Universidade Federal de Minas Gerais
Departamento de Estatística
Programa de Pós-Graduação

Sistema de vigilância espaço-temporal para a detecção de conglomerados emergentes em processos pontuais

Mestranda: Thais Rotsen Correa

thaisrc@est.mest.ufmg.br

Orientador: Renato Martins Assunção

assuncao@est.ufmg.br

Belo Horizonte, novembro de 2005

Resumo

Os métodos estatísticos existentes para a detecção de conglomerados espaço-temporais em eventos pontuais são retrospectivos, no sentido de que avaliam se existe evidência a favor da hipótese de interação espaço-tempo em um número fixo de eventos passados. Em contraste, métodos prospectivos tratam uma série de dados sequencialmente, com o intuito de detectar qualquer mudança o mais rápido possível. Neste trabalho propomos um sistema de vigilância prospectivo baseado na razão de verrosimilhanças do processo de Poisson. Este sistema supera duas limitações presentes em propostas anteriores: independência entre os eventos e conhecimento da verrossimilhnaça do processo após a mudança. Simulações mostram que em geral este sistema é eficiente na detecção de conglomerados espaço-temporais, apesar da velocidade de detecção depender de alguns parâmetros do método. O sistema de vigilância é ilustrado utilizando dados de linfoma de Burkitt previamente publicados.

Palavras-chave: Sistema de vigilância, interação espaço-tempo, conglomerado espaço-temporal.

Sumário

1 Vigilância **1**

1.1 Motivação 1

1.2 Revisão Bibliográfica 1

2 Revisão **3**

2.1 Conceitos básicos e notação 3

2.1.1 ARL^0 3

2.1.2 *Delay* 3

2.2 Teste de Knox 4

2.3 CUSUM 5

3 Estatística de Knox local **7**

4 Método de Shiryayev-Roberts **9**

4.1 Descrição do método 9

4.2 Vantagens e desvantagens 10

5 Proposta da dissertação **12**

5.1 Definições 12

5.1.1 Determinação da média $\mu(C_i)$ 14

5.2 Estatística de teste 15

5.3 Implementação do método 17

6 Simulações **20**

6.1 Cenários 20

6.1.1 Sem conglomerado 20

6.1.2 Com conglomerado 21

6.2 Discussão dos resultados 22

6.2.1 Cenários com conglomerado 25

6.2.2 Modelo Paramétrico 33

7 Aplicação aos dados de Burkitt em Uganda **35**

8 Considerações finais **40**

9 Referências Bibliográficas **42**

Lista de Figuras

1 Cilindro C_i 12

2 $\hat{\mu}(C_i)$ 15

3 Número médio de eventos até o alarme soar 25

4 Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.1$ 30

5 Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.2$ 30

6 Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.4$ 31

7 Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.5$ 31

8 Linfoma de Burkitt em West Nile, Uganda (aproximadamente 80 km $\times$ 170 km) 35

9 Aplicação do método proposto aos dados de Burkitt ($\varepsilon = 0.5$ e $\rho = 20$ km) 38

10 Comparação da estatística proposta para os quatro valores de ϵ ($\rho = 20$ km) 39

Lista de Tabelas

1 Número de simulações em que o alarme soa 23

2 Número de eventos até o alarme soar 24

3 Número de simulações em que o alarme soa - Início 26

4 Número de eventos do conglomerado até o alarme soar - Início 26

5 Número de simulações em que o alarme soa - Meio 27

6 Número de eventos do conglomerado até o alarme soar - Meio 27

7 Número de simulações em que o alarme soa - Fim 28

8 Número de eventos do conglomerado até o alarme soar - Fim 28

9 Resultados do Sistema de Vigilância 37

1 Vigilância

1.1 Motivação

A necessidade de métodos e sistemas apropriados para a detecção prospectiva de conglomerados espaço-temporais de forma rápida e eficiente tem sido amplamente reconhecida em diferentes áreas do conhecimento (Raubertas, 1989; Rogerson, 1997; Järpe, 1999; Kulldorff, 2001). Nas áreas de saúde pública e social, em especial, estes métodos são particularmente importantes, uma vez que permitem a adoção de ações preventivas adequadas.

Os tradicionais métodos estatísticos existentes para detecção de conglomerados espaço-temporais em eventos pontuais são retrospectivos. Ou seja, estes métodos objetivam identificar se há evidências a favor da hipótese de interação espaço-tempo em algum conjunto de dados com um número fixo de eventos passados.

No entanto, a maioria dos registros, doenças por exemplo, são constantemente atualizados. A ocorrência de um aumento repentino no risco em uma determinada região geográfica pode indicar o início de uma epidemia. Neste caso, o ideal seria um sistema de monitoramento capaz de detectar este novo conglomerado o mais rápido possível, independentemente de sua localização e tamanho. Esta rápida detecção seria benéfica tanto para os indivíduos quanto para a sociedade, por exemplo, no sentido de reduzir despesas com medicamentos ou evitar que a doença se espalhe.

Vários métodos prospectivos de monitoramento têm sido propostos com o intuito de desenvolver um sistema que soe um alarme assim que um conglomerado emergente é detectado. Ao mesmo tempo, deseja-se minimizar o número de alarmes falsos.

1.2 Revisão Bibliográfica

Sem dúvida, os principais métodos de detecção de mudança de média em processos são a carta de controle de Shewhart e o método de soma acumulada (CUSUM). Estes dois métodos, antigos e tradicionais, têm sido bastante utilizados em técnicas de vigilância.

O mais popular e aparentemente o mais simples método de vigilância é a carta de controle de Shewhart. A desvantagem deste método é que ele não é sensível a

mudanças pequenas ou médias.

O método de soma acumulada (CUSUM) foi desenvolvido para superar esta dificuldade (Page, 1954). CUSUM tem propriedades ótimas (Lorden, 1971; Moustakides, 1986; Ritov, 1990). Assintoticamente o CUSUM minimiza o tempo de espera pelo alarme, dado que houve uma mudança no processo.

Recentemente, muitos outros métodos de vigilância têm sido propostos na literatura. Kulldorff (2001) propôs o uso de uma estatística de varredura para monitoramento prospectivo de doenças; Järpe (1999) propôs o uso do modelo dinâmico de Ising para monitoramento de padrões espaço-temporais; Rogerson (2001) propôs o uso de uma estatística de Knox local para monitoramento de conglomerados espaço-temporais, dentre outros. Uma revisão dos métodos sugeridos na literatura para detecção seqüencial de mudanças no monitoramento de dados de saúde pública é apresentada por Sonesson e Bock (2003).

O método de Shiryayev-Roberts (SR), proposto por Kenett e Pollak (1996), destaca-se por apresentar algumas vantagens em relação aos métodos tradicionais. Ao contrário do CUSUM, este método não requer independência entre os eventos.

2 Revisão

2.1 Conceitos básicos e notação

Para avaliar métodos de vigilância, muitos tipos de medidas são usados para caracterizar o comportamento do processo sob controle e fora de controle. A seção 2.1.1 apresenta uma medida utilizada para descrever a performance do método quando o processo está sob controle. A seção 2.1.2 apresenta uma medida útil para avaliar a habilidade de detecção do teste dado que houve uma mudança no processo.

2.1.1 ARL^0

A principal medida utilizada para descrever a performance do método quando o processo está sob controle é o *average in-control run length* - ARL^0, definido como o número médio de observações até que uma mudança na média seja detectada, sob a hipótese nula de que não houve nenhuma mudança.

Quando o processo está sob controle, todos os alarmes são falsos. A distribuição do alarme falso é geralmente resumida pelo ARL^0, definido por

$$ARL^0 = E\left[t_A | \tau = \infty\right] \tag{1}$$

onde t_A é o tempo em que o alarme soou e τ é o tempo em que a mudança do processo ocorreu. Note que τ é desconhecido. Assim, para uma seqüência de observações X_1, $X_2, \ldots, X_{s-1}, X_s, X_{s+1}, \ldots$ dizemos que o processo está sob controle em X_s se $\tau > s$. No contexto de vigilância, o erro tipo I tem sido tradicionalmente caracterizado pelo ARL^0.

Nos casos em que o valor do ARL^0 é alto, teremos poucos alarmes falsos, mas mudanças reais não serão muito bem detectadas. Da mesma forma, para valores pequenos de ARL^0, as mudanças reais geralmente serão detectadas, mas alarmes falsos também serão mais freqüentes.

2.1.2 *Delay*

O *conditional expected delay* - *CED*, o tempo médio de atraso até que o alarme soe dado que realmente ocorreu uma mudança no processo, é bastante útil para avaliar o

desempenho de sistemas de vigilância, com base no poder de detecção. Esta medida caracteriza o comportamento de um processo fora de controle.

O CED no ponto t, uma vez que a habilidade de detecção depende do ponto em que a mudança ocorreu, é definido como

$$CED(t) = E\left[t_A - \tau | t_A \geq \tau = t\right] \tag{2}$$

Assumindo uma distribuição para τ, pode-se também considerar o *expected delay - ED*, que é a soma ponderada dos tempos entre a mudança e o tempo em que o alarme soou motivadamente:

$$ED_\tau = \sum_{t=1}^{\infty} P(\tau = t)\, P(t_A \geq t)\, CED(t) \tag{3}$$

O *expected delay* é o tempo médio de atraso de um alarme motivado.

2.2 Teste de Knox

O teste proposto por Knox (1964) é um método puramente retrospectivo, voltado para testar globalmente a presença de conglomerados espaço-temporais em processos pontuais.

Este teste baseia-se na contagem do número de pares de eventos que ocorrem em intervalos críticos pré-especificados de tempo e distância. Considerando-se n pontos, existem $n(n-1)/2$ pares de pontos distintos.

Seja n_s o número observado de pares de eventos que são próximos no espaço (ou seja, pares separados por uma distância espacial menor que a distância crítica espacial). Seja n_t o número observado de pares de eventos que são próximos no tempo (ou seja, pares separados por uma distância temporal menor que a distância crítica temporal). As distâncias críticas devem ser definidas pelo usuário de acordo com seu conhecimento sobre o processo.

A estatística de teste é simplesmente n_{st}, o número observado de pares de eventos que são próximos no espaço e no tempo simultaneamente. A estatística de teste excede seu valor esperado $2n_s n_t/n(n-1)$ quando pontos que são próximos no espaço são mais próximos no tempo que o esperado.

Esta estatística é comparada com uma distribuição de referência (sob a hipótese nula de que o processo não apresenta interação espaço-tempo) que é obtida através de permutações aleatórias dos índices de tempo dos eventos originais. Portanto, um valor-p pequeno é uma evidência a favor da hipótese de interação espaço-tempo.

No contexto de vigilância, poderíamos a princípio pensar em realizar o teste de Knox a cada vez que uma nova observação estivesse disponível. Porém, isto faria com que o erro do tipo I aumentasse consideravelmente devido aos sucessivos testes, levando a falsas indicações de interação espaço-tempo.

2.3 CUSUM

Seja x_t a observação no tempo t e assuma que x_t vem de uma distribuição normal com média μ e variância σ^2. Assuma também que a seqüência de observações não apresenta nenhuma correlação serial. Então, a soma acumulada no tempo t é dada por

$$S_t = \max\left(0, S_{t-1} + x_t - \mu - k\sigma\right) \tag{4}$$

onde k é um parâmetro, usualmente igual a $1/2$. Então, esta soma acumula desvios da média que excedem k desvios padrão. Quando S_t excede um valor crítico $h\sigma$, é um sinal de que ocorreu uma mudança na média subjacente. Note que, na prática, os parâmetros μ e σ^2 são desconhecidos. A média μ deve ser estimada através da média amostral $\overline{x}$ e a variância σ^2 deve ser estimada através da variância amostral s^2.

A escolha de h está relacionada ao ARL^0, definido como o número médio de observações até que uma mudança na média seja detectada, sob a hipótese nula de que não houve nenhuma mudança. Valores altos de h estão associados com um ARL^0 longo e, nestes casos, teremos poucos alarmes falsos, mas mudanças reais não serão muito bem detectadas. Da mesma forma, valores baixos de h caracterizam um ARL^0 pequeno e, nestes casos, as mudanças reais geralmente serão detectadas, mas alarmes falsos também serão mais freqüentes.

O cálculo de h baseia-se na suposição de que o comprimento RL (*Run Length*) tem, aproximadamente, uma distribuição exponencial com média ARL^0 sob a hipótese nula. Ou seja, $RL \sim \exp\left(ARL^0\right)$.

Logo, $\alpha = P(RL < n) \approx 1 - \exp(-n/ARL^0)$ e, portanto $ARL^0 = -n/\log(1-\alpha)$. Calculado o ARL^0, o valor de h pode ser obtido através da aproximação de Siegmund (1985),

$$ARL^0 \approx 2\{\exp(h + 1.166) - h - 2.166\} \tag{5}$$

A justificativa de se tomar $k = 1/2$ é que, para um dado valor de h, esta escolha minimiza o tempo para detecção de uma mudança real na média, quando esta mudança é de uma magnitude σ. De forma mais geral, sugere-se que o valor de k deve ser igual a metade da magnitude da mudança que se quer detectar.

No contexto de vigilância, a soma S_i irá exceder o valor crítico h quando observações que apresentam interação espaço-tempo começarem a acumular. Estas observações terão a seguinte característica: entre suas ligações com observações recentes (próximas no tempo), haverá mais ligações que o esperado com observações que são também próximas no espaço. Neste caso o alarme deveria soar.

Note que o CUSUM supõe independência entre os eventos, o que na maioria das vezes não acontece. No sistema de vigilância que vamos propor na seção 5, que também é uma soma acumulada, esta suposição não é mais necessária.

3 Estatística de Knox local

Quando o teste de Knox é significativo numa análise retrospectiva, é geralmente de interesse do pesquisador identificar os pares de eventos que estão próximos no espaço e no tempo simultaneamente, assim como determinar a significância individual dessas observações. Neste sentido, o teste de Knox é global: uma interação espaço-tempo significativa existe nos dados, mas a importância estatística dessas observações continua indeterminada.

Para detectar os pares de eventos responsáveis pela rejeição do padrão aleatório, Rogerson (2001) propôs uma versão local da estatística de Knox, que permite avaliar a significância de cada evento individualmente.

Seja $n_s(i)$ o número de eventos que são próximos no espaço ao evento i e $n_t(i)$ o número de eventos que são próximos no tempo ao evento i.

O valor observado da estatística de Knox local $N_{st}(i)$ é $n_{st}(i)$, o número de eventos que estão próximos ao evento i no espaço e no tempo simultaneamente.

A distribuição nula da estatística $N_{st}(i)$ é obtida fixando-se as localizações dos eventos e permutando os tempos, supondo que cada permutação seja igualmente provável. Assim, para uma dada permutação, a variável $N_{st}(i)$ tem distribuição hipergeométrica com parâmetros $n-1$, $n_s(i)$, $n_t^j(i)$, onde $n_t^j(i)$ é o número de eventos que são próximos no tempo ao evento i quando este evento recebe o j-ésimo tempo.

A distribuição nula de $N_{st}(i)$ é uma soma ponderada de distribuições hipergeométricas, onde os pesos são iguais, refletindo as permutações igualmente prováveis.

O teste de significância pode ser realizado utilizando-se uma aproximação normal para a distribuição de $N_{st}(i)$, sob a hipótese de independência entre espaço e tempo, com valor esperado $E\{N_{st}(i)\}$ e variância $Var\{N_{st}(i)\}$ dadas por:

$$E\{N_{st}(i)\} = \frac{2n_t n_s(i)}{n(n-1)} \tag{6}$$

$$Var\{N_{st}(i)\} = \frac{n_s(i)\{n_s(i)-1\}\{\sum_{j=1}^{n} n_t^j(i)^2 - 2n_t\}}{n(n-1)(n-2)} + \frac{2n_t n_s(i)}{n(n-1)} - \left(\frac{2n_t n_s(i)}{n(n-1)}\right)^2 \tag{7}$$

Para avaliar a significância estatística do valor observado $n_{st}(i)$, a variável $N_{st}(i)$

é padronizada e corrigida por um fator, já que $N_{st}(i)$ é uma variável discreta. A estatística de escore é dada por

$$z_i = \frac{n_{st}(i) - E\{N_{st}(i)\} - 0.5}{\sqrt{Var\{N_{st}(i)\}}} \tag{8}$$

e tem distribuição aproximadamente normal com média 0 e variância 1, sob a hipótese nula de que não há interação espaço-tempo no i-ésimo evento. Então os eventos críticos são todos aqueles que rejeitam a hipótese nula de não interação.

4 Método de Shiryayev-Roberts

Nesta seção descrevemos o método de vigilância SR, proposto por Kenett e Pollak (1996), e citamos os pontos positivos e negativos desta metodologia. O limite do alarme sugerido por esses autores será adotado na nossa proposta (seção 5).

4.1 Descrição do método

O método proposto por Kenett e Pollak (1996) aborda o problema clássico de trabalhar com o monitoramento de um processo temporal $[X_n]$, $n = 1, 2, \ldots$, para uma mudança na distribuição. O processo $[X_n]$ é observado seqüencialmente; as observações iniciais tem uma certa distribuição que muda para uma outra distribuição em um ponto desconhecido no tempo, digamos τ. O interesse é detectar a mudança rapidamente, minimizando os alarmes falsos.

Formalmente, P_τ denota a distribuição do processo $[X_n]$ quando a primeira observação após a mudança ocorre no tempo τ. A distribuição de $[X_n]$ quando não há mudança é denotada por P_∞ e E_τ denota a esperança sob P_τ. Um sistema de vigilância é essencialmente um tempo de parada N com respeito à seqüência $[X_n]$, no qual declara-se que ocorreu uma mudança no processo. N é uma variável aleatória, que depende apenas das observações entre o passado e o presente (inclusive). O sistema de vigilância deve atender à condição

$$E_\infty (N) \geq B \tag{9}$$

onde $E_\infty (N)$ é o ARL^0 e B reflete o ARL^0 aceitável para o usuário.

A estatística de teste de Shiryayev-Roberts (1996) é

$$R_n = \sum_{k=1}^{n} \frac{f_{(k)}(X_1, X_2, \ldots, X_n)}{f_{(\infty)}(X_1, X_2, \ldots, X_n)} \tag{10}$$

onde o numerador é a densidade conjunta de $X_1, \ldots, X_n$ sob a hipótese alternativa dado que a mudança aconteceu em $\tau = k$, e o denominador é a densidade conjunta de $X_1, \ldots, X_n$ sob a hipótese nula. Isto é, R_n é a soma em k das razões de verossimilhanças sob a hipótese de que $\tau = k$ e sob a hipótese nula de que $\tau = \infty$. Seja A o limite do alarme. O tempo N_A em que o alarme soa pela primeira vez, é dado por

$$N_A = \min\left[n \,|R_n \geq A\right] \tag{11}$$

A determinação da $E_\infty\left(N_A\right)$ baseia-se em um argumento de martingala que é válido mesmo quando as observações são dependentes. Seja

$$\Lambda_{k,n} = \frac{f_{(k)}(X_1, X_2, \ldots, X_n)}{f_{(\infty)}(X_1, X_2, \ldots, X_n)} \tag{12}$$

Sob P_∞, toda seqüência $[\Lambda_{k,n}]$, $n = 1, 2, \ldots$, de razão de verossimilhanças é uma martingala com esperança unitária. Então, tem-se que

$$R_n - n = \sum_{k=1}^{n} (\Lambda_{k,n} - 1) \tag{13}$$

é uma martingala com esperança zero (Kenett e Pollak, 1996). Pelo teorema da amostragem opcional (Karlin e Taylor, 1975), $E_\infty(R_{N_A} - N_A) = 0$. Então, $E_\infty\left(N_A\right) = E_\infty\left(R_{N_A}\right)$. Por definição , $R_{N_A} \geq A$, logo, $E_\infty\left(N_A\right) \geq A$. Uma vez que N_A é o primeiro tempo em que R_n excede A, o excesso é tipicamente pequeno, de forma que considerar $A = B$ gera um procedimento moderadamente conservativo que satisfaz à equação (9). Geralmente, a existência de uma constante $C > 1$ tal que $E_\infty\left(N_A\right) \geq CA$ pode ser provada e seu valor calculado em alguns casos simples, de forma que $A = B/C$ (Pollak, 1987). Então o alarme soa quando $R_n \geq A$ pela primeira vez e A é o valor desejado do ARL^0.

4.2 Vantagens e desvantagens

Assim como o CUSUM, o método SR também possui propriedades ótimas (Pollak,1985; Yakir, 1994). Assintoticamente, o método SR minimiza o tempo de espera por um alarme motivado dentre todos os tempos de parada que satisfazem uma dada restrição na taxa de alarmes falsos. Em termos de velocidade de detecção, os dois métodos são geralmente comparáveis (Shiryayev, 1963; Roberts, 1966; Pollak e Siegmund, 1985, 1991; Mevorach e Pollak, 1991).

A principal vantagem do método SR em relação ao método CUSUM é na relativa facilidade de sua aplicação sob suposições mínimas. Ao contrário do CUSUM, o método SR não requer independência entre as observações.

Resultados publicados sugerem que o método SR é no mínimo tão eficiente quanto os procedimentos clássicos mais conhecidos (Kenett e Pollak, 1996). Além disso, tais resultados mostram também que a carta de controle SR tem certas vantagens técnicas em relação ao CUSUM e à carta de Shewhart. 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.

No entanto, é fácil notar que o método SR depende do conhecimento prévio da distribuição conjunta de $X_1, \ldots, X_n$ após a mudança do processo. Esta dependência é explícita no numerador da estatística de teste. Na prática, geralmente tem-se muito pouca (ou nenhuma) informação sobre tal distribuição. Assim, a dependência do método SR ao conhecimento desta distribuição é uma limitação. O sistema de vigilância que vamos propor na próxima seção supera esta limitação.

5 Proposta da dissertação

Nesta seção definimos a estatística de teste para o nosso sistema de vigilância e descrevemos o funcionamento do mesmo. Este sistema, além de não assumir eventos independentes, independe do conhecimento prévio da verrossimilhnaça após a mudança do processo. Estas são as principais vatangens do nosso método em relação aos demais métodos de vigilância, em particular ao CUSUM e ao método SR.

5.1 Definições

Suponha que N seja um processo de Poisson em $\mathbb{R}^3$ parcialmente observado em uma região tridimensional finita $\mathcal{A} \times [0, T]$. Seja $N(C_i)$ o número de eventos em um cilindro C_i. A base deste cilindro é um círculo S_i no plano das coordenadas espaciais (x e y) centrado na i-ésima observação e com raio igual a uma distância crítica espacial ρ. Seja t_i o tempo da i-ésima observação e t_n o tempo da n-ésima observação. O cilindro C_i começa em $t = t_i$ e termina em $t = t_n$. Ou seja, a altura deste cilindro é dada pela diferença $t_n - t_i$. A Figura 1 ilustra o cilindro C_i.

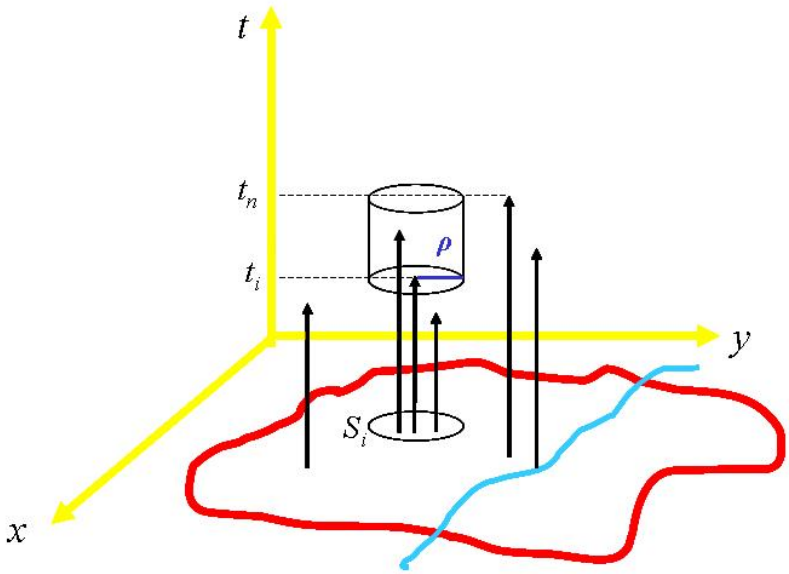

Figura 1: Cilindro C_i

Então $N(C_i)$ é uma variável aleatória que segue uma distribuição de Poisson com média desconhecida $\mu(C_i)$, o número médio de eventos no cilindro C_i.

Seja $\lambda(x, y, t)$ a função de intensidade de eventos na região tridimensional $\mathcal{A} \times [0, T]$. A média $\mu(C_i)$ é dada por

$$\mu(C_i) = \int_{C_i} \lambda(x, y, t)\, dx\, dy\, dt \tag{14}$$

Seja

$$\mu = E(N(\mathcal{A}, [0, T]) = \int_{\mathcal{A}} \int_{[0,T]} \lambda(x, y, t)\, dt\, dx\, dy \tag{15}$$

o número esperado de eventos na região tridimensional $\mathcal{A} \times [0, T]$.

Seja $\lambda_S(x, y)$ a função densidade de eventos no espaço e $\lambda_T(t)$ a função densidade de eventos no tempo. As densidades marginais são dadas por

$$\lambda_S(x, y) = \mu^{-1} \int_{[0,T]} \lambda(x, y, t)\, dt \tag{16}$$

e

$$\lambda_T(t) = \mu^{-1} \int_{\mathcal{A}} \lambda(x, y, t)\, dx\, dy \tag{17}$$

Note que

$$\int_{\mathcal{A}} \lambda_S(x, y)\, dx\, dy \; = \int_{[0,T]} \lambda_T(t)\, dt = 1 \tag{18}$$

Sob a hipótese nula de que não há interação espaço-tempo e portanto não há conglomerados emergentes, temos

$$\lambda(x, y, t) = \mu\, \lambda_S(x, y)\, \lambda_T(t) \tag{19}$$

Sob a hipótese alternativa de interação espaço-tempo, vamos supor que existe uma constante ε e um cilindro C tal que $\lambda(x, y, t)$ pode ser escrita como

$$\lambda(x, y, t) = \mu\, \lambda_S(x, y)\, \lambda_T(t)\, (1 + \varepsilon\, I_C(x, y, t)) \tag{20}$$

onde $I_C(x, y, t)$ é uma função que indica se o ponto (x, y, t) pertence ao cilindro C. Este modelo é semelhante ao modelo adotado por Assunção e Maia (2005). O modelo alternativo supõe que, fora de C, $\lambda(x, y, t)$ é o produto de funções separáveis do espaço e do tempo. Dentro de C, $\lambda(x, y, t)$ passa a ser este produto magnificado pela

constante multiplicativa $(1+\varepsilon)$. Assim, ε é a mudança relativa dentro de C entre os limites temporais do cilindro. ε pode ser visto também como um coeficiente de não separabilidade.

Sob a hipótese nula, temos $\varepsilon\, I_C(x,y,t) = 0$. Queremos comparar a hipótese nula (19) com a hipótese alternativa (20) onde espaço e tempo interagem.

5.1.1 Determinação da média $\mu(C_i)$

Modelo de Poisson Homogêneo

O modelo de Poisson homogêneo implica na independência entre espaço e tempo. Neste modelo, a média $\mu(C_i)$ é proporcional apenas ao volume do cilindro, independentemente de sua localização. Ou seja,

$$\mu(C_i) = \lambda v_i \tag{21}$$

onde v_i é o volume do cilindro C_i e λ é a função de intensidade de eventos num volume tri-dimensional unitário.

Modelo de Poisson Não Homogêneo

Suponha que N é um modelo de Poisson não homogêneo onde a suposição de independência entre espaço e tempo é válida. Nesta situação, a média $\mu(C_i)$ é dada por

$$\mu(C_i) = \int_{C_i} \lambda(x,y,t)\, dx\, dy\, dt \tag{22}$$

No modelo de Poisson não homogêneo, sob a hipótese nula de que não existe interação espaço-tempo, a intensidade de eventos é dada por (19). Neste caso, a média $\mu(C_i)$, definida em (22) pode ser escrita como

$$\mu(C_i) = \mu \int_{S_i} \lambda_S(x,y)\, dx\, dy \int_{[t_i,t_n]} \lambda_T(t)\, dt \tag{23}$$

Uma estimativa de $\mu(C_i)$ sob a hipótese nula é obtida através da equação

$$\hat{\mu}(C_i) = \frac{N(S_i \times [0,T])\, N(\mathcal{A} \times [t_i,t_n])}{n} \tag{24}$$

onde $N(S_i \times [0,T])$ é o número de eventos pertencentes ao círculo S_i considerando-se todo o eixo do tempo; $N(\mathcal{A} \times [t_i, t_n])$ é o número de eventos cujos tempos estão entre t_i e t_n considerando-se todo o plano e n é o número total de eventos.

Os termos $N(S_i \times [0,T])$ e $N(\mathcal{A} \times [t_i, t_n])$ são ilustrados na Figura 2. Neste caso temos $N(C_i) = 2$, $N(S_i \times [0,T]) = 3$, $N(\mathcal{A} \times [t_i, t_n]) = 4$ e $n = 6$.

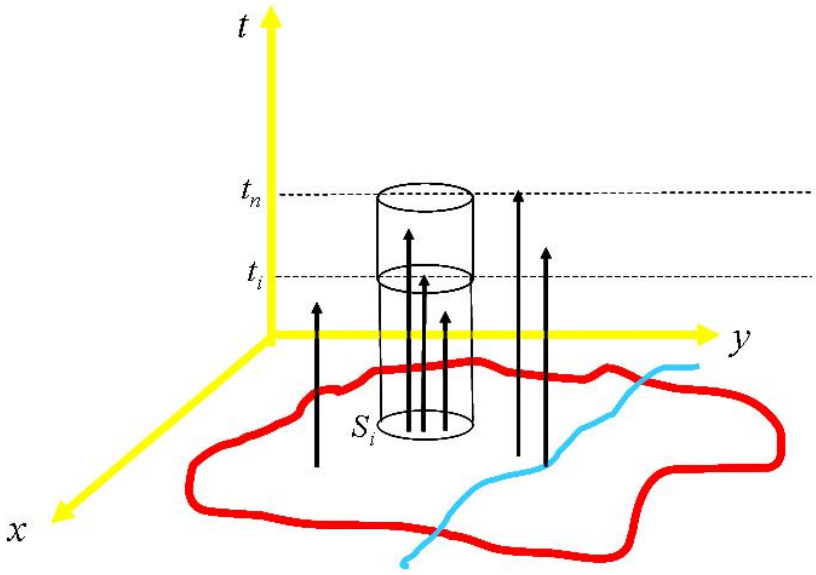

Figura 2: $\hat{\mu}(C_i)$

5.2 Estatística de teste

A estatística de teste aqui proposta se baseia na função de verossimilhança do Processo de Poisson espaço-temporal, que é dada por:

$$\left(\prod_{i=1}^{n} \lambda(x_i, y_i, t_i)\right) \exp\left(-\int_{R^3} \lambda(x,y,t)\, dx\, dy\, dt\right) \tag{25}$$

Assumindo que a mudança aconteceu em $t = \tau$, temos as verrossimilhanças sob a hipótese nula (L_∞) e sob a hipótese alternativa (L_τ):

$$L_\infty = \left(\prod_{i=1}^{n} \lambda(x_i, y_i, t_i)\right) \exp\left(-\int_{R^3} \lambda(x,y,t)\, dx\, dy\, dt\right) \tag{26}$$

$$L_\tau = \left(\prod_{i=1}^{n} \lambda(x_i, y_i, t_i)\,(1 + \varepsilon\, I_{C_\tau}(x_i, y_i, t_i)\right) \exp\left(-\int_{R^3} \lambda(x, y, t)\, dx\, dy\, dt\right) \exp\left(-\varepsilon \int_{C_\tau} \lambda(x, y, t)\, dx\, dy\, dt\right) \quad (27)$$

onde $\lambda(x, y, t) = \mu\, \lambda_S(x, y)\, \lambda_T(t)$ e C_τ é o cilindro que representa o conglomerado. A base deste cilindro é um círculo de raio ρ centrado na τ-ésima observação. O conglomerado C_τ começa em $t = t_\tau$ e termina em $t = t_n$, sendo $t_\tau \leq t_n$. A escolha de $t = t_n$ para o tempo final do conglomerado será justificada ainda nesta seção. Note que quando dizemos que $t = \tau$, os valores possíveis de τ são apenas os valores $t_1, t_2, \ldots, t_n$, o que não significa uma perda grande no poder do método.

A estatística de teste R_n é definida em função da razão dessas verossimilhanças:

$$R_n = \sum_{\tau=1}^{n} \frac{L_\tau}{L_\infty} = \sum_{\tau=1}^{n} \left\{ \left[\prod_{i=1}^{n} (1 + \varepsilon\, I_{C_\tau}(x_i, y_i, t_i)\right] \exp\left(-\varepsilon \int_{C_\tau} \lambda(x, y, t)\, dx\, dy\, dt\right) \right\} \quad (28)$$

Note que

$$\mu(C_\tau) = \int_{C_\tau} \lambda(x, y, t)\, dx\, dy\, dt \quad (29)$$

nada mais é que o número esperado de eventos no cilindro C_τ sob a hipótese nula, podendo ser estimado de forma semelhante à equação (24).

Então a equação (28) pode pode ser escrita como:

$$R_n = \sum_{\tau=1}^{n} (1 + \varepsilon)^{N(C_\tau)} \exp(-\varepsilon\, \mu(C_\tau)) \quad (30)$$

onde $\varepsilon > 0$ é conhecido e quantifica a intensidade da mudança, devendo seu valor ser fornecido pelo usuário de acordo com a mudança que ele deseja detectar. $\varepsilon = 0.3$ por exemplo, significa que dentro do conglomerado a intensidade de eventos é 30% a mais que fora do conglomerado.

Suponha que um conglomerado emergente começa na τ-ésima observação. A princípio não sabemos qual é a observação que identifica o final deste conglomerado. Por isso iremos supor que este conglomerado permanece até a observação corrente, a n-ésima observação. Afinal, se o conglomerado começou em τ e terminou em algum $m < n$, ele deveria ter sido detectado pelo sistema até m. Após este tempo m, o

conglomerado não existe mais e os eventos sucessivos não trazem mais informações relevantes.

Assim, temos que $\mu(C_\tau)$ pode ser estimado por

$$\hat{\mu}(C_\tau) = \frac{N(S_\tau \times [0,T])N(\mathcal{A} \times [t_\tau, t_n])}{n} \tag{31}$$

Dado τ, $N(S_\tau \times [0,T])$ é o número de eventos pertencentes ao cilindro $S_\tau \times [0,T]$. Este cilindro, que é espacialmente centrado no ponto (x_τ, y_τ), começa na primeira observação e termina na última, ou seja, começa em $t = t_1$ e termina em $t = t_n$. $N(\mathcal{A} \times [t_\tau, t_n])$ é o número de eventos que ocorrem no intervalo $[t_\tau, t_n]$ considerando-se todo o plano espacial e n é o número total de eventos.

Dessa forma, o sistema de vigilância espaço temporal calcula R_{n+1} à medida que o $(n+1)$-ésimo evento é adicionado ao conjunto de dados, sendo que R_n é redefinido com $\hat{\mu}(C_\tau)$ em vez de $\mu(C_\tau)$ em (30). O alarme soa quando $R_n \geq A$ pela primeira vez e A é o valor desejado do ARL^0.

5.3 Implementação do método

Nesta seção apresentaremos alguns detalhes sobre a implementação do método proposto. O algoritmo abaixo descreve passo a passo o cálculo da estatística R_n. Note que o parâmetro ε de incremento relativo da intensidade deve ser fornecido pelo usuário para usar a estatística de teste (30). Na seção 6.2 vamos discutir o impacto da escolha de ε.

Algoritmo R_n

Entrada

matriz de n linhas e 3 colunas com as coordenadas espaciais e temporais

coluna 1: coordenada x, coluna 2: coordenada y, coluna 3: coordenada t

Parâmetros

ε: magnitude da mudança

ρ: raio do círculo base do cilindro

A: limite do alarme

Ordene os eventos por tempo

Inicializações

R: vetor de n posições que contém os valores da estatística de teste

$RMax$: vetor de n posições que contém o valor da maior parcela de R

$RMaxInd$: vetor de n posições que contém o valor de τ correspondente à maior parcela de R

Para i **de 1 até** n # i é o índice de R $(R_1, \ldots, R_n)$

Para τ **de 1 até** i

Inicializações

$N(C_\tau)$: número de eventos no cilindro C_τ

$N(\mathcal{A} \times [t_\tau, t_n])$: número de eventos na faixa $\mathcal{A} \times [t_\tau, t_n]$

$N(S_\tau \times [0, T])$: número de eventos no cilindro $S_\tau \times [0, T]$

calcula $N(C_\tau)$

calcula $N(\mathcal{A} \times [t_\tau, t_n])$

calcula $N(S_\tau \times [0, T])$

calcula $\hat{\mu}(C_\tau)$: número de eventos esperados no cilindro sob a hipótese nula

calcula $RTemp$: $(1 + \varepsilon)^{N(C_\tau)} * \exp(-\varepsilon\, \hat{\mu}(C_\tau))$

Fim Para

identifica e armazena o maior valor de $RTemp$ e o τ correspondente

calcula e armazena o valor da estatística de teste R_i
(somatório de $RTemp$ para todo τ)

Fim Para

Inicializa o contador de número de alarmes

Para i **de 1 até** n

se R_i é maior ou igual ao limite do alarme

armazena o valor de R_i e o índice i correspondente

fim se

Fim Para

calcula o número de alarmes no processo

Saída

Plota os valores $R_1, \ldots, R_n$

Adiciona o limite do alarme ao gráfico

Escreve um arquivo com:
parâmentros (ε, ρ, A),

estatísticas de teste,
valor da maior parcela das estatísticas de teste e valor de τ correspondente,
número de alarmes,
índices dos casos críticos.

Na prática, os valores $N(C_\tau)$, $N(\mathcal{A} \times [t_\tau, t_n])$ e $N(S_\tau \times [0, T])$ foram calculados a partir da matriz das distâncias espaciais (distância euclidiana) dos eventos. Através deste artifício conseguimos tornar o programa extremamente eficiente.

O cálculo das distâncias temporais entre os eventos não foi necessário, uma vez que a estatística R_n depende somente da ordem em que os eventos ocorrem, e o primeiro passo do algoritmo é justamente ordenar os eventos por tempo.

6 Simulações

O método proposto foi testado via simulação em alguns cenários pré-definidos. Em cada cenário foram feitas 1000 simulações.

A seção 6.1 contém uma descrição completa de cada um destes cenários. Os resultados das simulações são mostrados na seção 6.2.

6.1 Cenários

A estatística de teste foi avaliada em um cenário de completa aleatoriedade (sem conglomerado) e em algumas variações de um cenário onde existe um conglomerado em forma de paralelepípedo. O cenário sem conglomerado e os cenários com conglomerado são descritos detalhadamente nas seções 6.1.1 e 6.1.2, respectivamente.

Para os dois tipos de cenário (sem conglomerado e com conglomerado) foram testados quatro valores para o parâmetro ε (magnitude da mudança): $\varepsilon_1 = 0.1$, $\varepsilon_2 = 0.2$, $\varepsilon_3 = 0.4$ e $\varepsilon_4 = 0.5$.

6.1.1 Sem conglomerado

O principal objetivo da utilização de um cenário em que não existe nenhum conglomerado é avaliar se a decisão de tomar $A = B$ para fixar o o limite do método de Shiryayev-Roberts é adequada e com isto, avaliar o ARL^0. De acordo com esta aproximação, espera-se que, sob a hipótese nula, o alarme soe em média na A-ésima observação (considerando que as observações estão ordenadas no tempo), onde A é o limite do alarme.

Foram feitas 1000 simulações independentes e, em cada simulação, utilizou-se 1000 eventos sucessivos. Os limites A testados foram 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 950, 1000.

Em todas as simulações utilizou-se raio crítico espacial ρ (usado para definir observações próximas no espaço) igual a 2, aproximadamente 14% da distância máxima possível dentro da região $\mathcal{A}$ de observação.

Considerou-se uma região em forma de um paralelepípedo $10 \times 10 \times 1000$ e foram gerados sempre 1000 pontos/eventos distribuídos uniformemente nesta região.

Note que neste caso (sob a hipótese nula) o valor correto para a magnitude da mudança é $\varepsilon = 0$.

6.1.2 Com conglomerado

O principal objetivo da utilização de um cenário onde existe um conglomerado é avaliar a capacidade de detecção do sistema de vigilância aqui proposto. Como τ é conhecido, uma possibilidade seria adotar o CED, definido na equação (2), para tal avaliação. Porém, para este problema, esta não é uma boa escolha, já que todos os eventos que ocorrem entre τ e t_A, pertencendo ou não ao conglomerado, contribuem para esta estatística. Portanto, ao invés do CED, utilizamos o número médio de eventos pertencentes ao conglomerado até o alarme soar para avaliar a capacidade de deteção do sistema. O ideal é que este número seja o menor possível.

Foram feitas 1000 simulações independentes e, em cada simulação, utilizou-se 500 eventos sucessivos. Os limites A testados foram 50, 100, 150, 200, 250, 300, 350, 400, 450, 500. Aqui os limites são diferentes do caso anterior (sem conglomerado) porque o tamanho da amostra é menor.

Em todos os cenários com conglomerado, foram gerados sob a hipótese nula 500 eventos no paralelepípedo $10 \times 10 \times 500$. Ou seja, foram gerados sempre 500 pontos/eventos distribuídos uniformemente nesta região.

Em seguida incluímos o conglomerado, também em forma de paralelepípedo. Em todos os casos a base deste paralelepípedo foi um quadrado de lado igual a 1. Já a altura variou de acordo com o tempo de início do conglomerado. Considerou-se sempre um conglomerado que começava em algum momento do tempo e permanecia "vivo"até o final do estudo. Ou seja, a altura do paralelepípedo referente ao conglomerado é dada por (500 - tempo de início do conglomerado).

Em relação ao tempo de início do conglomerado foram considerados três casos:
i) conglomerado que aparece logo no início do estudo (começa na 50^a observação);
ii) conglomerado que aparece um pouco antes do meio do estudo (começa na 150^a observação);
iii) conglomerado que aparece um pouco depois do meio do estudo (começa na 300^a observação).

Assim, no caso i) o conglomerado foi representado por um paralelepípedo $1 \times 1 \times 450$; no caso ii) o conglomerado foi representado por um paralelepípedo $1 \times 1 \times 350$; no caso iii) o conglomerado foi representado por um paralelepípedo $1 \times 1 \times 200$.

O número de eventos gerados dentro do conglomerado é dado por (1/5 * altura do conglomerado). Então, no caso i) gerou-se 90 eventos dentro do cluter (450/5); no caso ii) gerou-se 70 eventos dentro do conglomerado (350/5); no caso iii) gerou-se 40 eventos dentro do conglomerado (200/5). Assim espera-se que, dentro do conglomerado, a intensidade observada de eventos seja aproximadamente a intensidade fora do conglomerado multiplicada por (1+1/5). Isto é, neste caso (sob a hipótese alternativa), o valor correto para a magnitude da mudança é $\varepsilon = 0.2$.

Para simplificar a notação do texto, o caso i) será identificado apenas como Início, o caso ii) apenas como Meio e o caso iii) apenas como Fim.

Resumindo, gerou-se sempre:
- 500 eventos distribuídos uniformemente em todo o paralelepípedo $\mathcal{A}$;
- (500 - tempo de início do conglomerado)/5 eventos distribuídos uniformemente dentro do conglomerado em forma de paralelepípedo.

Note que mesmo depois do surgimento do conglomerado continuam aparecendo eventos fora do mesmo.

Em todas as simulações utilizou-se raio crítico espacial ρ (usado para definir observações próximas no espaço) igual a 1, valor correspondente ao lado do quadrado que é base do paralelepípedo referente ao conglomerado.

6.2 Discussão dos resultados

A Tabela 1 mostra, para cada valor de ε, o número de simulações em que o alarme soa em cada um dos limites A testados no cenário sem conglomerado.

Vale lembrar que o total de simulações é 1000 e que neste caso (sob a hipótese nula) o valor correto da magnitude da mudança é $\varepsilon = 0$.

Note que, independente do valor de ε, à medida que o limite aumenta, o número de simulações em que o alarme soa diminui. Ou seja, quanto mais alto é o limite, mais o alarme demora para soar.

Observa-se também que para valores de ε maiores, simulações em que o alarme

não soa aparecem em limites mais baixos. Para $\varepsilon_1 = 0.1$, por exemplo, o alarme soa nas 1000 simulações para todos os limites menores que 900. Já para $\varepsilon_2 = 0.2$, o alarme soa em todas as simulações apenas para limites menores que 600. Isto é, o ARL^0 é função de ε, o parâmetro de mudança relativa da intensidade considerado pelo modelo alternativo. Quanto maior o valor ε dessa mudança relativa considerado pelo modelo, maior será o ARL^0, ultrapassando rapidamente o limite nominal A.

Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	1000	1000	1000	1000
100	1000	1000	1000	1000
200	1000	1000	1000	1000
300	1000	1000	998	947
400	1000	1000	702	515
500	1000	1000	257	224
600	1000	940	109	115
700	1000	209	49	67
800	1000	21	25	38
900	289	4	18	24
950	14	3	10	22
1000	2	1	7	21

Tabela 1: Número de simulações em que o alarme soa

Na Tabela 2 estão representados a média e o desvio padrão do número de eventos até o alarme soar. Estes valores foram calculados a partir dos dados disponíveis, ou seja, foram calculados com base no número de simulações em que o alarme soa (Tabela 1). Isto significa que, para os casos em que o alarme não soa em todas as 1000 simulações, tanto a média quanto o desvio estão na verdade subestimados. Esta subestimação será discutida na seção 6.2.3.

	Média				Desvio Padrão			
Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	48.997	48.543	49.956	51.637	0.538	1.022	2.167	2.885
100	98.212	99.476	111.424	122.121	1.151	2.428	6.947	11.141
200	199.360	212.344	290.637	351.760	2.630	6.696	34.225	69.181
300	303.907	341.492	557.938	639.397	4.855	14.817	115.612	158.561
400	412.298	490.773	765.150	757.957	7.625	28.721	141.228	157.265
500	524.481	663.338	816.611	800.049	11.607	51.978	129.238	136.293
600	640.954	856.162	841.642	811.270	15.384	75.671	110.603	127.861
700	761.390	932.512	849.939	827.821	20.798	61.698	104.506	111.935
800	887.741	927.190	853.440	842.211	25.960	47.629	87.595	107.787
900	979.439	954.500	848.944	840.792	19.274	44.223	95.226	97.969
950	980.786	959.667	872.600	847.864	23.192	51.733	104.128	91.242
1000	969.500	908.000	842.000	846.381	27.577	—	108.870	91.464

Tabela 2: Número de eventos até o alarme soar

Claramente, ignorando-se a subestimação causada pela censura do experimento em 1000 eventos, verifica-se que tanto a média quanto o desvio padrão do número de eventos até o alarme soar aumentam à medida que o limite A aumenta. E para um dado limite, essas duas estatísticas aumentam à medida que o valor de ε cresce. Ou seja, quanto mais afastado o valor de ε está de seu valor verdadeiro ($\varepsilon = 0$), mais o alarme demora para soar.

Para $\varepsilon_2 = 2$ e limite=1000 não foi possível calcular o desvio padrão, já que neste caso o alarme soa somente em uma simulação (Tabela 1).

A Figura 3 mostra o número médio de eventos até que o alarme soe em cada um dos limites A testados. A reta em vermelho representa o limite $A = B$ proposto por Kenett e Pollak. Para $\varepsilon_1 = 0.1$ (valor de ε mais próximo do verdadeiro) o limite $A = B$ parece ser válido. Mas à medida que o valor de ε aumenta (se afasta de seu valor verdadeiro), este limite vai se mostrando cada vez mais inadequado.

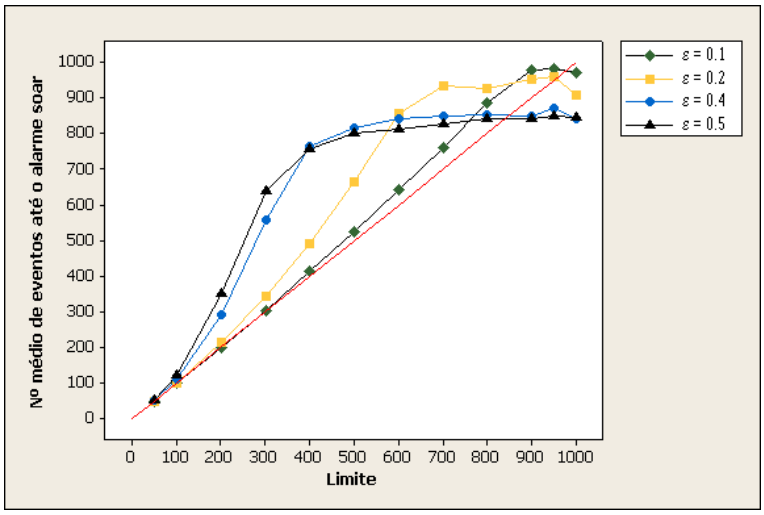

Figura 3: Número médio de eventos até o alarme soar

6.2.1 Cenários com conglomerado

As Tabelas 3, 5 e 7 mostram o número de simulações em que o alarme soa para o conglomerado que começa no início, no meio e no fim do estudo, respectivamente.

As Tabelas 4, 6 e 8 mostram a média e o desvio padrão do número de eventos do conglomerado até o alarme soar, usado para avaliar a capacidade de detecção do sistema, para o conglomerado que começa no início, no meio e no fim do estudo, respectivamente. Essas duas estatísticas foram calculadas com base no número de simulações em que o alarme soa em cada caso (Tabelas 3, 5 e 7). Como antes, para os casos em que o alarme não soa em todas as 1000 simulações, o número médio de eventos até o alarme soar está na verdade subestimado.

Alguns símbolos são usados nas Tabelas 4, 6 e 8:

* : alarme soa sempre antes do início do conglomerado;

a: alarme soa depois do início do conglomerado em 2 simulações;

b: alarme soa depois do início do conglomerado em 69 simulações;

c: alarme soa depois do início do conglomerado em 838 simulações;

d: alarme soa depois do início do conglomerado em 967 simulações.

Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	1000	1000	1000	1000
100	1000	1000	1000	1000
150	1000	1000	1000	1000
200	1000	1000	1000	1000
250	1000	1000	1000	1000
300	1000	1000	1000	1000
350	1000	1000	1000	1000
400	1000	1000	1000	941
450	1000	1000	839	505
500	1000	1000	424	323

Tabela 3: Número de simulações em que o alarme soa - Início

	Média				**Desvio Padrão**			
Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	*	*	*	*	*	*	*	*
100	7.483	6.871	6.229	6.062	2.274	2.063	1.792	1.686
150	15.428	14.368	13.344	13.172	3.064	2.788	2.533	2.606
200	23.501	22.090	21.304	21.584	3.540	3.289	3.807	4.757
250	31.611	30.074	30.210	31.575	3.834	3.828	5.893	7.806
300	39.724	38.324	40.469	43.035	4.113	4.553	8.403	11.399
350	47.997	46.927	51.816	56.116	4.363	5.346	11.381	15.129
400	56.431	55.967	64.194	69.274	4.551	6.359	14.327	18.434
450	64.809	65.189	73.983	67.952	4.719	7.419	16.335	18.939
500	73.330	74.811	72.226	67.567	4.849	8.394	16.261	18.874

Tabela 4: Número de eventos do conglomerado até o alarme soar - Início

Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	1000	1000	1000	1000
100	1000	1000	1000	1000
150	1000	1000	1000	1000
200	1000	1000	1000	1000
250	1000	1000	1000	1000
300	1000	1000	1000	1000
350	1000	1000	1000	1000
400	1000	1000	1000	1000
450	1000	1000	1000	1000
500	1000	1000	997	996

Tabela 5: Número de simulações em que o alarme soa - Meio

	Média				**Desvio Padrão**			
Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	*	*	*	*	*	*	*	*
100	*	*	*	*	*	*	*	*
150	*	*	1.000^a	0.348^b	*	*	1.414^a	0.614^b
200	6.642	5.909	5.597	5.714	2.221	1.937	1.552	1.448
250	14.274	12.438	10.557	10.036	2.714	2.030	1.398	1.465
300	21.551	18.365	14.459	13.274	2.796	1.924	1.778	2.126
350	28.576	23.823	17.837	16.081	2.895	2.077	2.697	3.214
400	35.480	29.029	20.973	18.742	2.926	2.528	3.816	4.496
450	42.345	34.166	24.011	21.323	3.076	3.341	5.314	6.100
500	49.282	39.414	27.149	23.765	3.198	4.455	6.976	7.679

Tabela 6: Número de eventos do conglomerado até o alarme soar - Meio

Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	1000	1000	1000	1000
100	1000	1000	1000	1000
150	1000	1000	1000	1000
200	1000	1000	1000	1000
250	1000	1000	1000	1000
300	1000	1000	1000	1000
350	1000	1000	1000	1000
400	1000	1000	1000	1000
450	1000	1000	1000	1000
500	1000	1000	1000	1000

Tabela 7: Número de simulações em que o alarme soa - Fim

	Média				**Desvio Padrão**			
Limite A	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$	$\varepsilon_1 = 0.1$	$\varepsilon_2 = 0.2$	$\varepsilon_3 = 0.4$	$\varepsilon_4 = 0.5$
50	*	*	*	*	*	*	*	*
100	*	*	*	*	*	*	*	*
150	*	*	*	*	*	*	*	*
200	*	*	*	*	*	*	*	*
250	*	*	*	*	*	*	*	*
300	*	0.000^a	1.409^c	3.169^d	*	0.000^a	1.357^c	1.759^d
350	6.049	5.774	7.250	7.678	2.009	1.813	1.560	1.557
400	13.467	11.897	10.661	10.166	2.243	1.714	1.547	1.656
450	20.041	16.645	12.945	11.837	2.133	1.570	1.689	1.847
500	26.185	20.546	14.723	13.164	1.976	1.634	1.919	2.001

Tabela 8: Número de eventos do conglomerado até o alarme soar - Fim

De acordo com a Tabela 3, verifica-se que para ε_3 - limite = 450, 500 e ε_4 - limite = 400, 450, 500, o alarme soa apenas numa parcela das 1000 simulações.

Já para o conglomerado que começa no meio do estudo (Tabela 5), o alarme só não soa nas 1000 simulações para $\varepsilon_3 = 0.4$ e $\varepsilon_4 = 0.5$ com limite = 500. Mas nestes casos, o número de simulações em que o alarme não soa é extremamente pequeno, de forma que seu impacto no número médio de eventos pertencentes ao conglomerado até o alarme soar é mínimo.

No caso do conglomerado que começa no fim do estudo (Tabela 7), o alarme soa nas 1000 simulações em todas as combinações de ε e limite testadas.

Verifica-se que, para $\varepsilon_3 = 0.4$ e $\varepsilon_4 = 0.5$, o número de simulações em que o alarme soa é maior quando o conglomerado está mais próximo do final do estudo, ou seja, quanto mais no início do estudo o conglomerado aparece, mais o alarme demora para soar. Já para valores menores de ε ($\varepsilon_1 = 0.1$ e $\varepsilon_2 = 0.2$), o alarme soa sempre em todas as simulações, independentemente do tempo de início do conglomerado.

A situação em que o alarme soa antes do início do conglomerado (representada por * nas Tabelas 4, 6 e 8) não é relevante para a análise. Nestes casos, o limite A estabelecido é menor que o tempo de início do conglomerado e portanto já esperamos que o alarme vá soar falsamente antes do início do conglomerado.

Para o conglomerado que começa no meio e no fim do estudo (Tabelas 6 e 8), verifica-se que o número médio de eventos do conglomerado até o alarme soar diminui à medida que o valor de ε aumenta. Ou seja, o desempenho do método melhora quando o valor de ε fornecido pelo usuário é um pouco maior que o valor verdadeiro deste parâmetro.

As Figuras 4, 5, 6 e 7 mostram o número médio de eventos do conglomerado até o alarme soar para ε_1, ε_2, ε_3 e ε_4, respectivamente.

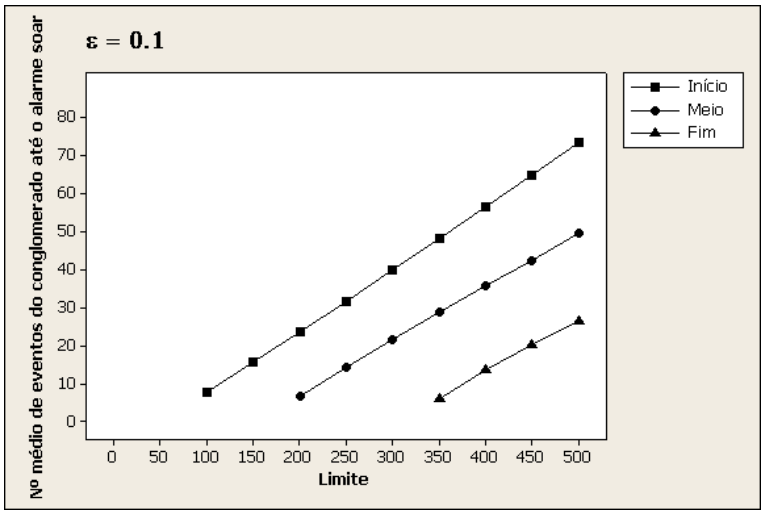

Figura 4: Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.1$

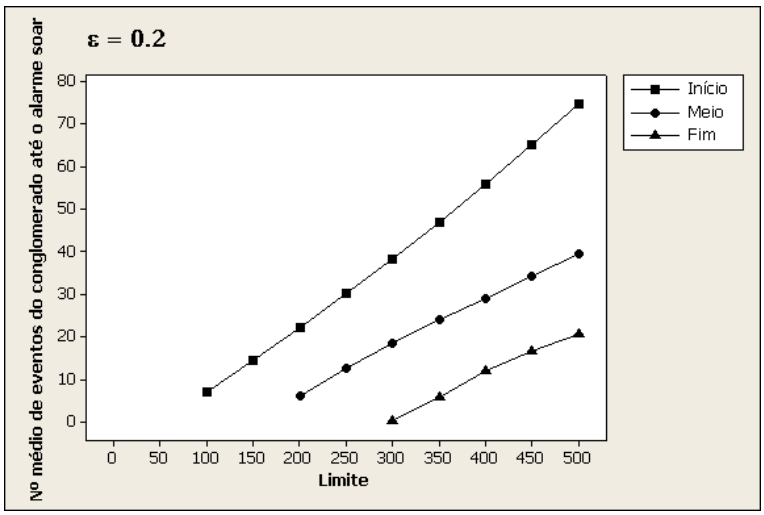

Figura 5: Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.2$

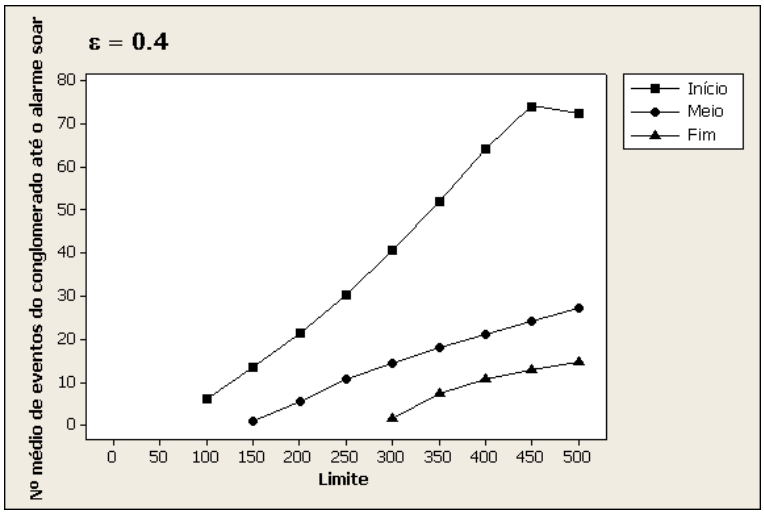

Figura 6: Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.4$

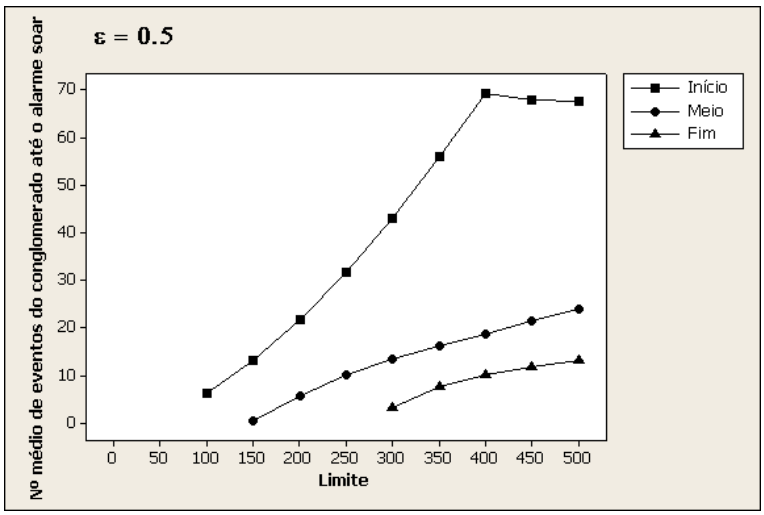

Figura 7: Número médio de eventos do conglomerado até o alarme soar: $\varepsilon = 0.5$

Observe que, independente do valor de ε, o número médio de eventos do conglomerado até o alarme soar é sempre maior para o conglomerado que começa no início, intermediário para conglomerado que começa no meio e menor para o conglomerado que começa no fim do estudo. Ou seja, quanto mais perto do final do estudo o conglomerado começa, maior é a capacidade de detecção do sistema. Isto significa que quanto maior é o acúmulo de informações (sob a hipótese nula) antes do surgimento do conglomerado, melhores são os resultados.

Note que as curvas da Figura 4 são retas quase paralelas. O mesmo acontece na Figura 5. Nestes dois casos o alarme soa em 100% das simulações nas três curvas. Nas Figuras 6 e 7 o padrão de retas paralelas é quebrado pela curva Início, onde o alarme soa apenas numa parcela das 1000 simulações.

Seja $N(t)$ o número de eventos até o tempo t e $H = \sum_{i=1}^{N(t)} \delta_i$ onde

$$\delta_i = \begin{cases} 1 & \text{se } i \in \text{ao conglomerado,} \\ 0 & \text{caso contrário.} \end{cases}$$

Seja τ o tempo de início do conglomerado. Então, a Figura 4 mostra que $E(H - \tau | H > \tau) \approx \alpha_{\varepsilon_1}(\tau) + \beta_{\varepsilon_1}(\tau) * (A - 350) \approx \alpha_{\varepsilon_1}(\tau) + 0.14666 * (A - 350)$. Isto é, para $\varepsilon \approx 0$, o tempo médio de espera para o alarme soar é função linear do limite A.

O valor $\beta_{\varepsilon_1}(\tau) \approx \beta_{\varepsilon_1} \approx 0.14666$ foi obtido tirando-se a média dos ajustes de mínimos quadrados nas três curvas da Figura 4. O valor da altura da reta quando $A = 350$ é dado por $\alpha_{\varepsilon_1}(\tau)$ que, pela estimação de mínimos quadrados é igual a $\alpha_{\varepsilon_1}(\text{Início}) = 48.2645$, $\alpha_{\varepsilon_1}(\text{Meio}) = 28.3071$, $\alpha_{\varepsilon_1}(\text{Fim}) = 6.3882$. Seja $\alpha_{\varepsilon_1}(\tau) = a + b * \tau = 55.289 - 0.16582 * \tau$, onde $a = 55.289$ e $b = -0.16582$ são as estimativas de mínimos quadrados. Então, a cada aumento de uma unidade no tempo τ de início do conglomerado, a altura da reta quando $A = 350$ diminui cerca de 0.16582 unidades. Assim, os $\beta_{\varepsilon_1}(\tau)$ são insensíveis ao momento de início do conglomerado: o impacto de mudar o limite A é de aumentar o tempo médio de espera pelo alarme soar e este aumento independe do momento em que o conglomerado surge.

Quando ε aumenta passando para o valor $\varepsilon_4 = 0.5$, o padrão de $E(H - \tau | H > \tau)$ não é mais tão simples. A quebra estrutural para limites A grandes para Início é devido à subestimação causada pela censura do experimento em 500 eventos. Ignorando esta quebra, o padrão é grosseiramente de três retas com inclinações dadas por

$\beta_{\varepsilon_4}(\text{Fim}) = 0.048298 < \beta_{\varepsilon_4}(\text{Meio}) = 0.064497 < \beta_{\varepsilon_4}(\text{Início}) = 0.17676$. Note que a diferença entre os valores $\beta_{\varepsilon_4}(\text{Fim})$ e $\beta_{\varepsilon_4}(\text{Meio})$ é pequena em relação às diferenças entre $\beta_{\varepsilon_4}(\text{Fim})$ e $\beta_{\varepsilon_4}(\text{Início})$ e entre $\beta_{\varepsilon_4}(\text{Meio})$ e $\beta_{\varepsilon_4}(\text{Início})$. Isto implica em mais sensibilidade de $E(H - \tau | H > \tau)$ a mudanças em A quando no Início.

6.2.2 Modelo Paramétrico

Como visto na seção 6.2.1, em várias combinações ε-limite testadas no cenário sem conglomerado, o número médio de eventos até o alarme soar não é uma estimativa fiel da realidade. Quando o alarme não soa em todas as simulações, este número é calculado com base apenas no número de amostras em que o alarme soa, sendo bem menor que seu valor verdadeiro. O ideal nestas situações seria corrigir esta estimativa adotando um modelo paramétrico. O mesmo ocorre para algumas estimativas do número médio de eventos do conglomerado até o alarme soar nos cenários com conglomerado. Porém, nestes casos este problema é bem menos acentuado.

Uma distribuição geralmente usada para modelar o tempo até a falha é a distribuição exponencial. No nosso contexto a falha ocorre quando o alarme soa. No entanto, para os nossos dados, o teste de aderência da distribuição exponencial não indica um bom ajuste deste modelo. Testou-se também vários outros modelos paramétricos, mas nenhum deles se mostrou adequado.

A capacidade de detecção do sistema é representada pelo número médio de eventos do conglomerado até o alarme soar, calculados nos cenários com conglomerado. Nestes cenários o problema da subestimação desta estatística é bastante brando.

O objetivo principal do cenário sem conglomerado não é avaliar a capacidade de detecção do sistema de vigilância, mas sim testar se o limite $A = B$ do método de Shiryayev-Roberts é adequado. Mesmo assim, para limites menores ou iguais a 500, o problema da subestimação do número médio de eventos até o alarme soar não é tão grave.

Assim, os esforços feitos no sentido de corrigir a estimativa do número médio de eventos até o alarme soar (cenário sem conglomerado) e do número médio de eventos do conglomerado até o alarme soar (cenários com conglomerado) se limitaram às tentativas de ajuste de um modelo paramétrico. Mesmo a subestimação destas

estatísticas sendo inegavelmente uma questão importante na avaliação do sistema de vigilância, não achamos que valesse a pena trabalhar neste ponto. No entanto, este problema pode servir como motivação para trabalhos futuros.

7 Aplicação aos dados de Burkitt em Uganda

Williams (1978) fornece dados com locais e datas de diagnóstico de casos de linfoma de Burkitt entre 1961 e 1975. A região de estudo foi o distrito de West Nile em Uganda (Figura 8). São 188 casos para os quais existem tanto a localização quanto a data de diagnóstico. Uma análise de conglomerado espaço-temporal desses dados (por Williams (1978) e por Bailey e Gatrell (1995), entre outros) encontrou evidência a favor de tal conglomerado para localizações particulares no tempo e espaço. O monitoramento da estatística de Knox local (por Rogerson (2001)), conduzido para as mesmas combinações de distâncias críticas temporais e espaciais usadas por Williams (1978), encontrou resultados semelhantes. O limite usado por Rogerson (2001) foi determinado adotando-se uma probabilidade de alarme falso de 0.1 durante o período de estudo de 188 observações.

Figura 8: Linfoma de Burkitt em West Nile, Uganda (aproximadamente 80 km $\times$ 170 km)

O sistema de vigilância aqui proposto foi aplicado aos dados de Burkitt com os mesmos valores de ε utilizados nas simulações. Os valores de ρ testados foram os mesmos usados por Rogerson (2001). O limite A adotado (161) foi determinado a partir de permutações aleatórias dos índices de tempo dos eventos originais, de forma que em aproximadamente 10% das vezes o alarme soa. Ou seja, manteve-se

a probabilidade de alarme falso de 0.1 sugerida por Rogerson (2001). Foram feitas 999 permutações aleatórias dos índices de tempo utilizando-se a média dos valores de ε testados (0.3) para a magnitude da mudança. Sejam s_0 e t_0 as distâncias críticas espacial e temporal usados por Rogerson (2001), respectivamente. Rogerson (2001) testou cinco valores para cada um desses parâmetros. O raio $\rho = 21.875$ km usado nas permutações é uma média ponderada dos valores de s_0, onde os pesos são dados pelo número de valores de t_0 em que encontrou-se evidência a favor de um conglomerado espaço-temporal em Rogerson (2001).

A Tabela 9 sumariza os resultados. Para $\rho = 2.5$, 5, 10 e 20 km, quanto menor o valor de ε, mais o alarme demora para soar. Já para $\rho = 40$ km acontece exatamente o oposto: quanto maior o valor de ε, mais o alarme demora para soar. Talvez isso ocorra porque 40 km é um valor muito grande para o raio crítico espacial.

ε	ρ (em km)	Número da observação em que o alarme soa
0.1	2.5	155
0.1	5	155
0.1	10	154
0.1	20	158
0.1	40	163
0.2	2.5	150
0.2	5	151
0.2	10	148
0.2	20	156
0.2	40	175
0.4	2.5	144
0.4	5	148
0.4	10	147
0.4	20	155
0.4	40	—
0.5	2.5	142
0.5	5	147
0.5	10	146
0.5	20	148
0.5	40	—

Tabela 9: Resultados do Sistema de Vigilância

A Figura 9 fornece um exemplo do comportamento da estatística R_n para $\varepsilon = 0.5$ e $\rho = 20$ km. Neste caso o alarme soa na observação 148 (Fevereiro de 1973). Um segundo alarme ocorre na observação 155 (Maio de 1973) e um terceiro alarme ocorre na observação 174 (Fevereiro de 1975), sendo que ambos são alarmes breves. O alarme soa pela quarta vez na observação 179 (Junho de 1975), permanecendo até o final do estudo (Outubro de 1975). Em Rogerson (2001), para esta mesma distância crítica espacial (20 km), o sinal de conglomerado espaço-temporal mais persistente começa na observação 146 (Janeiro de 1973).

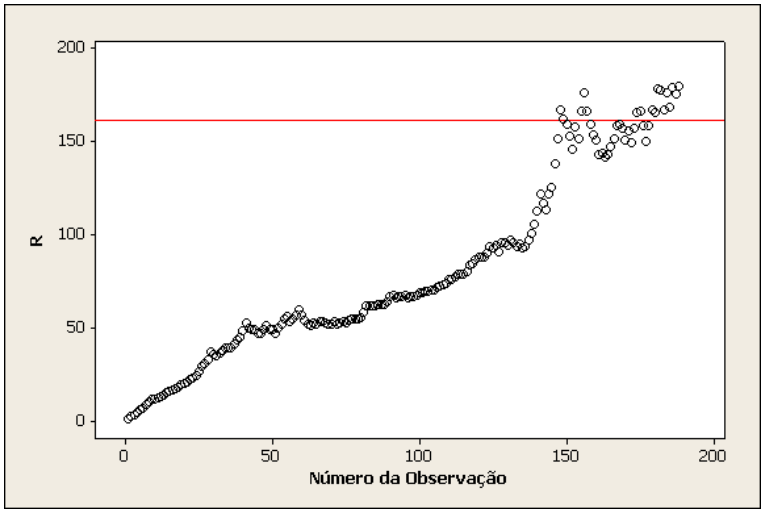

Figura 9: Aplicação do método proposto aos dados de Burkitt ($\varepsilon = 0.5$ e $\rho = 20$ km)

A Figura 10 compara o comportamento da estatística de teste proposta para os quatro valores de ϵ com $\rho = 20$ km, usando os dados de Burkitt. Até aproximadamente a observação de número 60, a estatística R_n é praticamente a mesma para os quatro valores de ϵ. A partir daí aparecem dois padrões diferentes: um seguido pelas curvas $\epsilon = 0.1$ e $\epsilon = 0.2$ e outro seguido pelas curvas $\epsilon = 0.4$ e $\epsilon = 0.5$.

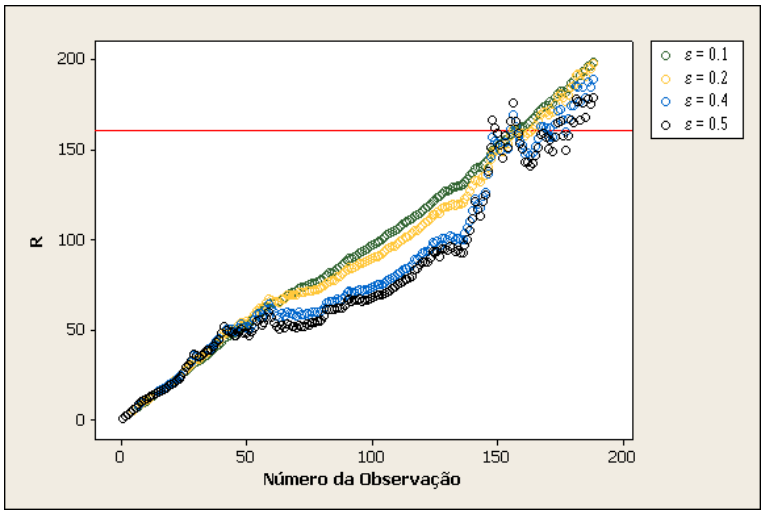

Figura 10: Comparação da estatística proposta para os quatro valores de ϵ ($\rho = 20$ km)

8 Considerações finais

De modo geral, o sistema de vigilância proposto é eficiente na detecção de conglomerados espaço-temporais. No entanto, alguns fatores influenciam o tempo que o sistema leva para tal detecção.

O parâmetro ε afeta consideravelmente o desempenho do método. Vimos que, com exceção do conglomerado que começa no início do estudo, a utilização de um valor de ε um pouco maior que o verdadeiro melhora a capacidade de detecção do sistema. Na prática o verdadeiro valor de ε é desconhecido. Especialistas podem sugerir valores de ε razoáveis para dados de sua área. Dependendo do tipo de dado, mudanças de magnitudes menores que determinado valor podem ser de pouco interesse. O ideal é testar mais de um valor para o parâmetro ε, incluindo sempre valores um pouco maiores que o valor que realmente se pretende testar. A incerteza sobre ε poderia ser descrita utilizando uma abordagem bayesiana em trabalhos futuros.

A escolha do limite do alarme é sem dúvida a parte em que deve-se ter um maior cuidado. Se possível, o melhor é utilizar mais de um limite. Atenção especial deve ser dada à limites muito altos, principalmente se o parâmetro ε também for alto. A utilização de tais limites deve ser evitada, pois pode fazer com que o alarme não soe ou demore muito para soar, mesmo na presença de um conglomerado. É importante lembrar que valores altos de ε levam a testes com $ARL^0 \gg A$.

O momento em que o conglomerado surge também tem um impacto significativo no desempenho do sistema de vigilância: quanto maior é o acúmulo de informações (sob a hipótese nula) antes do surgimento do conglomerado, melhores são os resultados. Ou seja, a capacidade de detecção do sistema aumenta à medida que o tempo de início do conglomerado se afasta do início do estudo. Assim, o usuário deve estar ciente de que, quando o conglomerado aparece logo no início do estudo, a capacidade de detecção do sistema é um pouco menor. Nestes casos, o alarme pode demorar mais para soar ou nem soar, principalmente para valores altos de ε e limite. Um possibilidade para tentar amenizar este problema seria utilizar um limite variável, que aumentasse com o passar do tempo. Assim, teríamos sinais progressivos de mudança, que poderiam ser representados em mapas de risco de alarme. Esta idéia poderia ser explorada em trabalhos futuros.

Neste trabalho o valor do raio crítico espacial ρ foi fixado. Acreditamos que a escolha deste parâmetro depende da natureza do processo, de forma que para um especialista não é difícil fornecer uma estimativa adequada. Entretanto, a influência deste parâmetro poderia ser avaliada em trabalhos futuros. Outra questão que também poderia ser explorada é a geometria do conglomerado. Aqui propomos um método que detecta conglomerados cilindricos, mas poderia-se considerar outras formas geométricas. Situações em que o número de conglomerados emergentes é maior que um, não avaliadas neste trabalho, também poderiam motivar trabalhos futuros.

9 Referências Bibliográficas

1. ASSUNÇÃO, R. e MAIA, A. (2005) A note on testing separability in spatial-temporal marked point processes. Artigo aceito pela *Biometrics*.

2. BAILEY, A. e GATRELL, A. (1995) Interactive Spatial Data Analysis. Logman, London.

3. JÄRPE, E. (1999) Surveillance of the interaction parameter of the Ising model. *Communs Statist. Theory Meth.*, 28, pp. 3009-3027.

4. KARLIN, S. e TAYLOR, H. M. (1975) A First Course in Stochastic Processes. 2nd Ed., Academic Press, New York.

5. KENETT, R. e POLLAK, M. (1996) Data-analytic aspects of the Shiryayev-Roberts control chart: surveillance of a non-homogeneos Poisson process. *Journal of Applied Statistics*, 23, pp. 125-137.

6. KNOX, G. (1964) The detection of space-time interactions. *Appl. Statist.*, 13, pp. 25-29.

7. KULLDORFF, M. (2001) Prospective time periodic geographical disease surveillance using a scan statistic. *J. R. Statisti. Soc. A*, 164, pp. 61-72.

8. LORDEN, G. (1971) Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42, pp. 1897-1908.

9. MEVORACH, Y. e 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, pp. 277-298.

10. MOUSTAKIDES, G. V. (1986) Optimal stopping times for detecting changes in distribution. *Annals of Statistics*, 14, pp. 1379-1387.

11. PAGE, E. S. (1954) Continuous inspection schemes. *Biometrika*, 41, pp. 100-115.

12. POLLAK, M. (1985) Optimal detection of a change in distribution. *Annals of Statistics*, 13, pp. 206-227.

13. POLLAK, M. (1987) Average run lengths of an optimal method of detecting a change in distribution. *Annals of Statistics*, 15, pp. 749-779.

14. POLLAK, M. e SIEGMUND, D. (1985) A diffusion process and its application to detecting a change in the drift of Brownian motion. *Biometrika*, 72, pp. 267-280.

15. RAUBERTAS, R. (1989) An analysis of disease surveillance data that uses geographical locations of the reporting units. *Statist. Med.*, 8, pp. 267-271.

16. RITOV, Y. (1990) Decision theoretic optimality of the Cusum procedure. *Annals of Statistics*, 18, pp. 1464-1469.

17. ROBERTS, S. W. (1966) A comparison of some control chart procedures. *Technometrics*, 8, pp. 411-430.

18. ROGERSON, P. (1997) Surveillance systems for monitoring the development of spatial patterns. *Statist. Med.*, 16, pp. 2081-2093.

19. ROGERSON, P. (2001) Monitoring point patterns for the development of space-time clusters. *J. R. Statist. Soc A*, 164, pp. 87-96.

20. SHIRYAYEV, A. N. (1963) On the detection of disorder in a manufacturing process. *Theory of Probability and its Applications*, 8, pp. 247-265.

21. SIEGMUND, D. O (1985) Sequential Analysis: Tests and Confidence Intervals. New York: Springer.

22. SONESSON, C. e BOCK, D. (2003) A review and discussion of prospective statistical surveillance in public health. *J. R. Statist. Soc. A*, 166, Part 1, pp. 5-21.

23. WILLIAMS, E. H., SMITH, P. G., DAY, N. E., GESER, A., ELLICE, J. e TUKEI, P. (1978) Space-time clustering of Burkitt's lymphoma in the West Nile District of Uganda: 1961-1975. *Br. J. Cancer*, 37, pp. 109-122.

24. YAKIR, B. (1994) Optimal detection of a change in distribution when the observations are independent. *Technical Report*, Department of Statistics, Hebrew University of Jerusalem.