VII Simpósio Brasileiro de Geoinformática, Campos do Jordão, Brasil, 20-23 novembro 2005, INPE, p. 281-291.

Sistema de vigilância para detecção de interações espaço-tempo de eventos pontuais

Taynãna C. Simões, Renato M. Assunção

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

Abstract. *The System of Surveillance proposed consists of the monitoring of a group of point events, in order to detecting quickly the formation of space-time clustering, as soon as begin to appear. It bases on the analysis of prospective statistical methods (on-line studies). In this article, it is to adapt the technique presented by Rogerson (2001), with the intention of detecting the existent cluster, through the visualization of the position in that was formed. Besides, we believed this mechanism can to avoid that observations that don't belong to the cluster make the alarm to sound falsely.*

Resumo. *O Sistema de Vigilância proposto consiste no monitoramento de um conjunto de eventos pontuais, a fim de detectar rapidamente a formação de conglomerados espaço-tempo, assim que estes começam a surgir. Baseia-se na análise de métodos estatísticos prospectivos (Estudos on-line). Neste artigo, objetiva-se adaptar a técnica apresentada por Rogerson (2001), com o intuito de detectar o conglomerado existente, através da visualização da posição em que foi formado. Além disso, acredita-se que a técnica proposta evite que observações que não pertençam ao conglomerado façam o alarme soar falsamente.*

1. Introdução

Compreender a distribuição espacial de dados oriundos de fenômenos ocorridos no espaço constitui hoje um grande desafio para a elucidação de questões centrais em diversas áreas do conhecimento. Tais estudos vêm se tornando cada vez mais comuns, devido à disponibilidade de Sistemas de Informação Geográfica (SIG) de baixo custo e com interfaces amigáveis, que
permitem a visualização espacial de fenômenos sob estudo.

A ênfase da análise espacial é mensurar propriedades e relacionamentos, levando em conta a localização espacial do fenômeno em estudo de forma explícita. Usualmente, o processo de modelagem é precedido de uma fase de análise exploratória, associada à apresentação visual dos dados sob forma de gráficos e mapas e a identificação de padrões de dependência espacial no fenômeno em estudo. No caso de análise de padrão de pontos, o objetivo de interesse é a própria localização espacial dos eventos.

Detectar aglomerados espaço-tempo de forma rápida, eficiente e em tempo real é uma necessidade em várias áreas do conhecimento. Nas áreas de saúde e social, principalmente, técnicas e ferramentas que possibilitem fazê-lo têm uma importância particular, pois permitem que ações preventivas ou de controle sejam realizadas de forma eficiente.

O Sistema de Vigilância em estudo consiste no monitoramento de um conjunto de eventos pontuais, a fim de detectar rapidamente a formação de conglomerados espaço-temporais, assim que estes começam a surgir. Baseia-se na análise de métodos estatísticos prospectivos, que são aqueles estudos em que há uma análise repetida de dados acumulados ao longo do tempo seqüencialmente, com o objetivo de detectar rapidamente qualquer mudança que ocorra na série. Este tipo de estudo também é chamado de caso *on-line*.

Utilizando a estatística de detecção de conglomerados espaço-tempo de Knox e métodos de Soma Acumulada (CUSUM), Rogerson (2001) propõe um sistema de vigilância que detecte conglomerados geograficamente ativos. No entanto, um problema observado é que pode ser que a soma acumulada esteja tão próxima do limiar a partir do qual o processo passe a estar fora de controle, que uma observação não pertencente ao conglomerado, faça o alarme soar falsamente. Pretende-se, desta forma, tentar reformular a técnica apresentada por Rogerson (2001), com o intuito de evitar que estas observações, que não pertençam ao conglomerado, façam o alarme soar quando não deveria. No entanto, neste artigo, o principal objetivo se limitará a apresentar um mecanismo de detecção do conglomerado, através da visualização da localização em que foi formado.

2. Vigilância Estatística

Basicamente, um Sistema de Vigilância monitora mudanças quando novas observações tornam-se disponíveis, no decorrer do estudo. Por sua vez, vigilância estatística significa um monitoramento *on-line* de um processo estocástico X = {X(t); t=1, 2, ...} com o objetivo de detectar uma mudança importante no processo, em um tempo desconhecido τ, tão rápida e precisamente possível.

A cada ponto de tempo s de decisão, deve-se discriminar entre dois estados no sistema monitorado: sob-controle (D(s)) e fora-de-controle (C(s)). Para que isso ocorra, utilizam-se as observações acumuladas $X_s=\{X(t);\ t \leq s\}$ para formar conjuntos alarme A(s), tais que se $X_s \in A(s)$, essa é uma indicação que o processo está no estado C(s) e um alarme é soado. Usualmente, isso é feito usando uma função alarme $p(X_s)$ e um limite de controle g(s). O tempo de um alarme t_A é escrito como

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

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

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

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

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

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

3. Sistema de Vigilância – Rogerson (2001)

Rogerson (2001) propõe um sistema de vigilância no qual combina métodos de Soma Acumulada (CUSUM) com uma estatística de detecção de conglomerados espaço-tempo para um conjunto de dados pontuais (Teste de Knox). A aproximação conta com uma estatística de Knox local que é útil em análises retrospectivas para detectar quando e onde a interação espaço-tempo ocorre.

3.1. Teste de Knox

O teste de Knox é um teste baseado na contagem do número de pares de eventos que ocorrem dentro de intervalos críticos pré-especificados de tempo (T) e distância (D). O mecanismo pode ser ilustrado através da Figura 1.

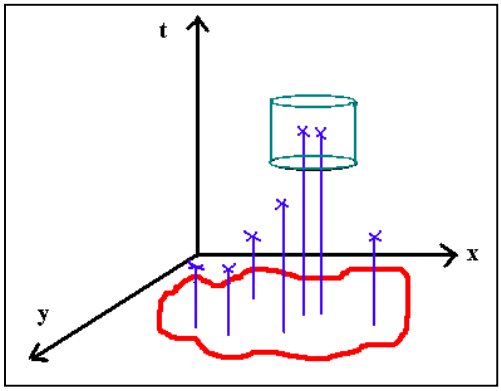


Figure 1. Visualização do mecanismo de contagem de pares próximos no Teste de Knox.

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

Dado n pontos localizados no tempo e no espaço, existem $\binom{n}{2} = \frac{n(n-2)}{2}$ pares distintos que podem ser divididos em uma matriz 2 *x* 2, indicando os pares que estão próximos no espaço e/ou no tempo. A estatística N_{st} excederá seu valor esperado, quando os pontos próximos no espaço (tempo) forem mais próximos que o esperado no tempo (espaço).

No contexto dos métodos prospectivos de detecção de conglomerados, uma proposta possível é a de que fosse utilizado um novo teste a cada vez que uma nova observação fosse acrescentada no banco de dados. No entanto, este procedimento resultaria em um erro do Tipo I muito maior que o especificado, levando a indicações falsas de interação espaço-tempo. Assim, uma alternativa para este tipo de problema, segundo Rogerson (2001), seria utilizar métodos de Soma Acumulada (CUSUM).

3.2. Métodos de Soma Acumulada (CUSUM)

Os métodos de Soma Acumulada são muito usados para monitorar processos industriais com o objetivo de detectar rapidamente uma mudança indesejável no processo. É utilizado, desta forma, no controle de um processo estatístico seqüencial de uma variável para potenciais desvios da média esperada.

Seja x_t a observação no tempo t, $X_t \sim N(\mu, \sigma^2)$, $\sigma^2 < \infty$. Assumindo que não há correlação na série de observações, a soma acumulada no tempo t é dada através da relação:

$$S_t = \max(0, S_{t-1} + x_t - \mu - k\sigma); \;\; S_0 = 0$$

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

Os parâmetros h e k são expressos em termos do desvio-padrão das observações. E a escolha de h, depende do número médio de observações até que ocorra uma mudança, sob a hipótese de que o processo esteja sob controle (ARL_0). Espera-se que o ARL seja longo, quando o processo está sob controle e curto após o processo ter experimentado uma mudança.

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

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

3.3. Estatística de Knox Local

Quando o teste de Knox global é significativo, freqüentemente é de interesse encontrar os pares específicos responsáveis por esta interação. Desta forma, implementa-se uma estatística de Knox Local, que é um teste retrospectivo.

Sejam $n_s(i)$ o número de eventos que são próximos à observação i no espaço, $n_t(i)$ o número de eventos que são próximos no tempo a i e $n_{st}(i)$ ser o número de eventos que são conjuntamente próximos no tempo e espaço à observação i.

Para encontrar a distribuição de $N_{st}(i)$, sob a hipótese nula de não interação espaço-tempo, supõe-se que cada permutação aleatória dos índices dos tempos e posições fixas é igualmente provável. Seja $n_t^j(i)$ o número de pontos que estão próximos no tempo do evento i quando a este é associado o j-ésimo valor do tempo. Assim, para uma dada permutação de tempos através das localizações espaciais, mostra-se que a distribuição de $N_{st}(i)$ é hipergeométrica com parâmetros n-1, $n_s(i)$ e $n_t^j(i)$.

Alternativamente, uma aproximação Normal pode ser utilizada Uma implementação prévia do método evidenciou que a variância da variável aleatória $N_{st}(i)$ apresentada no artigo do Rogerson (2001) está incorreta.

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

$$z_i = \frac{n_{st}(i) - E\{N_{st}(i)\} - 0{,}5}{\sqrt{Var\{N_{st}(i)\}}} \approx N(0,1)$$

Assim, em estudos prospectivos, deseja-se verificar quando há evidência de aumento na interação espaço-tempo. Portanto, a cada nova observação, é de interesse comparar qualquer aumento na estatística de Knox com o quê seria esperado sob a hipótese nula.

Suponha que foi observada a informação em i-1 casos. Para uma nova observação i, sejam $n_s(i)$, $n_t(i)$ e $n_{st}(i)$ o número de i-1 vizinhos das observações anteriores, que estão próximos no espaço, no tempo e em ambos, respectivamente. Pode-se comparar o valor da estatística de Knox após o caso i (denotado K_i) com o valor que seria esperado sob a hipótese nula, condicionado ao valor da estatística de Knox após a observação i-1 e valores observados de $n_s(i)$ e $\{n_t^j(i); j=1,...,i\}$. Logo a estatística escore é dada por:

$$z_i = \frac{K_i - E\{K_i \mid K_{i-1}, n_s(i), n_t(i)\} - 0{,}5}{\sqrt{Var\{K_i \mid K_{i-1}, n_s(i), n_t(i)\}}}$$

A informação obtida nesta comparação de K_i com sua esperança condicional é unicamente para especificar a contribuição da observação i na estatística de Knox. Na prática, utiliza-se a estatística escore z_i da última equação, com métodos de Soma Acumulada, tal que $n_{st}(i) = K_i - K_{i-1}$, com as mesmas esperança e variância da estatística de Knox global, substituindo n por i.

A Soma Acumulada correspondente à observação i é dada por:

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

Logo, a soma excederá seu valor crítico ($S_i > h$) quando observações que apresentarem interações espaço-tempo começarem a acumular.

4. Problemas com a metodologia de Rogerson (2001)

Um problema observado com o método desenvolvido por Rogerson (2001) é que pode ser que a soma acumulada esteja tão próxima do limiar h, que uma observação que não pertença ao conglomerado, faça o alarme soar falsamente. Além disso, a técnica apresentada não permite que o conglomerado seja localizado espacialmente.

Verifica-se que esta situação possa acontecer perfeitamente, dado que eventualmente, observações que não pertençam ao conglomerado possam exceder k desvios padrões da média. Desta forma, pretende-se reformular a técnica de Soma Acumulada apresentada, com o intuito de evitar que estes "ruídos" (observações que não pertencem ao conglomerado) façam o alarme soar, além de poder identificar o conglomerado.

A idéia proposta é observar a Soma Acumulada como uma superfície resultante (Superfície Acumulada), na qual cada evento contribui com uma superfície de Kernel (Estimação por densidade de Kernel). A partir desta técnica, a superfície i+1 seria formada por muitas curvas, de forma que na posição em que estivesse o conglomerado, se formaria uma curva muito maior, resultante das contribuições dos Kernels de cada um dos eventos. A cada período de tempo observado, a curva pode ser calculada como:

$$S_{i+1}(x,y) = \max(0, S_i + z_i K\{(x,y) - (x_i,y_i)\};\ S_0(x,y) = 0$$

Seja N_{st}, o número de eventos conjuntamente próximos no espaço e no tempo que, como mostrado por *Rogerson* (2001), pode ser padronizada, de forma que $z_i \sim N(0,1)$, sob a hipótese de aleatoriedade. Assim, aplicando-se a função de kernel à estatística de interação espaço-tempo padronizada (escore z_i), a contribuição de cada observação para a superfície pode ser obtida através de uma curva de kernel, centrada em uma posição arbitrária do evento i e "altura" z_i, como mostrado na Figura 2.

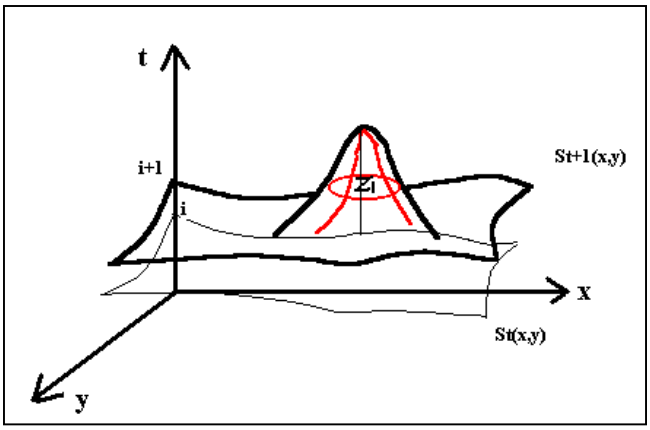


Figure 2. Visualização da técnica de superfície acumulada para detecção de conglomerados espaço-tempo.

Verifica-se, desta forma, que à medida que uma nova observação é acrescentada ao banco de dados, a Superfície Acumulada passa a ser obtida através da soma das superfícies anteriores com a contribuição da estimativa da densidade de kernel avaliada nesta nova observação coletada e padronizada.

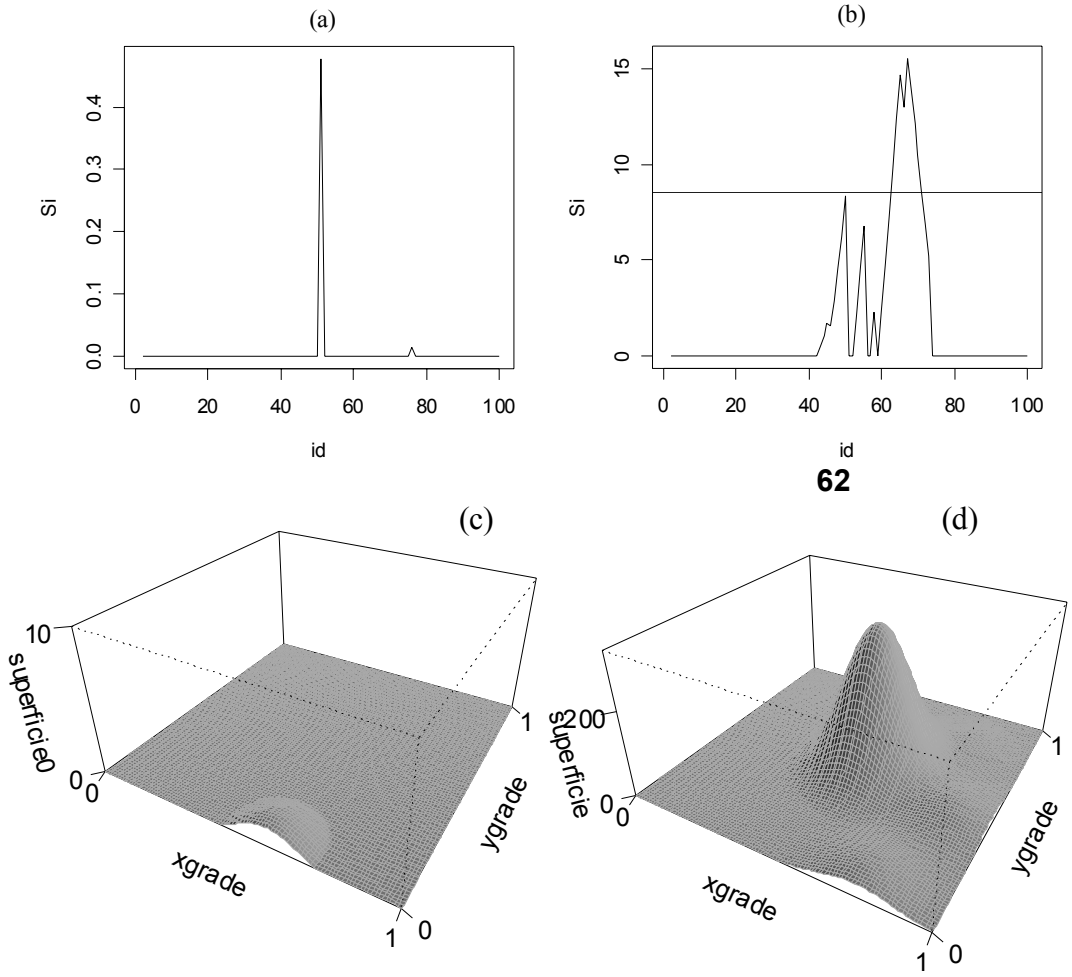
Acredita-se que, desta forma, o alarme soe "menos" falsamente se comparado ao modelo proposto por *Rogerson* (2001), no sentido que observações que não pertençam ao conglomerado contribuirão com uma curva na sua localização de origem, não influenciando efetivamente na curva formada por elementos do conglomerado. A superfície formada deve aparecer com vários "calombos" resultantes das curvas de kernel formadas por cada observação, de forma que no intervalo de tempo em que se encontra o conglomerado, observasse-ia um "calombo" muito maior se comparado aos demais, permitindo assim a localização espacial do conglomerado.

Obs: O método de estimação de densidade por kernel consiste em estimar a densidade de uma distribuição em pontos determinados, usando pontos empiricamente observados, por meio de interpolação. A intuição é que a função de Kernel é composta por uma soma ponderada dos pontos observados, em que o fator de ponderação cai rapidamente à medida que cada *x* se afasta de X_i. Existe ainda, um parâmetro τ que

regula o grau de suavidade das curvas de kernel, chamado *bandwidth*. Neste artigo, foram utilizadas a função de kernel Gaussiana e o *bandwidth* ótimo fornecido por Härdle (1999).

5. Simulando a Superfície Acumulada no software R

Para se ter uma idéia do comportamento da Superfície Acumulada, no caso bidimensional, foram observados 100 pontos com coordenadas (x_k,y_k,t_k), das quais 80 observações tinham coordenadas x e y geradas de uma distribuição uniforme U(0;1) e valores do tempo em que o evento ocorreu t gerados de uma uniforme U(0;10). Outras 20 observações foram geradas, com coordenadas x e y de uma uniforme U(0.5;0.6) e tempo t de uma uniforme U(5;6), de forma que formassem um conglomerado nesta região, uma vez que estão próximos no espaço e no tempo. Foram pré-estabelecidos, os parâmetros críticos de D=0.1 e T=1.0 para o cálculo das estatísticas de Knox. Aplicando-se a metodologia da Superfície Acumulada, têm-se as figuras seguintes referentes a uma das várias situações observadas:



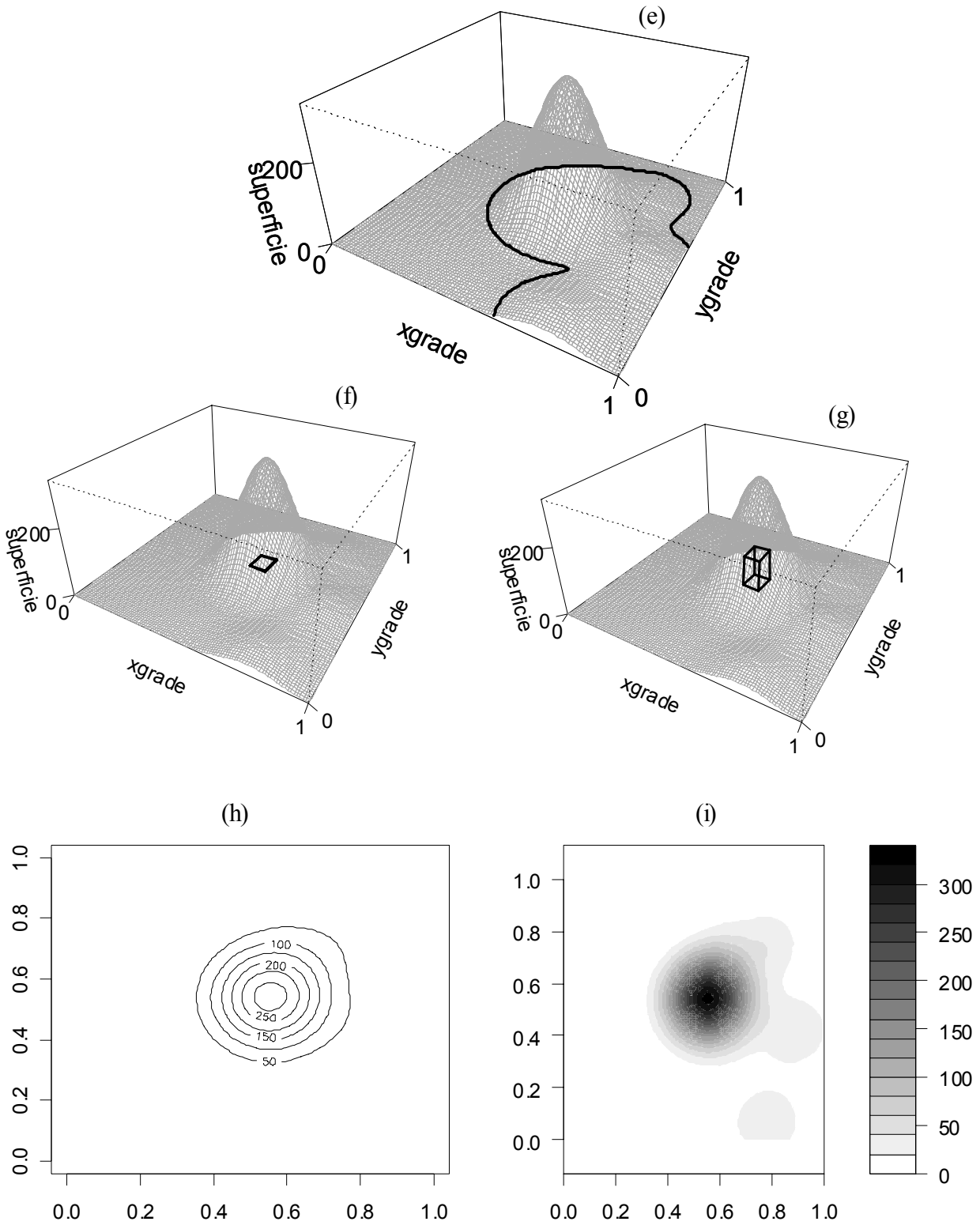


Figure 3. Simulando a técnica de superfície acumulada para detecção de conglomerados espaço-tempo.

Da forma como o conglomerado foi gerado, inclui as observações 40-42,44-45, 47-50,53-55,58,60-65 e 67.

As Figuras 3 (a) e 3 (b) mostram o mecanismo do Rogerson (2001), quando não há a presença de conglomerado e por isso nenhum alarme é soado, e quando o alarme soa na observação 62, respectivamente. A Figura 3 (c) mostra a técnica de Superfície Acumulada sob H_0, ou seja, sem a presença de conglomerados. Verifica-se apenas uma pequena elevação resultante da contribuição dos kernels. Por sua vez, as figuras seguintes fornecem um mecanismo de visualização do conglomerado, no momento em que o alarme é soado no sistema do Rogerson (2001). A maior curva se forma exatamente na posição em que o conglomerado verdadeiro foi criado (entre as posições 0.5 e 0.6 dos eixos x e y). Para este conjunto de dados, o alarme foi soado na observação 50 na Superfície e, portanto antes do método do Rogerson.

Uma maneira de avaliar o método de monitoramento é verificando se o alarme foi soado muito antes/depois do conglomerado verdadeiro ter surgido e através da probabilidade de alarmes falsos, sob H_0. Para isso, foram feitas 1000 simulações do mesmo cenário acima.

Seja d_i o tempo entre o primeiro alarme e o início do conglomerado. Se $d_i<0$, a proporção destes valores dentre as simulações equivale à probabilidade de se ter um alarme falso, entre o início do processo e o tempo do primeiro alarme. Analogamente, se $d_i>0$, é possível calcular o número esperado de observações até que um alarme ocorra, dado que o processo esteja fora de controle, ou seja, existe um conglomerado (ARL_1).

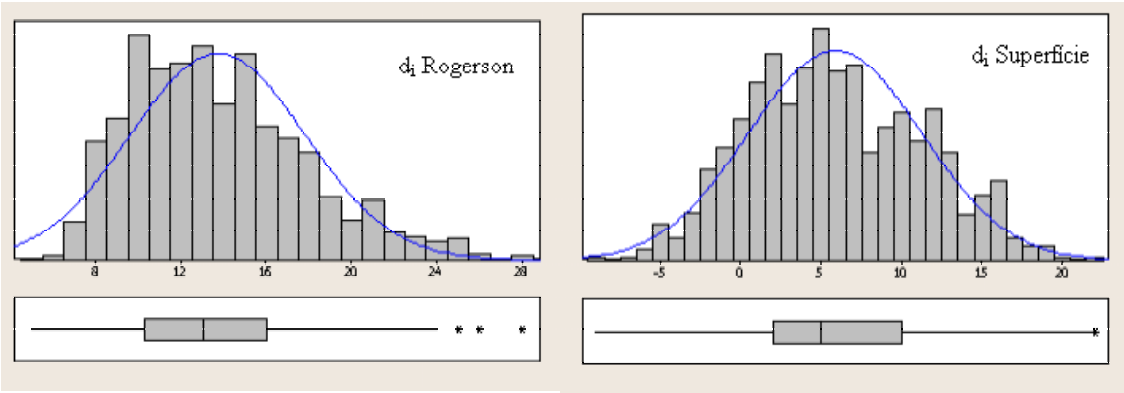


Figure 4. Histogramas dos d_i's nos dois métodos.

Tabela 1. Informações relevantes obtidas nas simulações.

	Informações sobre o di						
Método	Média	Desvio-padrão	Min	1ºQuartil	Mediana	3ºQuartil	Max
Rogerson	13,722	4,104	5,000	10,250	13,000	16,000	28,000
Superfície	5,868	5,394	-9,000	2,000	5,000	10,000	22,000

Método	P(Alarme falso)	ARL1	Teste Mann-Whitney
Rogerson	0	13,72	w = 1253217,0
Superfície	0,115	6,97	p-valor=0,000

Através da Figura 4 e Tabela 1 observa-se que, em geral, o alarme soa mais rápido no método da superfície (Mann-Whitney significativo). No entanto, a probabilidade de alarmes falsos (Sob H_0) é muito maior que o esperado, se comparado ao Rogerson. Acredita-se que este impacto possa ser devido ao fato do limiar utilizado ser inadequado na Superfície.

Além do limiar apresentado por Rogerson, foram implementados mais dois tipos empiricamente observados, sob H_0: o primeiro (lim1) foi obtido como o percentil 90 dos valores máximos das superfícies, em cada etapa da soma acumulada. Enquanto o segundo foi obtido pela média dos percentis 90 dos valores da superfície, em cada etapa da soma acumulada. No entanto, nenhuma conclusão foi tirada a respeito de um valor ótimo, como mostrado na Tabela 2.

Tabela 2. Informações obtidas nas simulações (100), com outros limiares.

Método	P(Alarme falso)	ARL1
Rogerson	0	13,79
Super(h roger)	0,31	7,47
Super(lim 1)	0,87	6,85
Super(lim 2)	0,97	6,38

6. Conclusões Preliminares

Verifica-se que o método proposto de Superfícies Acumuladas, com base no Sistema de Vigilância do Rogerson (2001), é um excelente identificador do conglomerado, uma vez que detecta a posição espacial do mesmo, através da visualização das representações gráficas das superfícies. Contribuição esta de grande relevância, uma vez que, na prática, a localização do conglomerado possa levar a várias tomadas de decisão importantes.

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

7. Referências

Rogerson, P.A..(2001) "Monitoring point patterns for the development of space-time clusters". Jornal Royal Statistical Society (2001) 164, Part 1, 87-96. University at Bualo, USA.

Siegmund, D., O..(1985) "Sequential Analysis: Tests and Condence Intervals". New York: Springer.

Härdle,W..(1990) "Smoothing Techniques". Louvain-La-Neuve.

Sonesson, C.; Bock, D..(2002). "A review and discussion of prospective statistical surveillance in public health. Göteborg University, Sweden. Jornal Royal Statistical Society
(2003) 166, Part 1, pp 5-21.

Diggle,P.J..(1983). "Statistical Analysis of Spatial Point Patterns". Academic Press Inc. Londres.

Frisen,M..(2003). "Statistical surveillance. Optimality and methods". International Statistical Review, *71*, 403-434.

Camara,G.; Monteiro,A.,M.; Fuks,S.; Camargo,E.; Felgueiras. (2001). "Análise Espacial". INPE.

Knox, E. G..(1964). "The detection of space-time interactions".. Appl. Statist.. 13, 25-29.

Montgomery, D. C..(2000). "Introduction to Statistical Quality Control".. 4th Edition, New York : John Wiley, 2000.