not find clear evidence of a linear trend due to the large variability present in this plot. As in Figure 16, the average p-value is around 0.5 and its distribution agrees with a uniform distribution in $(0,1)$. Since this is a prospective situation, the vertical axis in the third plot is labeled as proportion of alarms. It shows the proportion of times the 1000 simulated time series of length n had a p-value smaller or equal than $\alpha = 0.05$. For each fixed n, this proportion should be around 0.05. Comparing with Figure 16, we see less serial correlation in this plot. In situation (a) the candidate cluster is often the same when we increment the time series from length n to $n+1$. In contrast, in situation (b), the candidate cluster must change from length n to $n+1$ since it must include the most recent observation of the time series. This weakens the serial correlation of the p-values observed in situation (a).

The first plot of Figure 18 is almost identical to the corresponding plot in Figure 16. The values in this graph refer only to the reference distribution used to evaluate the p-value of a given scan statistic. Since the reference distributions are almost the same in situations (a) and (c), one can anticipates that the average thresholds will be also almost the same.

Substantial differences between the situations (a) and (c) start to show up in the second plot, where we can see the average p-value and the lower band $L(n)$ increasing towards 1 as the length n increases. The upper band $U(n)$ is equal to 1 for all n. The reference distribution in these two situations are almost the same but the test statistic is different. In situation (a), the test statistic is S_n. In situation (c), it is S^a_n. This leads to very different p-values in these two situations. Because of the apparent p-value convergence to 1, there could be a positive probability that the alarm never goes off, even if we allow the time series length n goes to ∞. This would imply an $ARL = \infty$, which illustrates our argument that the error type I probability α is not meaningful in the prospective context. The third plot in Figure 18 shows that, for all

length n, there is no relationship between the nominal value $\alpha = 0.05$ and the probability of having an alarm. The discreteness of this last time series of p-values is due to the rareness with which an alarm sounds off making the values to jump between 0, 0.001, 0.002, etc.

The situation (d) is illustrated in Figure 19 and it is similar to situation (c) shown in Figure 18. The differences between the two situations are due to the fixed number of 100 previous analysis used to adjustment in situation (d). Comparing Figures 18 and 19, we see that the average threshold range in the first plot is shorter in situation (d). The reason for that is the larger number of candidate clusters considered in the reference distribution in situation (c). For the same reason, the second plot shows that the average p-value and the lower band $L(n)$ do not increase with the length n, in contrast with the situation in Figure 18. This is the explanation for the higher proportion of alarms in the third plot.

Figures 18 and 19 show that, for the prospective situations with adjustments, the distribution of the p-value is concentrated at the higher end of the $[0,1]$ interval and the proportion of alarms is significantly less than $\alpha = 0.05$ for all length n. For situation (c), the minimum proportion of alarms is 0.000 and the maximum is 0.005, but most part of the values are between 0.000 and 0.003. For situation (d), the minimum proportion of alarms is 0.000 and the maximum is 0.009, but most part of the values are between 0.001 and 0.004. That means that the probability of having an alarm when the process is in control is not controlled when applying the prospective scan with adjustment for all or for a fixed number of previous analysis.

The proportion of simulations with p-value less or equal than the confidence level $\alpha = 0.05$ for at least one length n is 0.18 in situation (a), while the proportion of simulations in which the alarm goes off for at least one length n is 0.90 in situation (b). It is slightly puzzling that these proportions are so different since the average p-value in these two situations is very similar and around 0.5 (see second plots in Figures 16 and 17). We will explain the reason for this difference using Figure 20 in the next paragraph. The proportion of simulations in which the alarm goes off for at least one length n is 0.07 and 0.13 for situations (c) and (d), respectively. Given that the average p-value in these two situations is around 0.95 (according to the second plots in Figures 18 and 19), this is an expected behavior.

For each one of the 1000 independently simulated counting time series $Y_1, \ldots Y_{400}$, there is an associated time series of p-values $p_1, \ldots p_{400}$. The first column of plots of Figure 20 shows the typical behavior of these p-value time series p_t. Situation (a) is in the first row and we show five time series of p-values, associated with five independent simulations of the counting time series Y_t we are monitoring. The horizontal line represents the confidence level $\alpha = 0.05$. The second and third graphs in each row present the distribution of the range and

the variance for the 1000 time series of p-values. The prospective situations (b), (c) and (d) are shown in rows 2, 3, and 4, respectively. Only one p-value time series p_t is shown in the first plot in the last three rows. Given the stable behavior of these p_t time series, a single realization is enough to provide a very good idea of their patterns.

For the retrospective situation (a), the first graph shows that the p-value time series p_t has a very high serial correlation. When the counting time series Y_t is updated increasing its length from n to $n+1$ by the arrival of a new observation, the most likely cluster typically remains the same as that at length n. Usually, the most likely cluster does not includes the last observation available. By not changing the most likely cluster, the test statistic S_{n+1} is almost the same as S_n and, as a consequence, the p-value does not change substantially. However, occasionally, the p-value time series p_t has a very sharp decline. That is, the p-value of the most likely cluster at time $n+1$ becomes much smaller than the p-values of the clusters that had been the most likely ones at the previous times. This collapse happens when the newly arriving observation is such that the most likely cluster changes completely and now includes this observation at length $n+1$. If this completely changed cluster at time $n+1$ is much more likely than the previous cluster at time n, the p-value at $n+1$ is much smaller than the p-value at time n and this explains its occasional sharp decline.

For the prospective situation (b), the most likely cluster changes every time a new observation comes in. This is so because the most likely cluster must be alive and, therefore, it must change by, at least, including the newly arrived observation. This leads to a change in the test statistic S_n^a and its associated p-value. For the prospective situations (c) and (d), the p-value time series has a very high serial correlation, with most values being exactly equal to 1. Eventually, the p-value can have a huge drop but it quickly returns to 1. Having a p-value equal to 1 at time n means that the observed test statistic S_n^a is the minimum among the 999 values of $S_{n,n}^a$ composing the reference distribution in situation (c). In our opinion, this shows how drastic is the adjustment undertaken. The statistic $S_{n,n}^a = \max\{S_n^a, S_{n-1}^a, S_{n-2}^a, \ldots\}$ used to build the reference distribution typically generates much higher values than the observed test statistic S_n^a. In rare occasions, the observed counting time series Y_t will have a few very large counts in the last positions leading to large values of S_n^a and smaller p-values. However, as we saw in Figures 18 and 19, these p-values hardly reach below 0.05.

The plots in the second and third columns of Figure 20 show that, for situation (b), both the range and the variance of the p_t time series are concentrated at higher values when compared to situation (a). As a consequence, there is a larger number of the p_t series crossing the horizontal line $\alpha = 0.05$. For the prospective situations (c) and (d) the number of p_t series

crossing this line is small, since these series are typically concentrated at values around 1.

Only situations (c) and (d) are truly prospective, in the sense that they take into account the prospective time periodic surveillance procedure. One of the main characteristics of sequential methods is the *ARL* and this motivated us to study the distribution of the run length *RL* and the recurrence interval for the prospective situations (c) and (d). The Recurrence Interval (*RI*) is a measure with interpretation similar to the *ARL*. *RI* is defined, under the in control state, as the length of time for which the expected number of alarms is 1. Woodall et al. (2008) criticized the *RI* in the prospective case, since it is not affected by dependences in the counts between regions or between the monitoring statistic values over time.

The proportion of simulations in which the alarm went off at least once is equal to 0.07 and 0.13 for situations (c) and (d), respectively. These small proportions make it difficult to study the *RL* and *RI*. Therefore we adopt longer time series of maximum length $n = 20000$ to perform this study. With this new maximum time series length, the time required to run all the simulations would be extremely long. For example, it took 48 hours to simulate one single time series of maximum length $n = 20000$ and the corresponding p-values for situation (d), using an Intel Core 2 Duo 1.5 GHz processor. To make the analysis possible, we adopt a different procedure to obtain the p-value associated with the scan statistic without using Monte Carlo replications. We will refer to the p-value calculated by this procedure as the p*-value, to differ it from the *SaTScan*$^{\text{TM}}$ p-value calculated by Monte Carlo replications. The procedure to obtain the p*-value has the following steps:

- We generate 266 time series of iid $Y_1, \ldots, Y_{20000}$ random variables with Poisson(3) distribution.
- For each time series, a sequential analysis is carried out. We begin with a time series of length $n = 1$ and increased it sequentially until length $n = 20000$.
- For each realized counting time series and each length n:
  - We obtain the value of the test statistic S_n^a from the *SaTScan*$^{\text{TM}}$ software.
  - We calculated $S_n^* = \max_i\{S_i^a\}$, $1 \leq i \leq n$ for situation (c) and $n - 99 \leq i \leq n$ for situation (d).
- For each time series and length n, we obtain the p*-value as the proportion of time series in which $S_n^* \geq S_n^a$.

Figure 21 shows that the p*-value is a very good approximation for the *SaTScan*^{TM} p-value. This graph refers to situation (d) and it is based on 1000 counting time series of maximum length $n = 100$. We used this small maximum length of $n = 100$ to verify the quality of p*-value as an estimator for the *SaTScan*^{TM} p-value. The good quality of the p*-value as an approximation for the *SaTScan*^{TM} p-value was also observed in situation (c).

Using the 266 p*-values for the counting time series of maximum length $n = 20000$, we obtained RL^* and RI^*. RL^* is the smallest length n so that the p*-value is less or equal than the confidence level $\alpha = 0.05$. That is, the RL^* is the smallest length n so that the alarm goes off. The RI^* is a function of the p*-value associated with this first alarm and the number A of previous analysis to adjust for:

$$RI^* = \frac{1}{1-(1-\text{p*-value})^{(1/A)}}\,.$$

For situation (c), the proportion of alarms is 0.38. Then, the probability of having RL^* larger than 20000 is 62% and there could be a positive probability that the alarm never goes off. This implies $ARL^* = E(RL^*) = \infty$. We already mentioned this behavior when discussing the second plot of Figure 18.

Figure 22 shows the graphs of RL^* versus RI^* for the prospective situations (c) and (d). In both plots the vertical and the horizontal axis are in the logarithm scale. Considering the 38% of the time series in which the alarm went off in situation (c), the first plot of Figure 22 shows that there is a clear linear relationship between RL^* and RI^*. However, the scales in the two axes are completely different and we can conclude that the recurrence interval is not a good estimate of the average run length in this situation. For situation (d), the alarm went off in all simulations but there is no relationship between RL^* and RI^* . The values for RI^* are concentrated at different levels. This occurs because the p*-values used to obtain RI^* are also concentrated at different values. We had already observed this discreteness in the third plot of Figure 19.

8.5 Final Considerations

We analyzed the characteristics of the prospective scan statistic using in control simulated counting time series of random variables with Poisson distribution with mean equal to 3. We consider only this particular case due to the long time required for the simulations. We strongly believe that the results we observed for this particular case represents the general characteristics of the scan statistic, including the space-time context.

The results we present in section 8.4 show that, in the prospective context, the nominal confidence significance level α is not meaningful and there is no relationship between α and the recurrence interval or the average run length. The *ARL* could even be equal to ∞ in some cases. The proposed adjustments for the previous analysis do not solve the problem of the prospective time periodic tests. We hope that the problems with the prospective scan statistic that we showed in this paper can motivate future studies to improve the surveillance methods based on the scan statistic.

9 Sistemas Alternativos

No trabalho apresentado nesta seção, nós propomos três sistemas prospectivos para vigilância espaço-tempo em dados pontuais. A seção 9.1 traz uma revisão de algumas propriedades do passeio aleatório com barreiras. Estas propriedades serão utilizadas para determinar um dos parâmetros dos sistemas que iremos propor. Na seção 9.2 apresentamos os três sistemas. A seção 9.3 traz um estudo do desempenho dos sistemas propostos via simulação. Comparamos a performance destes sistemas com a do sistema proposto em Assunção and Correa (2009).

9.1 Passeio aleatório com barreiras

Nas seções 9.1.1 e 9.1.2 consideramos o passeio aleatório com uma barreira absorvente e uma barreira refletora de uma forma geral. Na seção 9.1.3 consideramos esta mesma situação no contexto de vigilância.

9.1.1 Passeio aleatório com uma barreira absorvente em 0 e uma barreira refletora em $b > 0$

Considere uma partícula que está na posição inicial u no eixo x ($0 < u \leq b$ inteiro) no tempo inicial $t = 0$. A cada instante de tempo $t = 1, 2, \ldots$, a partícula dá um passo para a direita ou para a esquerda. Cada passo é de uma unidade. Os passos são independentes. A barreira em $x = 0$ é absorvente e a barreira em $x = b$ é refletora. Isto significa que quando a partícula atinge a barreira 0 ela é absorvida e o processo termina. Quando a partícula atinge a barreira b existe uma probabilidade p dela permanecer aí no próximo instante de tempo e existe uma probabilidade q dela mover uma unidade para a esquerda. A partícula dá um passo no sentido da barreira absorvente (neste caso um passo para a esquerda) com probabilidade q. A partícula dá um passo no sentido da barreira refletora (neste caso um passo para a direita) com probabilidade p ($q + p = 1$).

Seja T o número de passos até que a partícula seja absorvida. De acordo com Weesakul (1961):

$$E(T) = \begin{cases} u + u(2b - u) & \text{se} \quad p = q = 0,5 \\ \\ \frac{u}{q-p} + \frac{p^{b+1}}{q^b(q-p)^2}[1 - (q/p)^u] & \text{se} \quad p \neq q \end{cases}$$

9.1.2 Passeio aleatório com uma barreira absorvente em $b > 0$ e uma barreira refletora em 0

Considere agora o problema descrito anteriormente, porém com uma barreira absorvente em $x = b$ e uma barreira refletora em $x = 0$. A partícula dá um passo no sentido da barreira absorvente (neste caso um passo para a direita) com probabilidade q. A partícula dá um passo no sentido da barreira refletora (neste caso um passo para a esquerda) com probabilidade p. Neste caso, se p=q=0,5

$$\begin{aligned} E(T) &= (b-u)+(b-u)(2b-(b-u)) \\ &= b-u+b^2-u^2\,. \end{aligned}$$

Se $p \neq q$,

$$E(T) = \frac{b-u}{q-p} + \frac{p^{b+1}}{q^b(q-p)^2}[1-(q/p)^{b-u}]\,.$$

Em particular, se a posição inicial da partícula é $u = 0$ temos:

$$E(T) = \begin{cases} b+b^2 & \text{se} \quad p = q = 0,5 \\ \frac{b}{q-p} + \frac{p^{b+1}}{q^b(q-p)^2}[1-(q/p)^b] & \text{se} \quad p \neq q \end{cases}$$

9.1.3 Passeio aleatório no contexto de vigilância

Considere o problema do passeio aleatório descrito anteriormente, com uma barreira absorvente em $x = A$ e uma barreira refletora em $x = 0$. A posição inicial da partícula é $u = 0$.

Seja

$$B_n = \begin{cases} -1 & \text{com probabilidade} \quad p \\ 1 & \text{com probabilidade} \quad q \end{cases}$$

Seja $S_n = \max(0, S_{n-1} + B_n)$ a estatística de teste, cujo valor inicial é $S_0 = 0$.

Seja $T = \min\{n | S_n \geq A\}$, ou seja, T é o número de passos até que a barreira A seja atingida. Logo,

$$E(T) = \begin{cases} A + A^2 & \text{se} \quad p = q = 0,5 \\ \\ \frac{A}{q-p} + \frac{p^{A+1}}{q^A(q-p)^2}[1-(q/p)^A] & \text{se} \quad p \neq q \end{cases}$$

No contexto de vigilância, $p = q = 0,5$ representa um processo sob controle; $q > p$ representa um processo fora de controle.

Então, $ARL = E(T|p = q) = A + A^2$ e $A^2 + A - ARL = 0$. Isto implica em

$$A = \frac{-1 \pm \sqrt{1+4ARL}}{2} \approx \sqrt{ARL}\,.$$

Ou seja, a barreira A deve ser igual à raiz quadrada do valor desejado para o ARL. Note que, se o processo está sob controle ($p = q = 0,5$), o tempo médio de espera até o alarme cresce com o quadrado do limite A. Se o processo está fora de controle ($q > p$), o tempo médio de espera até o alarme é linear em A. Assim, adotando o limite $A = \sqrt{ARL}$, o tempo médio de espera até que o alarme soe, dado que o processo está sob controle, é bem maior que este mesmo tempo quando o processo está fora de controle (veja Figura 23). Note que, nos quatro gráficos desta figura, a escala do eixo das abcissas é a mesma, mas a escala do eixo $E(T)$ é bem diferente.

9.2 Sistemas de vigilância para dados pontuais

Nesta seção descrevemos os três sistemas de vigilância que estamos propondo: sistema Binário, sistema Padronizado e sistema Padronizado com Constante Ótima. Considere o processo pontual espaço-temporal $X(t) = \{N(x,y,t)\}$, onde $N(x,y,t)$ é o número de eventos ocorridos na posição (x,y) desde o início do processo em $t = 0$ até um instante arbitrário t.

Seja C_n o cilindro de raio ρ centrado espacialmente no n-ésimo evento, cujo tempo de ocorrência é t_n. A altura deste cilindro é δ, $0 < \delta \leq t_n$. Seja $N(C_n)$ o número de eventos no cilindro C_n. Seja λ_0 a intensidade de eventos, dado que o processo está sob controle. Seja λ_1 a intensidade de eventos, dado que o processo está fora controle ($\lambda_1 > \lambda_0$). Se o processo está sob controle, assumimos $N(C_n) \sim$ Poisson($\lambda_0 \delta \rho^2 \pi$). Caso contrário, $N(C_n) \sim$ Poisson($\lambda_1 \delta \rho^2 \pi$).

O alarme soa quando a estatística de teste, que será definida a seguir, ultrapassa um limite

A estabelecido. O usuário deve especificar valores para δ, ρ, λ_0 e A. A intensidade λ_0 pode ser estimada usando-se dados históricos de um processo sob controle.

Sejam

$$Z_n = \frac{N(C_n) - \lambda_0 \rho^2 \pi \delta}{\sqrt{\lambda_0 \rho^2 \pi \delta}}$$

e

$$I_n = \begin{cases} -1 & \text{se} \quad Z_n \leq 0 \\ +1 & \text{se} \quad Z_n > 0 \end{cases}$$

9.2.1 Sistema Binário

A estatística de teste do sistema Binário é dada por

$$S_n = \max(0, S_{n-1} + I_n),$$

onde $S_0 = 0$. Adotamos o limite $A = \sqrt{ARL}$, conforme a seção 9.1.3. Logo, quando $S_n \geq \sqrt{ARL}$ pela primeira vez, o sistema Binário soa um alarme.

A variável I_n assume apenas os valores 1 e -1, exatamente como na situação descrita na seção 9.1.3. Porém, I_n leva em conta apenas o sinal da diferença entre $N(C_n)$ e seu valor esperado, e desconsidera a informação sobre a magnitude desta diferença.

9.2.2 Sistema Padronizado

A estatística de teste do sistema Padronizado é dada por

$$S_n = \max(0, S_{n-1} + Z_n),$$

onde $S_0 = 0$. Adotamos o limite $A = \sqrt{ARL}$, conforme a seção 9.1.3. Logo, quando $S_n \geq \sqrt{ARL}$ pela primeira vez, o sistema Padronizado soa um alarme.

A variável Z_n, ao contrário da variável I_n do método Binário, leva em conta a magnitude da diferença entre $N(C_n)$ e seu valor esperado. No entanto, Z_n não se comporta conforme a situação descrita na seção 9.1.3, ou seja, Z_n não assume apenas os valores 1 e -1.

9.2.3 Sistema Padronizado com Constante Ótima

Um dos problemas com o método Padronizado é a utilização da variável Z_n na estatística de teste S_n, uma vez que Z_n não se comporta conforme descrito na seção 9.1.3. Já no método

Binário, este problema é contornado com a utilização de I_n ao invés de Z_n. Porém, ao utilizarmos I_n, levamos em conta apenas o sinal de Z_n, e perdemos informação sobre o valor de Z_n.

No método Padronizado com Constante Ótima, tentamos encontrar uma variável Z_n^* que seja próxima de uma variável binária que assume valores 1 e -1 mas que leve em conta o valor de Z_n, e não apenas o sinal. A idéia é encontrar uma constante $c > 0$ tal que

$$Z_n^* = \frac{N(C_n) - \lambda_0 \rho^2 \pi \delta}{c\sqrt{\lambda_0 \rho^2 \pi \delta}}$$

seja o mais próximo possível (em distribuição) da variável I_n. O valor de c que satisfaz esta condição é $c = \sqrt{\pi/2}$ (veja seção 9.2.4).

A estatística de teste do método Padronizado com Constante Ótima é dada por

$$S_n = \max(0, S_{n-1} + Z_n^*),$$

onde $S_0 = 0$. Adotamos o limite $A = \sqrt{ARL}$, conforme a seção 9.1.3. Logo, quando $S_n \geq \sqrt{ARL}$ pela primeira vez, o sistema Padronizado com Constante Ótima soa um alarme.

Vamos nos referir ao método Padronizado com Constante Ótima apenas como método Padronizado com Constante.

9.2.4 Determinação da constante c

Se o processo está sob controle, assumimos $G = N(C_n) \sim \text{Poisson}(\mu)$, onde $\mu = \lambda_0 \delta \rho^2 \pi$. Para μ suficientemente grande, a distribuição de G pode ser aproximada por uma Normal com média e variância μ, ou seja, $G \sim \text{Normal}(\mu, \mu)$ para μ suficientemente grande. Neste caso sejam

$$Y = \frac{G - \mu}{c\sqrt{\mu}} \quad \text{e} \quad W = \begin{cases} -1 & \text{se} \quad G \leq \mu \\ +1 & \text{se} \quad G > \mu \end{cases}$$

Queremos encontrar $c > 0$ tal que $Y \approx W$.

$$
\begin{aligned}
\|Y-W\|^2 &= E(|Y-W|^2) \\
&= E\left(\left|\frac{G-\mu}{c\sqrt{\mu}}-W\right|^2\right) \\
&= \frac{1}{2}E\left(\left|\frac{G-\mu}{c\sqrt{\mu}}-1\right|^2 \mid G>\mu\right)+\frac{1}{2}E\left(\left|\frac{G-\mu}{c\sqrt{\mu}}+1\right|^2 \mid G<\mu\right) \\
&= \frac{1}{2}(E_1+E_2)
\end{aligned}
$$

onde $E_1 = E\left(\left|\frac{G-\mu}{c\sqrt{\mu}}-1\right|^2 \mid G>\mu\right)$ e $E_2 = E\left(\left|\frac{G-\mu}{c\sqrt{\mu}}+1\right|^2 \mid G\leq\mu\right)$.
Mas

$$
\begin{aligned}
E_1+E_2 &= \int_\mu^\infty \left(\frac{x-\mu}{c\sqrt{\mu}}-1\right)^2 f_G(x)dx+\int_{-\infty}^\mu \left(\frac{x-\mu}{c\sqrt{\mu}}+1\right)^2 f_G(x)dx \\
&= \int_\mu^\infty \left[\left(\frac{x-\mu}{c\sqrt{\mu}}\right)^2-2\left(\frac{x-\mu}{c\sqrt{\mu}}\right)+1\right] f_G(x)dx \\
&\quad +\int_{-\infty}^\mu \left[\left(\frac{x-\mu}{c\sqrt{\mu}}\right)^2+2\left(\frac{x-\mu}{c\sqrt{\mu}}\right)+1\right] f_G(x)dx \\
&= \int_\mu^\infty \left(\frac{x-\mu}{c\sqrt{\mu}}\right)^2 f_G(x)dx+\int_{-\infty}^\mu \left(\frac{x-\mu}{c\sqrt{\mu}}\right)^2 f_G(x)dx \\
&\quad +\int_\mu^\infty -2\left(\frac{x-\mu}{c\sqrt{\mu}}\right) f_G(x)dx+\int_{-\infty}^\mu 2\left(\frac{x-\mu}{c\sqrt{\mu}}\right) f_G(x)dx \\
&\quad +\int_\mu^\infty f_G(x)dx+\int_{-\infty}^\mu f_G(x)dx \\
&= \int_{-\infty}^\infty \left(\frac{x-\mu}{c\sqrt{\mu}}\right)^2 f_G(x)dx+\int_{-\infty}^\infty f_G(x)dx \\
&\quad +\frac{2}{c\sqrt{\mu}}\left[\int_\mu^\infty -(x-\mu)f_G(x)dx+\int_{-\infty}^\mu (x-\mu)f_G(x)dx\right] \\
&= \frac{1}{c^2\mu}\int_{-\infty}^\infty (x-\mu)^2 f_G(x)dx+1+\frac{2}{c\sqrt{\mu}}[I_1+I_2],
\end{aligned}
$$

sendo

$$\begin{aligned} I_1 &= \int_{\mu}^{\infty} -(x-\mu) f_G(x) dx \\ &= -\int_0^{\infty} u f_G(u+\mu) du \\ &= -\int_0^{\infty} u f_G(\mu-u) du \end{aligned}$$

e

$$\begin{aligned} I_2 &= \int_{-\infty}^{\mu} (x-\mu) f_G(x) dx \\ &= \int_{\infty}^{0} (-u) f_G(\mu-u)(-du) \\ &= \int_{\infty}^{0} u f_G(\mu-u) du \\ &= -\int_0^{\infty} u f_G(\mu-u) du\,. \end{aligned}$$

Então,

$$\begin{aligned} I_1 + I_2 &= -\int_0^{\infty} u f_G(\mu-u) du - \int_0^{\infty} u f_G(\mu-u) du \\ &= -2\int_0^{\infty} u f_G(\mu-u) du \end{aligned}$$

e

$$
\begin{aligned}
E_1+E_2 &= \frac{1}{c^2\mu}\int_{-\infty}^{\infty}(x-\mu)^2 f_G(x)dx+1 \\
&\quad +\frac{2}{c\sqrt{\mu}}\left[-2\int_0^{\infty} x f_G(\mu-x)dx\right] \\
&= \frac{\mu}{c^2\mu}+1-\frac{4}{c\sqrt{\mu}}\int_0^{\infty} x f_G(\mu-x)dx \\
&= \frac{\mu}{c^2\mu}+1-\frac{4}{c\sqrt{\mu}}\int_0^{\infty}\frac{x}{\sqrt{2\pi\mu}}\exp\left(\frac{-x^2}{2\mu}\right)dx \\
&= \frac{1}{c^2}+1-\frac{4}{c\sqrt{\mu}}\frac{\sqrt{\mu}}{2}\frac{\sqrt{2}}{\sqrt{\pi}} \\
&= \frac{1}{c^2}+1-\frac{2\sqrt{2}}{c\sqrt{\pi}}.
\end{aligned}
$$

Logo,

$$
\|Y-W\|^2 = f(c) = \frac{1}{2}\left\{\frac{1}{c^2}+1-\frac{2\sqrt{2}}{c\sqrt{\pi}}\right\}.
$$

Portanto, queremos encontrar $c>0$ que minimiza $f(c)$. Mas

$$
f'(c) = \frac{1}{2}\left\{-2c^{-3}+\frac{2\sqrt{2}}{\sqrt{\pi}}c^{-2}\right\}.
$$

Logo,

$$
f'(c)=0 \Rightarrow \frac{-\sqrt{\pi}+c\sqrt{2}}{c^3\sqrt{\pi}}=0.
$$

Então

$$
-\sqrt{\pi}+c\sqrt{2}=0 \Rightarrow c=\sqrt{\pi/2}.
$$

Além disso, $-\sqrt{\pi}+c\sqrt{2}<0 \Leftrightarrow c\sqrt{2}<\sqrt{\pi} \Leftrightarrow c<\sqrt{\pi/2}$. Ou seja, $f(c)$ decresce no intervalo $[0,\sqrt{\pi/2}]$ e cresce no intervalo $[\sqrt{\pi/2},\infty]$. Logo $c=\sqrt{\pi/2}$ é mínimo.

9.2.5 Esperança e Variância de Z_n e Z_n^*

Se o processo está sob controle ($\tau=\infty$), $N(C_n)\sim$ Poisson($\lambda_0\delta\rho^2\pi$). Então,

$$
E(Z_n|\tau=\infty)=0 \text{ e } Var(Z_n|\tau=\infty)=1.
$$

Se o processo está fora de controle, $N(C_n) \sim \text{Poisson}(\lambda_1 \delta \rho^2 \pi)$. Então,

$$E(Z_n|\tau < \infty) = \frac{\rho(\lambda_1 - \lambda_0)\sqrt{\pi\delta}}{\sqrt{\lambda_0}} > 0 = E(Z_n|\tau = \infty)$$

e

$$Var(Z_n|\tau < \infty) = \frac{\lambda_1}{\lambda_0} > 1 = E(Z_n|\tau = \infty).$$

Em relação à variável Z_n^* utilizada no sistema Padronizado com Constante,

$$E(Z_n^*|\tau = \infty) = 0 = E(Z_n|\tau = \infty)$$

e

$$Var(Z_n^*|\tau = \infty) = 2/\pi < Var(Z_n|\tau = \infty) = 1.$$

Além disso,

$$E(Z_n^*|\tau < \infty) = \sqrt{\frac{2}{\pi}} \frac{\rho(\lambda_1 - \lambda_0)\sqrt{\pi\delta}}{\sqrt{\lambda_0}} < \frac{\rho(\lambda_1 - \lambda_0)\sqrt{\pi\delta}}{\sqrt{\lambda_0}} = E(Z_n|\tau < \infty)$$

e

$$Var(Z_n^*|\tau < \infty) = \frac{\lambda_1}{\lambda_0}\frac{2}{\pi} < \frac{\lambda_1}{\lambda_0} = Var(Z_n|\tau < \infty).$$

Dado $\tau < \infty$, $E(Z_n)$, $Var(Z_n)$, $E(Z_n^*)$ e $Var(Z_n^*)$ aumentam à medida que aumenta a diferença entre λ_0 e λ_1.

9.3 Estudo de simulação

Foram feitas simulações dos três sistemas apresentados na seção anterior, denotados aqui por sistemas alternativos, e do sistema proposto em Assunção and Correa (2009). As simulações foram feitas tanto para processos sob controle quanto para processos fora de controle. Vamos nos referir ao sistema proposo em Assunção and Correa (2009) como sistema Spatial Shiryayev-Roberts (*SSR*).

Em todas as simulações foram gerados eventos no plano espacial de tamanho 5 por 5, onde tanto o eixo *x* quanto o eixo *y* variaram de 0 a 5. O tempo de ocorrência destes eventos variou de 0 até um valor inteiro grande o suficiente para que, no caso de um processo fora de controle, o alarme soasse em todas as simulações. A intensidade de eventos quando o processo está sob controle, λ_0, foi de 4 eventos por unidade de volume.

No caso de um processo fora de controle, além dos eventos gerados considerando-se $\lambda_0 = 4$ eventos por unidade de volume, geramos também, a partir do tempo $m = 5$, um determinado número de eventos dentro de um conglomerado artificial. Este conglomerado foi representado por um quadrado espacial de lado 1, cujo centro tem coordenadas $(x; y) = (2,5; 2,5)$. λ_1 é a intensidade de eventos dentro deste conglomerado ($\lambda_1 > \lambda_0$). Consideramos dois valores diferentes para a intensidade λ_1: 6 eventos por unidade de volume e 12 eventos por unidade de volume. Como o conglomerado artificial começa no tempo $m = 5$, foram gerados aproximadamente $5 \times 5 \times m \times \lambda_0 = 500$ eventos antes do processo passar ao estado fora de controle. Em todas as simulações, o valor especificado para o raio espacial ρ foi igual ao seu valor verdadeiro (0,5). O verdadeiro valor do parâmetro ε do método *SSR* é $(\lambda_1/\lambda_0) - 1$.

Nos sistemas alternativos, a altura de cada cilindro C_n é dada por $t_n - \delta$. Por este motivo, foram gerados $5 \times 5 \times \delta \times \lambda_0$ eventos no plano de tamanho 5 por 5 com tempos entre $-\delta$ e 0. Estes eventos foram usados apenas para que fosse possível calcular a estatística de teste S_n para eventos com tempo $t_n < \delta$.

9.3.1 Resultados preliminares

A Tabela 3 apresenta estatísticas referentes a 100 simulações de cada um dos sistemas (sistemas alternativos e sistema *SSR*) para processos sob controle. Nos sistemas alternativos utilizou-se $\delta = 3$. Foram analisadas duas possibilidades para o parâmetro ε do sistema *SSR*: $\varepsilon = 0,5$ e $\varepsilon = 2,0$. No método *SSR* o limite adotado foi $A = ARL$; nos métodos alternativos o limite adotado foi $A = \sqrt{ARL}$. O valor do *ARL* foi fixado em 650 eventos, ou seja, espera-se em média 1 alarme falso a cada 650 eventos. A coluna **% Alarmes** mostra o proporção de alarmes. A coluna seguinte mostra a média e o desvio padrão do número de eventos observados até o alarme. Neste caso, todos os alarmes são falsos, pois o processo está sob controle.

A Tabela 3 mostra que, em todos os sistemas, o alarme soa falsamente em praticamente todas as simulações. No método *SSR*, o número médio de eventos até o alarme é maior que o valor especificado para o *ARL* (650). Este comportamente já era esperado, uma vez que a adoção do limite $A = ARL$ neste sistema gera um procedimento conservador (veja Assunção and Correa, 2009). A Figura 24 mostra a distribuição do número de eventos até o alarme no método *SSR*, dado que o processo está sob controle. Esta distribuição parece ser simétrica.

Método	% Alarmes	Número de eventos até o alarme	
		Média	Desvio padrão
SSR $\varepsilon = 0,5$	100	998,00	111,48
SSR $\varepsilon = 2,0$	98	1150,09	401,19
Binário	100	507,73	359,20
Padronizado	100	314,52	245,95
Padronizado com Constante	100	408,07	275,57

Tabela 3: Estatísticas referentes a 100 simulações de um processo sob controle, onde $\delta = 3$ nos sistemas alternativos e $ARL = 650$ eventos. No método *SSR*, o limite $A = ARL$. Nos métodos alternativos, $A = \sqrt{ARL}$. A coluna **% Alarmes** mostra o proporção de alarmes.

A Tabela 3 mostra ainda que, nos três métodos alternativos, o número médio de eventos até o alarme, é inferior ao valor nominal 650. Esperaríamos em média 650 eventos até o alarme na situação do passeio aleatório descrito na seção 9.1.3. No entanto, nem todos os sistemas alternativos reproduzem exatamente esta situação. Nos métodos Padronizado e Padronizado com Constante, por exemplo, Z_n e Z_n^* não assumem apenas os valores 1 e -1. Além disso, a distribuição do número de eventos observados até o alarme nos sistemas alternativos é extremamente assimétrica à direita, como mostrado na Figura 25. Assim, a média é uma medida pouco representativa desta distribuição. Simulações do passeio aleatório com barreiras conforme descrito na seção 9.1.3, fora do contexto espaço-tempo, mostram que a distribuição do tempo até que a barreira seja cruzada pela primeira vez é extremamente assimétrica à direita. No caso dos sistemas alternativos, a consequência desta assimetria é que a probabilidade do alarme soar falsamente é muito alta.

A Tabela 4 ilustra a situação descrita acima. Ela apresenta as proporções de alarmes falsos e motivados em 100 simulações de cada sistema, considerando-se processos fora de controle. Nos sistemas alternativos utilizou-se $\delta = 3$. Para o sistema *SSR* utilizou-se $\varepsilon = (\lambda_1/\lambda_0) - 1$. Foram analisadas duas possibilidades para o valor de λ_1: $\lambda_1 = 6$ e $\lambda_1 = 12$. No método *SSR* o limite adotado foi $A = ARL$; nos métodos alternativos o limite adotado foi $A = \sqrt{ARL}$. O valor do *ARL* foi fixado em 650 eventos.

9.3.2 Modelo exponencial

Como dito anteriormente, nos métodos alternativos, a média do tempo de espera até que o alarme soe falsamente é uma medida pouco representativa da distribuição deste tempo.

Método	$\lambda_1 = 6$ - % Alarmes		$\lambda_1 = 12$ - % Alarmes	
	Falsos	**Motivados**	**Falsos**	**Motivados**
SSR	0	100	7	92
Binário	54	46	62	38
Padronizado	80	20	85	15
Padronizado com Constante	70	30	74	26

Tabela 4: Proporção de alarmes falsos e motivados em 100 simulações utilizando-se processos fora de controle, onde $\delta = 3$ nos métodos alternativos e $ARL = 650$ eventos. No método *SSR*, o limite $A = ARL$. Nos métodos alternativos, $A = \sqrt{ARL}$.

Assumindo que o tempo até o alarme soar falsamente, *RL*, tem distribuição exponencial com parâmetro λ, podemos encontrar λ tal que a probabilidade do alarme soar falsamente seja igual a α. Como o conglomerado artificial começa no tempo $m = 5$, se *RL* for menor que 500 então o alarme é falso. Na prática, o instante de tempo em que o conglomerado começa é desconhecido.

Seja $\alpha = 0,1$ e $RL \sim \text{Exp}(\lambda)$. Então

$$P(RL \leq 500) = 1 - \exp(-\lambda * 500) = 0,1 \Rightarrow \lambda = 0,00021\,.$$

Se $\lambda = 0,00021$, $E(RL) = ARL = 1/\lambda = 4745$. Neste caso a barreira seria $A = \sqrt{ARL} \approx 69$.

As Tabelas 5 e 6 mostram os resultados de simulações para processos fora de controle, onde $\delta = 3$ nos sistemas alternativos. Foram feitas 50 simulações de cada um dos sistemas alternativos e 100 simulações do sistema *SSR*. Os limites adotados nos métodos alternativos e no método *SSR* foram $A = 69$ e $A = 650$, respectivamente. Na Tabela 5, $\varepsilon = 0,5$ no sistema *SSR* e $\lambda_1 = 6$. Na Tabela 6, $\varepsilon = 2,0$ no sistema *SSR* e $\lambda_1 = 12$. O limite $A = 650$ adotado no método *SSR* é o valor desejado para o *ARL*; o limite $A = 69$ adotado nos métodos alternativos é o valor tal que o alarme deveria soar falsamente em aproximadamente 10% das simulações. Como estes dois limites não têm o mesmo significado, os resultados do método *SSR* não são diretamente comparáveis aos resultados dos métodos alternativos. Nestas tabelas, por questões práticas, o método Padronizado com Constante está representado apenas por Constante. A coluna **% Alarme** mostra: a proporção de simulações em que o alarme não soa (**Sem**), a proporção de simulações em que o alarme soa falsamente (**Falsos**), e a proporção de simulações em que o alarme soa motivadamente (**Motivados**). A proporção de simulações em que o alarme não soa é zero em todos os casos. A coluna **Alarmes Motivados** mostra

a média e o desvio padrão (DP) do número de eventos até que o alarme soe motivadamente. A coluna **Delay** mostra a média e o desvio padrão (DP) da estatística CED^* para os alarmes motivados. A estatística CED^* é aquela definida em Assunção and Correa (2009). CED^* é o número médio de eventos, pertencentes ao conglomerado espaço-tempo, que ocorrem entre o instante τ e o instante T_A em que o alarme soa, dado $T_A > \tau$. A coluna **t.inicio** mostra a média e o desvio padrão (DP) para a estimativa do tempo de início do conglomerado nos casos em que o alarme é motivado. A coluna **% Dentro** mostra a proporção de simulações em que a estimativa da localização espacial do conglomerado está dentro do quadrado que representa o conglomerado artificial. Nos sistemas alternativos, esta estimativa é dada pelas coordenadas espaciais do próprio evento do alarme. No método *SSR* esta estimativa é dada pelas coordenadas espaciais do evento com maior contribuição para a estatística de teste (veja Assunção and Correa, 2009).

As Tabelas 5 e 6 mostram que, quando a diferença entre λ_0 e λ_1 é pequena ($\lambda_1 = 6$ e $\lambda_0 = 4$, na Tabela 5), a proporção de simulações em que a estimativa da localização espacial do conglomerado está dentro do quadrado que representa o conglomerado artificial é baixa em todos os métodos. Mesmo quando a diferença entre λ_0 e λ_1 é maior ($\lambda_1 = 12$ e $\lambda_0 = 4$, na Tabela 6), estas estimativas não são tão boas quanto gostaríamos, principalmente nos sistemas alternativos. Além disso, nos sistemas Padronizado e Padronizado com Constante, a proporção de alarmes falsos é maior que o valor nominal 10%. No método Binário, esta proporção está bem próxima de 10%, mas o número médio de eventos até o alarme é maior, se comparado ao sistema *SSR*. A proporção de alarmes falsos nos métodos alternativos foi avaliada em um número maior de simulações, uma vez que os resultados das Tabelas 5 e 6 são baseados em apenas 50 simulações.

As Tabelas 7 e 8 mostram os resultados de 1000 simulações para cada um dos sistemas alternativos considerando-se processos fora de controle ($\delta = 3$). O limite adotado foi $A = 69$, valor tal que o alarme deveria soar falsamente em aproximadamente 10% das simulações. Nestas tabelas, por questões práticas, o método Padronizado com Constante está representado apenas por Constante. A coluna **% Alarme** mostra: a proporção de simulações em que o alarme não soa (**Sem**), a proporção de simulações em que o alarme soa falsamente (**Falsos**) e a proporção de simulações em que o alarme soa motivadamente (**Motivados**). A coluna **Alarmes Motivados** mostra a média e o desvio padrão (DP) do número de eventos até que o alarme soe para os alarme motivados. Estas duas tabelas mostram que, nos métodos Padronizado e Padronizado com Constante, a taxa de alarmes falsos é realmente maior que o valor nominal 10%.

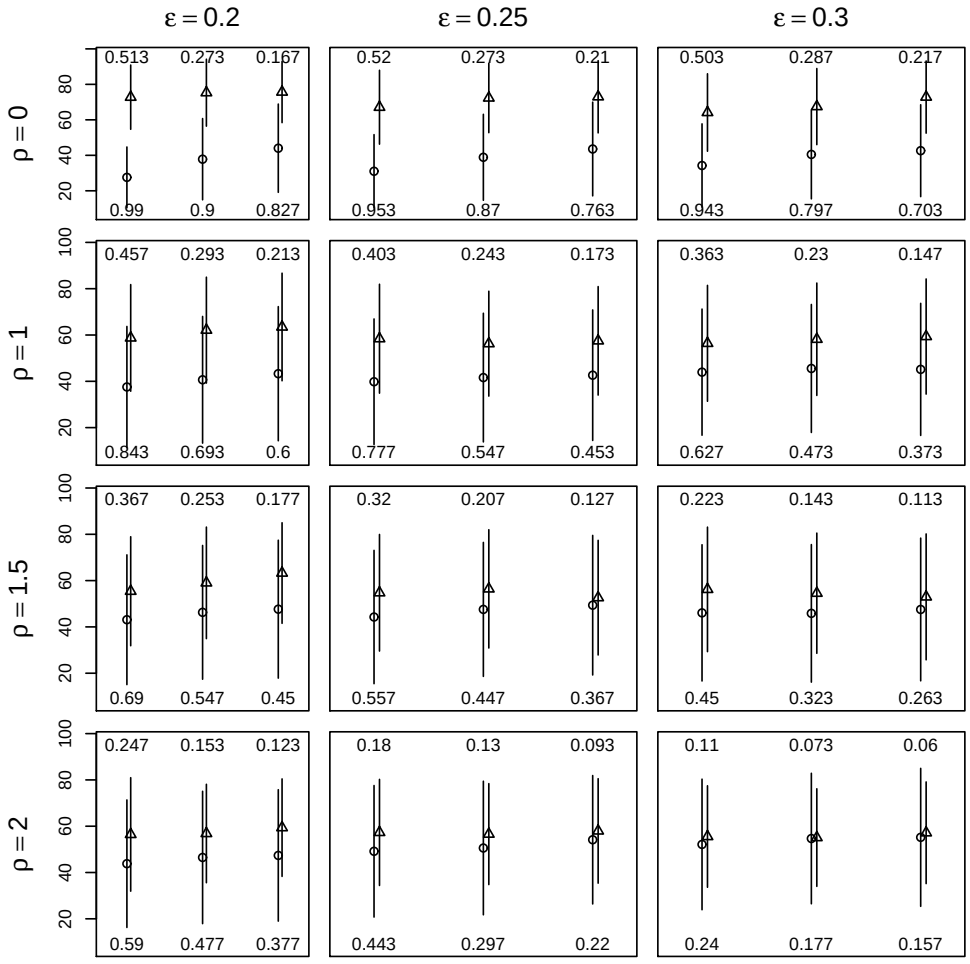

Figura 10: Summary statistics for the number of time periods until the alarm sounds off in 300 simulations, using an under control process. Rows $1,2,3,4$ correspond to $\rho = 0.0, 1.0, 1.5, 2.0$, respectively. Columns $1,2,3,4$ corresponds to $\varepsilon = 0.20, 0.25, 0.30$, respectively. The horizontal axis represents the threshold limit: $20, 30, 40$ time periods. The vertical axis represents the number of time periods until the alarm sounds off. The circles and the triangles represents the average value using $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$, respectively. The segments in each symbol represents one standard deviation above and below the mean. The numbers at the bottom and at the top correspond to the proportion of alarms when using $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$, respectively.

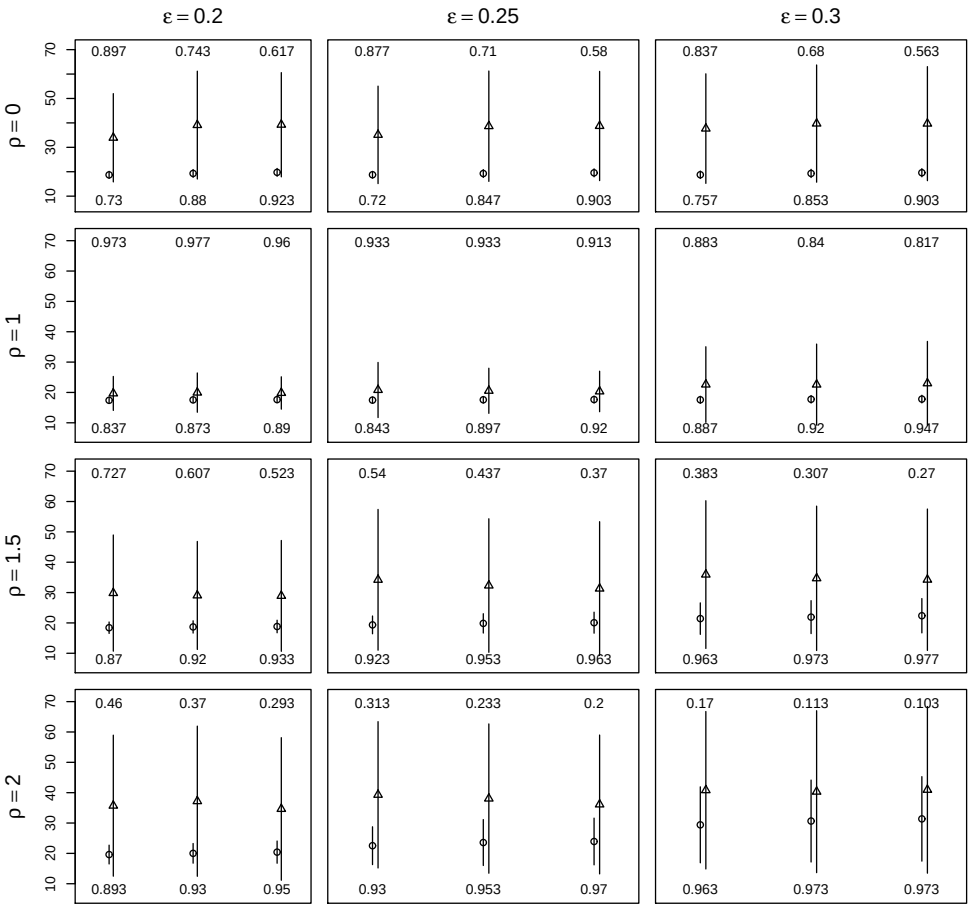

Figura 11: Summary statistics for the number of time periods until a motivated alarm in 300 simulations, using an out of control process. Rows $1,2,3,4$ correspond to $\rho = 0.0, 1.0, 1.5, 2.0$, respectively. Columns $1,2,3,4$ corresponds to $\varepsilon = 0.20, 0.25, 0.30$, respectively. The horizontal axis represents the threshold limit: $20, 30, 40$ time periods. The vertical axis represents the number of time periods until the alarm sounds off. The circles and the triangles represents the average value using $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$, respectively. The segments in each symbol represents one standard deviation above and below the mean. The numbers at the bottom and at the top correspond to the proportion of motivated alarms when using $\mu_{C_{j,k}}$ and $\widehat{\mu}_{C_{j,k}}$, respectively.

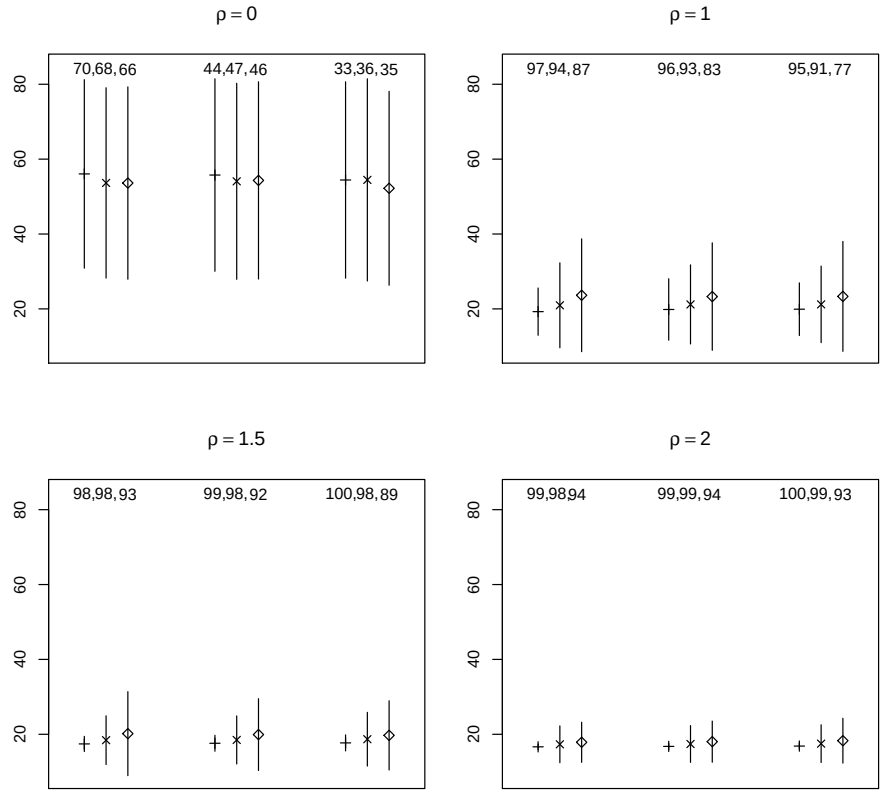

Figura 12: Summary statistics for the observed number of time periods until a motivated alarm in 300 simulations, considering an out of control process. Each plot corresponds to one value for ρ. In all plots, the horizontal axis represents the threshold limit: $20, 30, 40$ time periods. The vertical axis represents the observed number of time periods until the motivated alarm. The traces, crosses, and lozenges represent the average value for $\varepsilon = 0.20, 0.25, 0.30$, respectively. The segments in each symbol represents one standard deviation above and below the mean. The numbers at the top corresponds to the proportion of motivated alarms. For each threshold limit, the first, second and third numbers are related to $\varepsilon = 0.20, 0.25, 0.30$, respectively.

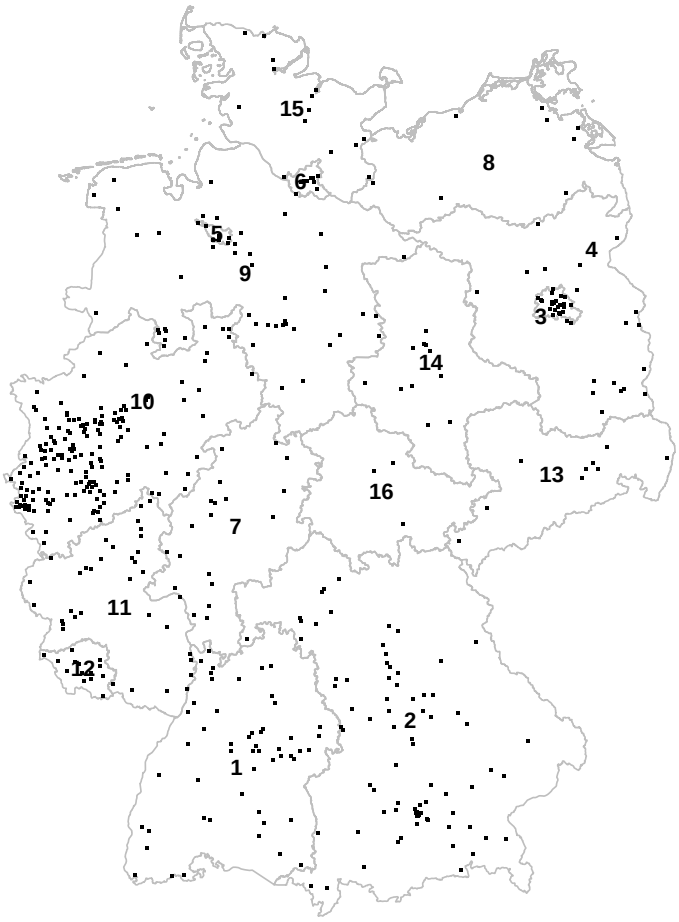

Figura 13: Meningococcal cases in Germany between 2002 and 2008. The federated states are: 1=Baden-Württemberg, 2=Bayern, 3=Berlin, 4=Brandenburg, 5=Bremen, 6=Hamburg, 7=Hessen, 8=Mecklenburg-Vorpommern, 9=Niedersachsen, 10=Nordrhein-Westfalen, 11=Rheinland-Pfalz, 12=Saarland, 13=Sachsen, 14=Sachsen-Anhalt, 15=Schleswig-Holstein, 16=Thüringen.

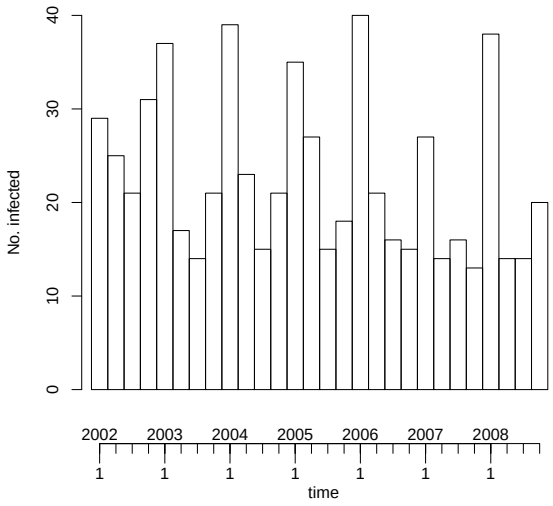

Figura 14: Number of cases per year/quarter.

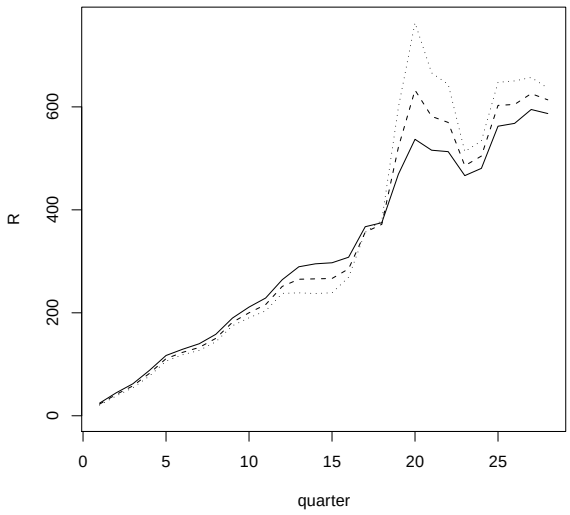

Figura 15: Monitoring statistic R at each quarter. The solid, dashed and dotted lines corresponds to $\varepsilon = 0.20, 0.25, 0.30$, respectively.

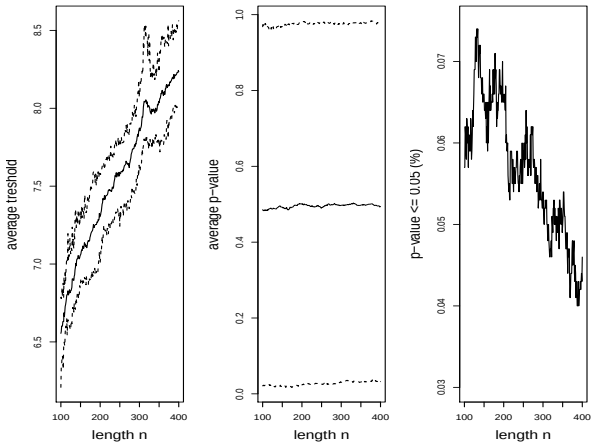

Figura 16: Results for the retrospective scan - situation (a). The first graph shows in solid line the average critical value for the scan statistic together with the pointwise 95% confidence bands $L(n)$ and $U(n)$ in dashed lines. The second graph shows in solid line the average p-value together with the 95% confidence bands for the observed p-value at each time series length n (in dashed lines). The third graph shows the proportion of simulations in which the p-value is at most α.

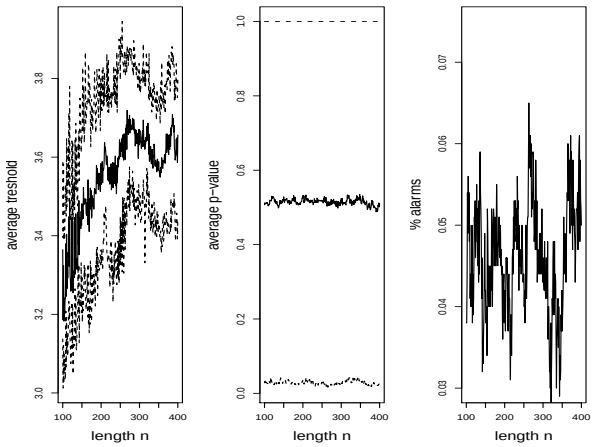

Figura 17: Results for situation (b) when the prospective scan has no adjustment for earlier analysis. In all the three graphs, the lines (solid and dashed) have the same definition as in Figure 16.

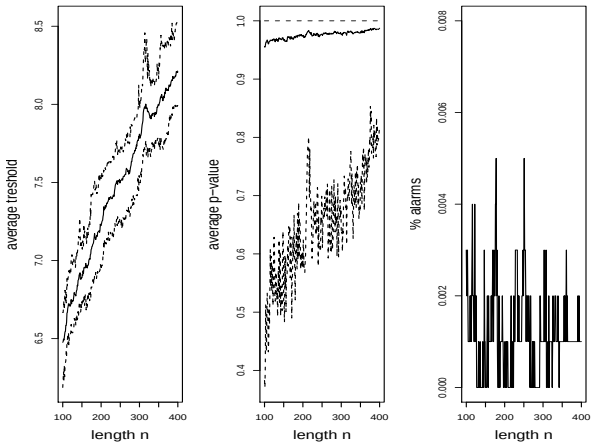

Figura 18: Results for situation (c) when the prospective scan adjusts for all previous analysis. Definitions of lines and plots are the same as in Figure 16.

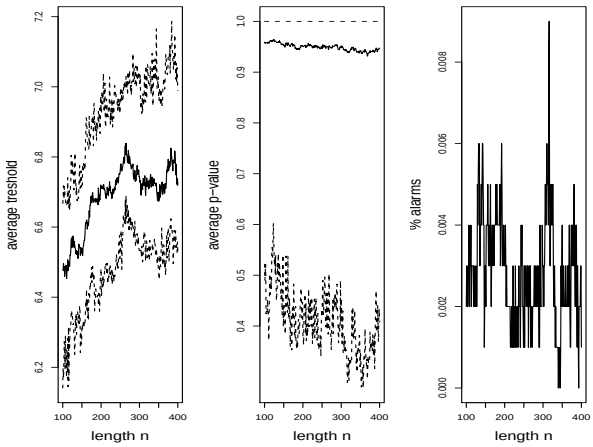

Figura 19: Results for situation (c) when the prospective scan adjusts for the last 100 analysis. Definitions of lines and plots are the same as in Figure 16.

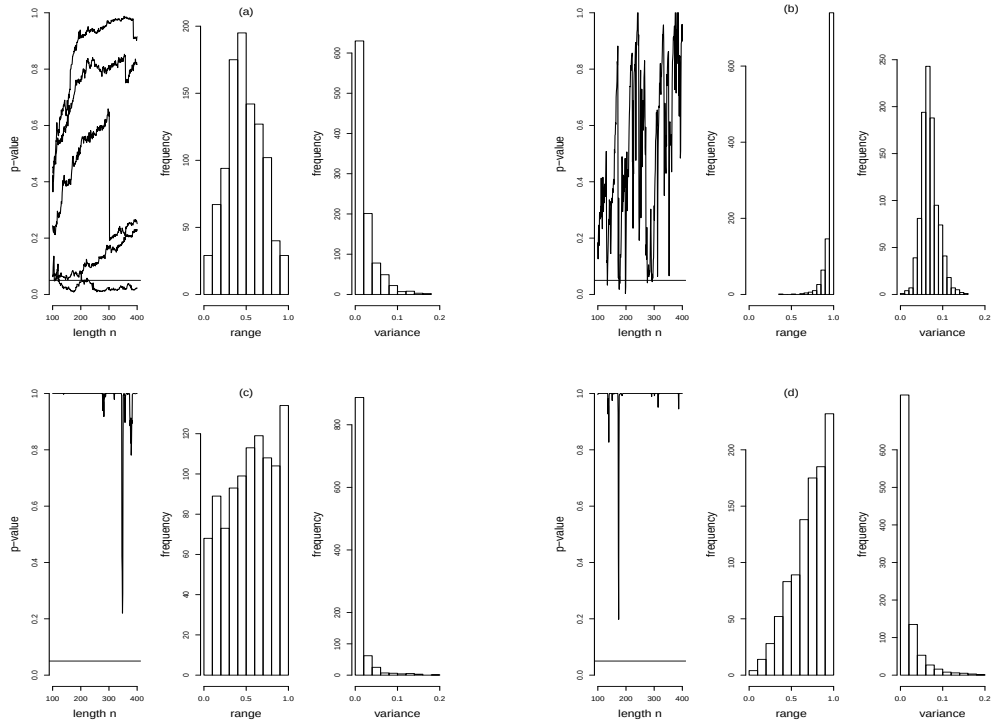

Figura 20: In each row, the first plot illustrates the typical behavior of the p-value time series p_t. The second and the third plots show the distribution of the range and the variance for all 1000 time series p_t, respectively.

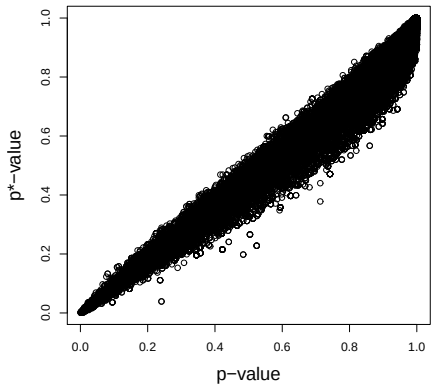

Figura 21: SaTScan p-value versus the p*-value for the prospective scan adjusting for all previous analysis.

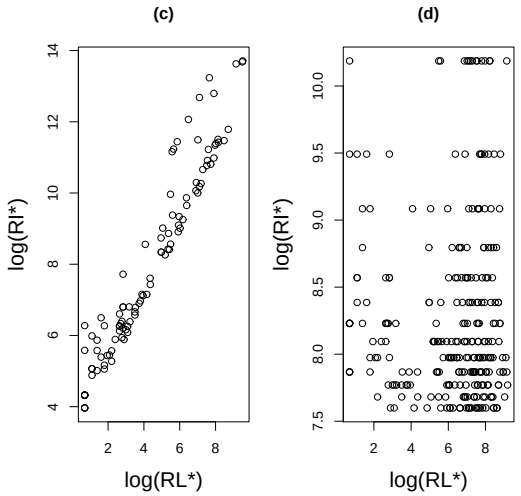

Figura 22: RL^* versus RI^* for the prospective situations (c) and (d). The axes are in the logarithm scale. The proportion of alarms is 0.38 and 1.00 for situations (c) and (d), respectively.

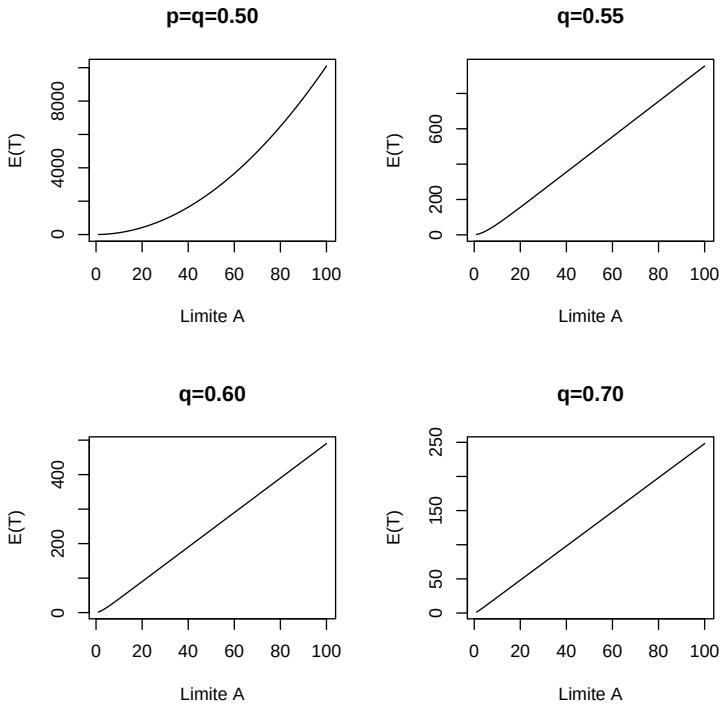

Figura 23: Tempo médio até o alarme *versus* limite *A*.

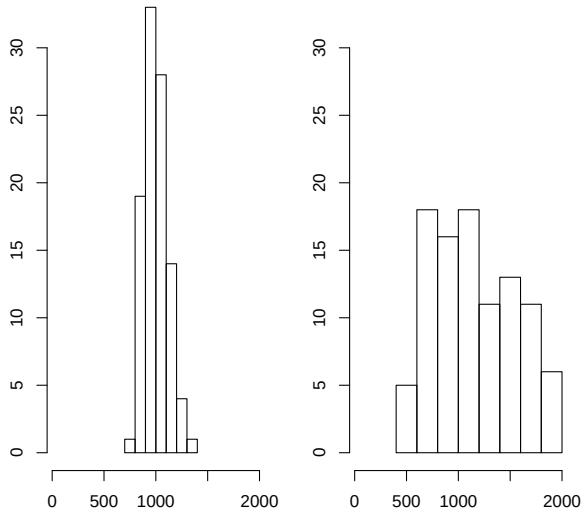

Figura 24: Distribuição do número de eventos até o alarme no método *SSR*, dado que o processo está sob controle. No gráfico da esquerda $\varepsilon = 0,5$; no gráfico da direita $\varepsilon = 2,0$. O limite $A = ARL = 650$ eventos.

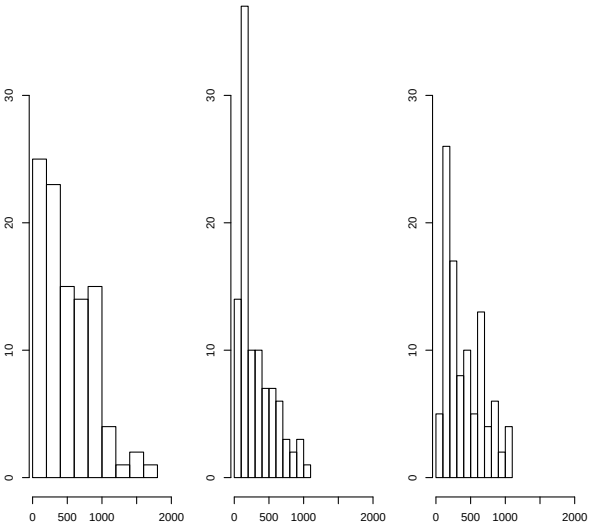

Figura 25: Distribuição do número de eventos até o alarme nos métodos alternativos, dado que o processo está sob controle. Da esquerda para a direita: método Binário, método Padronizado e método Padronizado com Constante. O limite $A = \sqrt{ARL} = \sqrt{650}$ eventos e $\delta = 3$.

Método	% Alarme			Alarmes Motivados		Delay		t.inicio		% Dentro
	Sem	Falsos	Motivados	Média	DP	Média	DP	Média	DP	
SSR	0	0	100	998,20	121,17	29,85	8,39	7,87	1,58	19,00
Binário	0	12	88	1294,02	455,25	48,43	26,01	9,78	3,52	9,09
Padronizado	0	28	72	845,11	188,81	21,72	12,09	5,44	1,89	8,33
Constante	0	18	82	992,88	344,61	30,76	19,22	6,88	3,52	19,51

Tabela 5: Estatísticas baseadas em 50 simulações dos métodos alternativos e 100 simulações do método *SSR* para processos fora de controle, onde $\varepsilon = 0,5$ no método *SSR*, $\delta = 3$ nos métodos alternativos, $ARL = 650$ e $\lambda_1 = 6$. No método *SSR*, o limite $A = ARL = 650$ eventos. Nos métodos alternativos, $A = 69$ eventos.

Método	% Alarme			Alarmes Motivados		Delay		t.inicio		% Dentro
	Sem	Falsos	Motivados	Média	DP	Média	DP	Média	DP	
SSR	0	3	97	856,00	319,72	39,35	34,82	7,64	2,94	59,79
Binário	0	6	94	951,21	257,71	51,60	29,75	6,21	2,53	25,53
Padronizado	0	24	76	715,21	89,22	25,13	11,49	4,11	0,99	55,26
Constante	0	14	86	740,93	100,05	27,53	12,48	4,29	1,09	44,19

Tabela 6: Estatísticas baseadas 50 simulações dos métodos alternativos e 100 simulações do método *SSR* para processos fora de controle, onde $\varepsilon = 2,0$ no método *SSR*, $\delta = 3$ nos métodos alternativos, $ARL = 650$ e $\lambda_1 = 12$ No método *SSR*, o limite $A = ARL = 650$ eventos. Nos métodos alternativos, $A = 69$ eventos.

Método	% Alarme			Alarmes Motivados	
	Sem	Falsos	Motivados	Média	DP
Binário	0	10	90	1381,56	619,63
Padronizado	0	34	66	893,02	305,93
Constante	0	24,2	75,8	980,75	342,52

Tabela 7: Estatística baseadas em 1000 simulações de processos fora de controle, onde $\delta = 3$ e $\lambda_1 = 6$. O limite $A = 69$ eventos.

Método	% Alarme			Alarmes Motivados	
	Sem	Falsos	Motivados	Média	DP
Binário	0	9,7	90	940,59	258,41
Padronizado	0	32,9	66	690,55	95,85
Constante	0	20,7	75,8	714,64	103,66

Tabela 8: Estatística baseadas em 1000 simulações de processos fora de controle, onde $\delta = 3$ e $\lambda_1 = 12$. O limite $A = 69$ eventos.

9.4 Considerações finais

Os sistemas alternativos não se mostraram eficientes no que diz respeito ao controle da taxa de alarmes falsos. Adotando-se o limite $A = \sqrt{ARL}$, a proporção de alarmes falsos é extremamente alta. Mesmo adotando-se o modelo exponencial, esta taxa ainda é maior que o valor nominal nos sistemas Padronizado e Padronizado com Constante. No sistema Binário, a taxa de alarmes falsos parece estar controlada com o uso do modelo exponencial. No entanto, o tempo de espera por um alarme motivado é maior, se comparado ao sistema *SSR*. Além disso, os sistemas alternativos se mostraram ruins no que diz respeito a estimar a localização espacial do conglomerado.

10 Considerações Finais

Os sistemas de vigilância propostos nestas tese atendem aos objetivos propostos no sentido que não exigem modelagem estatística complexa. Assim, este sistemas podem ser facilmente usados por usuários sem conhecimento estatístico, como por exemplo funcionários de agências de saúde pública.

Os sistemas baseados na estatística de Shiryayev-Roberts (Kenett and Pollak, 1996) se mostraram eficientes no sentido de controlar a taxa de alarmes falsos e detectar rapidamente um conglomerado emergente. No entanto, a velocidade desta detecção depende dos valores adotados para os parâmetros do sistema. Os sistemas alternativos se mostraram ainda pouco eficientes neste mesmo sentido, de forma que são necessarias melhorias neste sistemas.

Dentre as várias possibilidades de trabalhos futuros relacionados ao sistemas propostos, destacamos a adaptação dos sistemas para detecção de conglomerados com outra forma geométrica, diferente da forma cilíndrica.

A análise do sistema de vigilância proposto por Kulldorff (2001) mostra que este sistema também requer melhorias, e aponta os pontos que devem ser trabalhados. Esperamos que esta análise possa motivar trabalhos futuros no sentido de solucionar o problema do uso da estatística scan no contexto prospectivo.

Apêndice

Let $\mathbf{Y}$ be a matrix $H \times m$ with terms y_{it}. Let I_i be an indicator function such that $I_i = 1$ if $i \in S_j$ and $I_i = 0$ if $i \notin S_j$. Therefore,

$$E_{\tau=\infty}(\Lambda_{j,k,m+1}|\mathbf{Y})$$

$$= E_{\tau=\infty}\left(\prod_{i\in S_j}\prod_{t=k}^{m+1}(1+\varepsilon)^{y_{it}}\exp(-\varepsilon\mu_{it})|\mathbf{Y}\right)$$

$$= E_{\tau=\infty}\left(\prod_{i\in S_j}\prod_{t=k}^{m}(1+\varepsilon)^{y_{it}}\exp(-\varepsilon\mu_{it})\prod_{i\in S_j}(1+\varepsilon)^{y_{i(m+1)}}\exp(-\varepsilon\mu_{i(m+1)})|\mathbf{Y}\right)$$

$$= E_{\tau=\infty}\left(\Lambda_{j,k,m}\prod_{i\in S_j}(1+\varepsilon)^{y_{i(m+1)}}\exp(-\varepsilon\mu_{i(m+1)})|\mathbf{Y}\right)$$

$$= \Lambda_{j,k,m}E_{\tau=\infty}\left(\frac{\prod_{1\le i\le H}\left(\mu_{i(m+1)}(1+\varepsilon I_i)\right)^{y_{i(m+1)}}\left(y_{i(m+1)}!\right)^{-1}\exp\left(-\mu_{i(m+1)}(1+\varepsilon I_i)\right)}{\prod_{1\le i\le H}\left(\mu_{i(m+1)}\right)^{y_{i(m+1)}}\left(y_{i(m+1)}!\right)^{-1}\exp\left(-\mu_{i(m+1)}\right)}|\mathbf{Y}\right)$$

$$= \Lambda_{j,k,m}\sum_{y_{i(m+1)}=0}^{\infty}\frac{\prod_{1\le i\le H}\left(\mu_{i(m+1)}(1+\varepsilon I_i)\right)^{y_{i(m+1)}}\left(y_{i(m+1)}!\right)^{-1}\exp\left(-\mu_{i(m+1)}(1+\varepsilon I_i)\right)}{\prod_{1\le i\le H}\left(\mu_{i(m+1)}\right)^{y_{i(m+1)}}\left(y_{i(m+1)}!\right)^{-1}\exp\left(\mu_{i(m+1)}\right)}$$

$$\prod_{1\le i\le H}\left(\mu_{i(m+1)}\right)^{y_{i(m+1)}}\left(y_{i(m+1)}!\right)^{-1}\exp\left(-\mu_{i(m+1)}\right)$$

$$= \Lambda_{j,k,m}\sum_{y_{i(m+1)}=0}^{\infty}\prod_{1\le i\le H}(\mu_{i(m+1)}(1+\varepsilon I_i))^{y_{i(m+1)}}\left(y_{i(m+1)}!\right)^{-1}\exp(-\mu_{i(m+1)}(1+\varepsilon I_i))$$

$$= \Lambda_{j,k,m}\sum_{y_{1(m+1)}=0}^{\infty}\ldots\sum_{y_{H(m+1)}=0}^{\infty}(\mu_{1(m+1)}(1+\varepsilon I_1))^{y_{1(m+1)}}\left(y_{1(m+1)}!\right)^{-1}\exp(-\mu_{1(m+1)}(1+\varepsilon I_1))$$

$$\ldots(\mu_{H(m+1)}(1+\varepsilon I_H))^{y_{H(m+1)}}\left(y_{H(m+1)}!\right)^{-1}\exp(-\mu_{H(m+1)}(1+\varepsilon I_H))$$

$$= \Lambda_{j,k,m}\times 1\times\ldots\times 1$$

$$= \Lambda_{j,k,m}.$$

Referências

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

Assunção, R., Costa, M., Tavares, A., and Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25:723–742.

Assunção, R. and Maia, A. (2007). A note on testing separability in spatial-temporal marked point processes. *Biometrics*, 63:290–294.

Assunção, R., Tavares, A., Correa, T., and Kulldorff, M. (2007). Space-time cluster identification in point processes. *The Canadian Journal of Statistics*, 35:9–25.

Buckeridge, D., Burkom, H., Campbell, M., Hogane, W., and Moore, A. (2005). Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, 38:99–113.

Buehler, J., Hopkins, R., Overhage, J., Sosin, D., Tong, V., and Group, C. W. (2004). Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the cdc working group. *Morbidity and Mortality Weekly Report*, 7:1–11.

Corberán-Vallet, A. and Lawson, A. B. (2011). Conditional predictive inference for online surveillance of spatial disease incidence. *Statistics in Medicine*, 30:3095–3116.

Correa, T., Assunção, R., and Costa, M. (submitted in October2011). A close look on prospective surveillance using a scan statistic. *Biometrics*.

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

Duczmal, L. and Assunção, R. (2004). A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, 45:269–286.

Duczmal, L., Kulldorff, M., and Huang, L. (2006). Evaluation of the spatial scan statistics for irregular shaped clusters. *Journal of Computational and Graphical Statistics*, 15:428–442.

Farrington, C., Andrews, N., Beale, A., and Catchpole, M. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society, Series A*, 159:547–563.

Frisén, M. (2003). Statistical surveillance. optimality and methods. *International Statistical Review*, 71:403–Ö434.

Frisén, M., Andersson, E., and Schiöler, L. (2011). Sufficient reduction in multivariate surveillance. *Communications in Statistics - Theory and Methods*, 40:1821–1838.

Hardy, A. (2001). Methods of outbreak investigation in the "era of bacteriology"1880-1920. *Social and Preventive Medicine*, 46:355–360.

Henderson, D. (1999). The looming threat of bioterrorism. *Science*, 283:1279–1282.

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

Höhle, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biometrical Journal*, 51:961–978.

Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics and Data Analysis*, 52:4357–4Ö´368.

Höhle, M., Paul, M., and Held, L. (2009). Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health. *Preventive Veterinary Medicine*, 91:2–10.

Kenett, R. and Pollak, M. (1996). Data-analytic aspects of the Shiryayev-Roberts control chart: surveillance of a non-homogeneos poisson process. *Journal of Applied Statistics*, 23:125–137.

Kleinman, K. (2005). *Generalized linear models and generalized linear mixed models for small-area surveillance*. In Spatial & Syndromic Surveillance for Public Health, A. Lawson and K. Kleinman (eds), 77-94. Chichester, U.K.: John Wiley.

Kleinman, K., Lazarus, R., and Platt, R. (2004). A generalized linear models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology*, 159:217–224.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26:1481–1496.

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

Kulldorff, M. (2003). *Information Management Services, Inc. SaTScan^{TM} Version 3.1: Software for the Spatial and Space-time Scan Statistics*. Information Management Services, Boston.

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

Lawson, A. and Kleinman, K. (2005). *Spatial and Syndromic Surveillance for Public Health*. Wiley, New York.

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

Mevorach, Y. and Pollak, M. (1991). A small sample size comparison of the Cusum and the Shiryayev-Roberts approaches to change point detection. *American Journal of Mathematical and Management Sciences*, 11:277–298.

Montgomery, D. (1996). *Introduction to Statistical Quality Control*. New York: Wiley.

Neil, D. and Cooper, G. (2010). A multivariate bayesian scan statistic for early event detection and characterization. *Machine Learning*, 29:261–282.

Patil, G. and Taillie, C. (2003). Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science*, 18:457–Ö´465.

Pollak, M. (1985). Optimal detection of a change in distribution. *Annals of Statistics*, 13:206–277.

Pollak, M. and Siegmund, D. (1985). A diffusion process and its application to detecting a change in the drift of Brownian motion. *Biometrika*, 72:267–280.

Pollak, M. and Siegmund, D. (1991). Sequential detection of a change in a normal mean when the initial value is unknown. *Annals of Statistics*, 19:394–416.

Pollak, M. and Tartakovsky, A. (2009). Optimality properties of the Shiryayev-Roberts procedure. *Statistica Sinica*, 19:1729–1739.

Roberts, S. (1966). A comparison of some control chart procedures. *Technometrics*, 8:411–430.

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

Rogerson, P. (2001). Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society, Series A*, 164:89–96.

Ryan, T. (1989). *Statistical Methods for Quality Improvement*. New York: Wiley.

Shiryaev, A. (1963). On the detection of disorder in a manufacturing process. *Theory of Probability and its Application*, 8:247–265.

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

Takahashi, K., Kulldorff, M., Tango, T., and Yih, K. (2008). A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics*, 7:14.

Tango, T. and K., T. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geography*, 18:4–11.

Tango, T., Takahashi, K., and Kohriyamma, K. (2011). A space-time scan statistic for detecting emerging outbreaks. *Biometrics*, 67:106–115.

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., and Andrews, N. (2011). Statistical methods for the detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society, Series A*, 175:1–34.

Waller, L. and Gotway, C. (2004). *Applied spatial statistics for public health data*. John Wiley & Sons, New York.

Weesakul, B. (1961). The random walk between a reflecting and an absorbing barrier. *The Annals of Mathematical Statistics*, 32:765–769.

Wetherill, G. and Brown, D. W. (1991). *Statistical Process Control: Theory and Practice*. New York: Chapman and Hall.

Williams, E., Smith, P., Day, N., Geser, A., Ellice, J., and Tukei, P. (1978). Space-time clustering of burkitt's lymphoma in the west nile district of uganda: 1961–1975. *British Journal of Cancer*, 37:109–122.

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

Yakir, B. (1997). Optimal detection of a change in distribution. *Annals of Statistics*, 25:2117–2126.