Revista da Estatística UFOP, Vol III(2), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS

DESENVOLVIMENTO DE UMA MEDIDA DE ASSOCIAÇÃO ENTRE ESPAÇO E TEMPO

Thais Correa[1], Bráulio Veloso[2], Andrea Iabrudi[3]

Resumo: *O sistema STCD (Spatio-Temporal Cluster Detection) é um método para a detecção de conglomerados espaço-tempo em processos pontuais, de forma prospectiva e eficiente. Este método acomoda variações puramente temporais ou espaciais, podendo ser utilizado em diversas aplicações. No entanto, o sistema STCD não é capaz de detectar conglomerados simultâneos (múltiplos). Neste trabalho nós propomos uma adaptação deste sistema para a identificação de conglomerados múltiplos, sem a necessidade de prévia especificação do número de conglomerados. Resultados iniciais fornecem evidência de um desempenho desta adaptação para dois conglomerados simultâneos, com uma alta taxa de identificação alta e um atraso razoavelmente pequeno.*

Palavras-chave: *Conglomerado espaço-tempo, Conglomerados simultâneos, Sistema de vigilância.*

Abstract: *Here we propose a technique to efficiently detect multiple emergent clusters in space-time point process. There is no need to specify in advance the number of clusters, and the method accommodates purely temporal or spatial variations. Preliminary results provide evidence of good performance of the proposed method for two simultaneous clusters.*

Keywords: *Simultaneous clusters, Space-time cluster, Surveillance systems.*

1 Introdução

Métodos para a detecção prospectiva de conglomerados espaço-tempo, chamados de sistema de vigilância, são de extrema importância para aplicações em que respostas rápidas são crucias. Estas aplicações incluem saúde pública [5, 13], controle de tráfego [6], criminologia, mudança de comportamento em redes sociais [11, 16], dentre outras. Ao contrário das abordagens retrospectivas para detecção de conglomerados, o objetivo de um sistema de vigilância é detectar conglomerados "vivos", propiciando uma intervenção eficaz. A epidemiologia é tradicionalmente uma área de grande interesse em sistemas de vigilância, uma vez que a detecção de surtos de doenças é uma tarefa essencial, exigindo reação muito rápida dos agentes públicos.

Dados espaço-tempo estão cada vez mais disponíveis, uma vez que os procedimentos de georeferenciamento são frequentemente utilizados atualmente [15]. A comunidade de Sistemas de Informação Geográfica têm proposto métodos para tratar diferentes aspectos: armazenamento,

[1]DEST - UFMG. e-mail: *thaisrc@ufmg.br*
[2]DCC - UFMG. email: *brauliodcc141@ufmg.br*
[3]Jasper Design Automation Brasil.

recuperação de informação e métodos de visualização específicos para dados espaço-tempo [14]. A comunidade estatística por sua vez têm estudado métodos, tanto prospectivos quanto retrospectivos, para a detecção de conglomerados espaciais e espaço-tempo ([2, 17, 21]). A estatística scan espacial [8] é a principal técnica para detecção de conglomerados espacias. Em [9] o autor propõe um extensão dessa estatística para o contexto espaço-tempo. Diversas outras propostas utilizam variações da estatística scan espaço-Tempo([18, 19, 10]).

Mais recentemente, a detecção de conglomerados simultâneos têm sido alvo de diversos estudos. Em [20] a estatística scan espacial é ajustada para a presença de outros conglomerados na região de estudo. Em [12] os autores consideram a existência de conglomerados espacias simultâneos diretamente na hipótese alternativa. Uma estratégia baseada em trajetórias é adotada por [4] para detectar conglomerados espaço-tempo simultâneos, de formato arbitrário. Os dois primeiros métodos citados acima são puramente espacias. O terceiro detecta conglomerados espaço-tempo, porém de forma retrospectiva.

Este trabalho propõe um método eficiente para detecção prospectiva de conglomerados espaço-tempo. O método apresentado é um extensão do sistema proposto por [1], projetado para a detecção de um único conglomerado. A proposta, avaliada em diferentes cenários contendo dois conglomerados simultâneos, têm resultados promissores. Assume-se que os eventos são gerados por um processo de Poisson, com taxas que podem variar tanto no espaço quanto no tempo. Não é necessário especificar as intensidades de eventos marginais (espacial e temporal), e nem o número de conglomerados a serem detectados.

2 Material e métodos

A seção 2.1 traz uma descrição resumida do método proposto por [1] para a detecção prospectiva de um único conglomerado. A extensão para conglomerados simultâneos é apresentada na seção 2.2. As métricas utilizadas para avaliar a eficácia desta extensão são definidas na seção 2.3.

2.1 Detecção prospectiva de um único conglomerado espaço-tempo

Considere um processo pontual observado na região tri-dimensional $A \times (0, T]$, onde A representa o espaço e $(0, T]$ o tempo. O método STCD (Space Time Cluster Dection), proposto por [1], busca um conglomerado espaço-tempo vivo de formato cilíndrico. O raio da base circular, ρ, deve ser especificado pelo usuário. O método consiste em monitorar uma estatística que não depende das intensidades marginais de eventos no espaço e no tempo. Quando esta estatística supera um limiar fixo o método soa um alarme, e o conglomerado é identificado.

Suponha eventos observados sequencialmente nos tempos $t_1, t_2, \ldots$. As coordenadas espaciais do evento observado no tempo t_i são (x_i, y_i). Seja $C_{k,n} \in A \times (0, T]$ um cilindro de base circular cujo centro é dado por (x_k, y_k). A altura de $C_{k,n}$ é dada por $t_n - t_k$, onde t_n é o tempo do último evento observado. Considerando t_n como o tempo atual, $C_{k,n}$ é um cilindro vivo, um vez que ele atinge o tempo t_n. Seja $N(C_{k,n})$ o número de eventos dentro do cilindro $C_{k,n}$, e assuma $N(C_{k,n}) \sim \text{Poisson}(\mu(C_{k,n}))$.

Dado que não existe interação espaço-tempo (ou equivalentemente, não existe conglomerado espaço-tempo), a intensidade de eventos $\lambda(x, y, t)$ é separável. Neste caso $\lambda(x, y, t)$ pode ser escrita como o produto das intensidades marginais no espaço $(\lambda_s(x, y))$ e no tempo $(\lambda_t(t))$:

$$\lambda(x, y, t) = \mu\, \lambda_s(x, y)\, \lambda_t(t), \qquad \mu = \int_A \int_{(0,T]} \lambda(x, y, t)\, dt\, dx\, dy.$$

Se um conglomerado $C_{k,n}$ inicia no tempo t_k, a intensidade muda por uma constante $\varepsilon > 0$ tal que

$$\lambda(x, y, t) = \mu\, \lambda_s(x, y)\, \lambda_t(t)\, (1 + \varepsilon I_{C_{k,n}}(x, y, t))\,,$$

onde $I_{C_{k,n}}$ é uma função indicadora para $(x, y, t) \in C_{k,n}$ e ε representa o aumento na intensidade de eventos dentro do conglomerado. O aumento ε é um parâmetro de entrada do método, especificado pelo usuário.

Seja L_∞ a verossimilhança do processo de Poisson espaço-tempo quando não existe conglomerado e seja L_k a verossimilhança deste mesmo processo quando existe um conglomerado $C_{k,n}$, ambas para n eventos observados. A estatística de teste é a soma da razão destas verossimilhanças sob todas as possibilidades para o cilindro $C_{k,n}$:

$$R_n = \sum_{k=1}^{n} \frac{L_k}{L_\infty} = \sum_{k=1}^{n} \Lambda_{k,n} = \sum_{k=1}^{n} (1+\varepsilon)^{N(C_{k,n})} \exp(-\varepsilon\, \mu(C_{k,n})) \,.$$

O método usa uma estimativa não paramétrica para a média $\mu(C_{k,n})$. Assumindo que $\lambda(x, y, t)$ é separável, esta quantidade é estimada por:

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

onde $N(\ B_{k,\rho} \times (0, t_n]\ )$ é o número de eventos dentro da base circular do cilindro $C_{k,n}$ independente do tempo, $N(\mathcal{A} \times (t_k, t_n])$ é o número de eventos entre os tempo t_k e t_n independente do espaço e n é o número total de eventos observados.

Quando a estatística de teste supera um limiar L, o método soa um alarme, indicando que existe evidência empírica de um conglomerado vivo. Como esta estatística é uma soma sob todos os possíveis conglomerados vivos, a estimativa do conglomerado é aquele com maior contribuição para a estatística. Ou seja: se $R_n > L$, então $\Lambda_{k*,n} = \max\{\Lambda_{k,n}, 1 \leq k \leq n\}$ e o conglomerado estimado é o cilindro $C_{k*,n}$.

O limiar L deve ser especificado pelo usuário conforme sua tolerância a alarmes falsos. Assuma que o conglomerado inicia no tempo t_k. Se a estatística supera o limiar L em um tempo $t < t_k$, então o método soa um alarme falso. O método soa um alarme motivado quando o limiar L é superado em um tempo $t > t_k$. O controle de alarmes falsos é feito especificando-se o limiar L igual ao valor desejado para o número médio de eventos até um alarme falso (B). Isto garante que, em média, o usuário irá esperar pelo menos B eventos até que o alarme soe falsamente [7].

2.2 Extensão para conglomerados espaço-tempo simultâneos

Suponha agora que o método descrito acima está sendo usado e o alarme soa. Isto indica que existe evidência suficiente de que existe um conglomerado vivo. Mas neste caso, existe evidência suficiente também para um segundo conglomerado vivo? E para um terceiro? Estas questões são respondidas com uma extensão do método para o caso em que mais de um conglomerado iniciam no mesmo tempo t_k. O raio espacial ρ e o aumento na intensidade (ε) são os mesmo para todos os conglomerados. A extensão proposta, chamada de STCD-Sim (STCD Simultâneo), pode ser aplicada para qualquer número de conglomerados simultâneos, uma vez que este número não precisa ser previamente especificado.

Considere que o método STCD soa um alarme no tempo t_n. A estimativa do conglomerado é $C_{k*,n}$. A extensão proposta consiste em excluir o excesso de eventos dentro do cilindro $C_{k*,n}$ de maneira aleatória e reaplicar o método na base de dados reduzida. Este excesso de eventos é dado por: $\Delta(C_{k*,n}) = N(C_{k*,n}) - \hat{\mu}(C_{k*,n})$.

Após excluir o excesso de eventos, avalia-se a estatística de teste no tempo t_n para a base de dados reduzida. A estatística correspondente a base reduzida é denominada R'_n. Uma vez que o conglomerado $C_{k*,n}$ foi artificialmente eliminado, através da exclusão do excesso de eventos, se R'_n superar um novo limiar L' existe evidência de um segundo conglomerado vivo, diferente de $C_{k*,n}$. A estimativa para este segundo conglomerado é dada pelo cilindro com maior contribuição para a estatística R'_n, assim como no método original. O novo limiar L' é igual ao limiar L, exceto pelos eventos excluídos: $L' = L - \Delta(C_{k*,n})$. Este procedimento é repetido sucessivamente, até que o limiar não seja superado.

2.3 Métricas

Esta seção apresenta as métricas utilizadas para avaliar o desempenho do método STCD e da extensão proposta em bases de dados simuladas.

O método STCD foi utilizado em bases de dados com um único conglomerado. Para cada base de dados utilizada, o STCD foi aplicado sequencialmente no tempo até um *Alarme Correto*, ou até o último evento da base (n). Registrou-se o número total de alarmes. Cada alarme gerado foi classificado como *Alarme Incorreto* ou *Alarme Correto*. Um *Alarme Incorreto* ocorre quando $R_i > L$ para algum $i = 1, \ldots, n$ e o conglomerado estimado não tem interseção com o verdadeiro. Um *Alarme Correto* ocorre quando $R_i > L$ para algum $i = 1, \ldots, n$ e o conglomerado estimado tem alguma interseção com o verdadeiro. Os casos em que $R_i < L$ para todo $i = 1, \ldots, n$ foram chamados de *Sem Alarme*. Para cada base de dados utilizada, o método original foi aplicado sequencialmente no tempo até um *Alarme Correto*, ou até o último evento da base. Registrou-se o número total de alarmes. Se $R_i > L$, $R_{i+1} < L$, $R_{i+2} > L$, então os alarmes nos tempos i e $i+2$ foram considerados como alarmes diferentes. Quando $R_i > L$, $R_{i+1} > L$ e os conglomerados estimados nos tempos i e $i+1$ são iguais, considerou-se os alarmes nos tempos i e $i+1$ como o mesmo alarme. Se $R_i > L$, $R_{i+1} > L$ e os conglomerados estimados nos tempos i e $i+1$ são diferentes, os alarmes nos tempos i e $i+1$ foram considerados como alarmes diferentes. Os conglomerados estimados foram considerados como o mesmo quando a distância euclidiana entre seus centros é menor que 2ρ.

A Figura 1 ilustra estes conceitos. Em (a) o primeiro alarme é um alarme correto. Em (b) existe um período de alarmes incorretos antes do alarme correto. É possível também alternar entre períodos sem alarme e períodos com alarmes incorretos antes do alarme correto. Outra possibilidade é não se chegar a um alarme correto, como em (c). A situação (d) ilustra um caso de *Sem Alarme*.

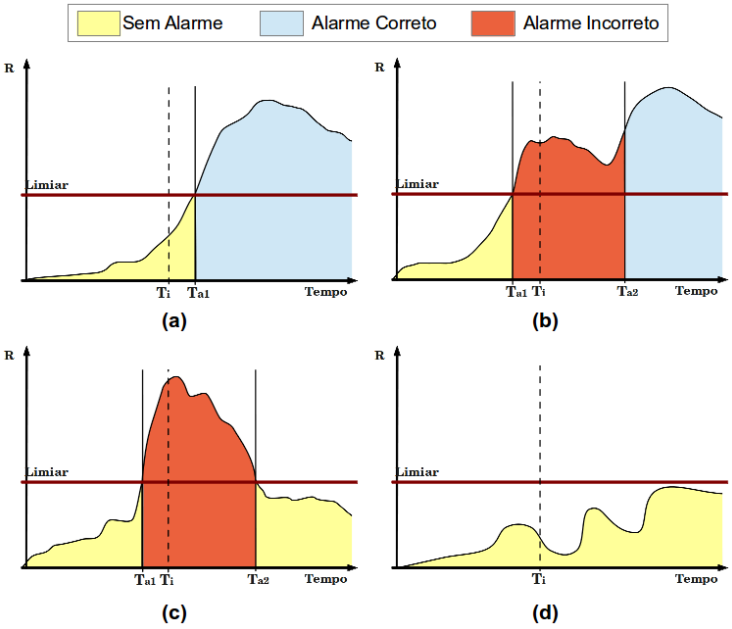

Figura 1: Alarmes para um único conglomerado.

Os casos em que chegou-se a um alarme correto, mesmo que antes deste alarme tenham sido registrados alarmes incorretos, foram chamados de *Identificação Correta*. Os casos em que obteve-se somente alarmes incorretos foram chamados de *Identificação Incorreta*. Os casos sem alarme foram chamados de *Identificação Nula*. As porcentagens de *Identificação Nula*,

Identificação Incorreta e *Identificação Correta* foram calculadas em relação ao número total de alarmes.

Registrou-se o atraso (*delay*) até o primeiro *Alarme Correto*, em unidades de tempo. Este atraso é a diferença entre o tempo do primeiro *Alarme Correto* e o tempo de início do conglomerado verdadeiro.

Utilizou-se também a métrica F1 para comparar o verdadeiro conglomerado (C') com aquele detectado pelo sistema STCD (C^*). Essa métrica é uma média harmônica entre **Precisão** e **Revocação** [3]. A Precisão (P), a Revocação (R) e a métrica F1 são dadas por:

$$P = \frac{|C' \cap C^*|}{|C^*|}; \qquad R = \frac{|C' \cap C^*|}{|C'|}; \qquad F1 = \frac{2PR}{P+R}.$$

A precisão mede o quão precisa foi a detecção do método. Se todos os eventos do conglomerado encontrado pertencerem ao conglomerado verdadeiro, então a precisão é total, $P = 1$. Se nem todos pertencerem ao conglomerado verdadeiro, a precisão fica proporcionalmente abaixo de um.

A revocação mede o quão completa foi a detecção. Se todos os eventos do conglomerado verdadeiro estão no conglomerado encontrado, então a revocação é total, $R = 1$. Se nem todos estão no conglomerado encontrado, então a revocação fica proporcionalmente abaixo de um.

Normalmente deseja-se uma métrica $F1 = 1$, alcançada com $P = 1$ e $R = 1$, para definir uma boa recuperação. No caso do STCD, espera-se que ele identifique o conglomerado o quanto antes, antes mesmo de todos os seus eventos ocorrerem. Esta detecção rápida implica em uma revocação sempre menor que 1. Por isso foi utilizada uma adaptação da métrica F1, chamada aqui de **F1 parcial**. A métrica F1 parcial considera apenas os eventos do conglomerado verdadeiro que ocorreram até o tempo do alarme, permitindo uma **revocação parcial** igual a 1.

A extensão proposta para conglomerados simultâneos foi aplicada a bases com dois conglomerados. Neste caso, os alarmes foram classificados como *Simples* ou *Duplos*. Um alarme é Simples quando $R_i > L$ e $R'_i < L'$. Um alarme Duplo ocorre quando $R_i > L$ e $R'_i > L'$. Em nenhuma das simulações realizadas o STCD-Sim gerou um alarme triplo, o que indica que a princípio este sistema é eficiente no sentido de identificar corretamente o número de conglomerados existentes.

A Figura 2 ilustra as possíveis situações dos alarmes para bases com dois conglomerados. Inicialmente não existe alarme (0). Se apenas o primeiro limiar é excedido (1), então existe um Alarme Simples. Se o primeiro e segundo limiares forem excedidos (2), então o alarme é Duplo. De (1), há três possibilidades. A primeira: o conglomerado encontrado é substituído por outro (3), então há um novo alarme simples. A segunda: o primeiro limiar não é mais excedido (4), então retorna-se para o caso de nenhum alarme. A última: o segundo limiar também é ultrapassado (5), então agora existe um Alarme Duplo. De (2) também há três possibilidades. Se pelo menos um dos dois conglomerados encontrado mudar (6), é um novo Alarme Duplo. Se o primeiro e o segundo limiares não forem mais excedidos (7), retornamos para o caso de Sem Alarme. Se apenas o segundo limiar não for mais ultrapassado (8), então é um Alarme Simples.

Um Alarme Simples pode ser Correto ou Incorreto. Um *Alarme Simples Correto* é aquele em que o conglomerado encontrado possui alguma intercessão com um dos conglomerados verdadeiros, nomeados aqui de $C1$ e $C2$. Se o conglomerado encontrado não tiver nenhuma intercessão com $C1$ ou $C2$, temos um caso de *Alarme Simples Incorreto*.

Um Alarme Duplo pode ser Correto, Incorreto ou Semi-correto. Um *Alarme Duplo Correto* ocorre quando um dos conglomerados encontrados tem alguma intercessão com $C1$ e o outro tem alguma intercessão com $C2$. Um *Alarme Duplo Incorreto* acontece quando os dois conglomerados encontrados não tem intercessão tanto com $C1$ quanto com $C2$. Se apenas um dos dois conglomerados encontrados tiver alguma intercessão com $C1$ ou $C2$, então temos um caso de *Alarme Duplo Semi-Correto*.

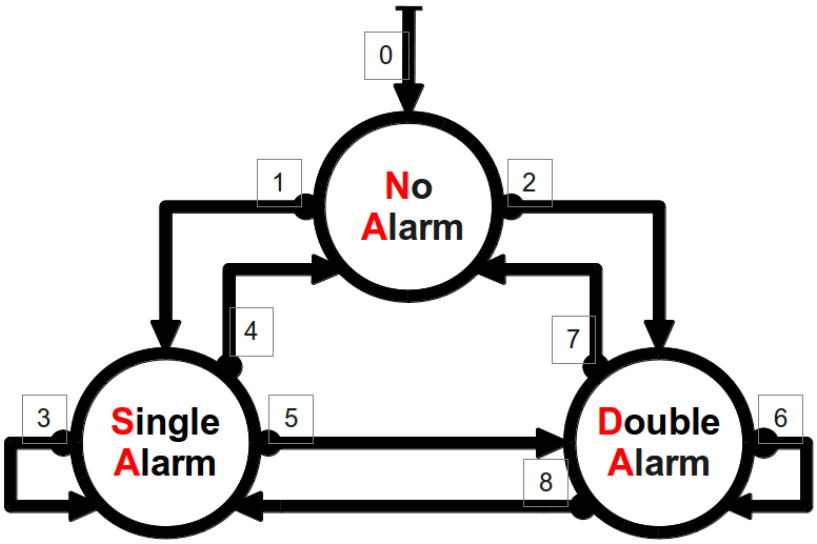

Figura 2: Alarmes para bases com dois conglomerados simultâneos.

Nas bases com dois conglomerados, a extensão proposta foi aplicada sequencialmente até um Alarme Duplo Correto, ou até o último evento da base. Foram analisadas as porcentagens de *Identificação Nula*, *Identificação Incorreta*, *Identificação Incompleta* e *Identificação Completa*, sempre em relação ao número total de alarmes. Considerou-se como *Identificação Completa* os casos de alarme Duplo Correto. Os casos em que ocorreram apenas alarmes Simples Corretos foram considerados como *Identificação Incompleta*. Tanto um alarme Simples Incorreto quanto um Duplo Incorreto foram considerados com *Identificação Incorreta*. Já um alarme Duplo Semi-correto contabilizou 1/2 *Identificação Incorreta* e 1/2 *Identificação Incompleta*. Os casos sem alarme foram classificados como *Identificação Nula*.

Analisou-se também o atraso de detecção, considerando-se algumas variações para este atraso. Foram analisados quatro *delays* diferentes, todos em unidades de tempo:

- **Delay Min** – diferença entre o tempo do primeiro alarme que detecta pelo menos um dos dois conglomerados ($C1$ e $C2$) e o tempo de início dos conglomerados;
- **Delay** $C1$ – diferença entre o tempo do primeiro alarme que detecta o conglomerado $C1$ e o tempo de início de $C1$;
- **Delay** $C2$ – diferença entre o tempo do primeiro alarme que detecta o conglomerado $C2$ e o tempo de início de $C2$;
- **Delay Duplo** – diferença entre o tempo do primeiro alarme duplo que detecta os conglomerados $C1$ e $C2$ e o tempo de início dos conglomerados.

Vale lembrar que o tempo de início dos dois conglomerados, $C1$ e $C2$, são iguais.

3 Resultados e discussões

Nesta seção são apresentados os resultados obtidos aplicando-se o método original e a extensão proposta em base de dados simuladas. O método original foi aplicado em bases de dados contendo um único conglomerado. Já a extensão foi aplicada em bases contendo dois conglomerados espacialmente separados, de mesmo raio e mesmo aumento na intensidade de

eventos. As bases de dados utilizadas são descritas na seção 3.1 e os resultados são apresentados nas seções posteriores.

3.1 Bases de dados

Para avaliar o desempenho do sistema STCD e da extensão proposta, foram utilizadas bases de dados simuladas. Em todas as bases considerou-se o plano espacial $A = [0, 10] \times [0, 10]$ e o intervalo de tempo $[0, 10]$. Dois tipos de bases foram utilizados: bases com um conglomerado espaço-tempo e bases com dois conglomerados simultâneos.

Os eventos foram distribuídos de forma homogênea no espaço e intervalo de tempo considerados, adotando-se uma taxa de 10 eventos por unidade de volume. Esta taxa foi aumentada ε vezes dentro do(s) conglomerado(s) (área circular de raio ρ com uma duração temporal Δt específica). Nas bases contendo conglomerados simultâneos gerou-se dois conglomerados de mesmo raio e mesmo aumento na intensidade de eventos, sem nenhuma interseção espacial.

Os valores adotados para os parâmetros ρ, ε e Δt foram: $\rho = \{0{,}5;\ 1{,}0;\ 1{,}5;\ 2{,}0\}$, $\varepsilon = \{1{,}0;\ 3{,}0;\ 10{,}0\}$ e $\Delta t = \{[5, 10]; [7, 10]; [8, 10]\}$. Para cada uma das 36 configurações analisadas – $(\rho \times \varepsilon \times \Delta t)$ – foram utilizadas 100 bases de dados diferentes.

Ao se executar o método, os verdadeiros valores de ρ e ε foram adotados como entrada para os respectivos parâmetros. Para cada base o limiar L foi definido como sendo igual ao número total de eventos: $L = n$. Note que o número total de eventos n não é exatamente o mesmo em todas as bases.

As configurações foram rotuladas de 1 a 36 da seguinte maneira: $X = (\rho, \varepsilon, |\Delta t|)$: $1 = (0.5, 1.0, 2)$; $2 = (0.5, 1.0, 3)$; $3 = (0.5, 1.0, 5)$; $4 = (0.5, 3.0, 2)$; $5 = (0.5, 3.0, 3)$; $6 = (0.5, 3.0, 5)$; $7 = (0.5, 10.0, 2)$; $8 = (0.5, 10.0, 3)$; $9 = (0.5, 10.0, 5)$; $10 = (1.0, 1.0, 2)$; $11 = (1.0, 1.0, 3)$; $12 = (1.0, 1.0, 5)$; $13 = (1.0, 3.0, 2)$; $14 = (1.0, 3.0, 3)$; $15 = (1.0, 3.0, 5)$; $16 = (1.0, 10.0, 2)$; $17 = (1.0, 10.0, 3)$; $18 = (1.0, 10.0, 5)$; $19 = (1.5, 1.0, 2)$; $20 = (1.5, 1.0, 3)$; $21 = (1.5, 1.0, 5)$; $22 = (1.5, 3.0, 2)$; $23 = (1.5, 3.0, 3)$; $24 = (1.5, 3.0, 5)$; $25 = (1.5, 10.0, 2)$; $26 = (1.5, 10.0, 3)$; $27 = (1.5, 10.0, 5)$; $28 = (2.0, 1.0, 2)$; $29 = (2.0, 1.0, 3)$; $30 = (2.0, 1.0, 5)$; $31 = (2.0, 3.0, 2)$; $32 = (2.0, 3.0, 3)$; $33 = (2.0, 3.0, 5)$; $34 = (2.0, 10.0, 2)$; $35 = (2.0, 10.0, 3)$; $36 = (2.0, 10.0, 5)$.

3.2 Detecção de um único conglomerado

A Figura 3 apresenta as porcentagens de *Identificação Nula*, *Identificação Incorreta* e *Identificação Correta* obtidas aplicando-se o método STCD em bases com um conglomerado. As porcentagens médias são: *89,92%* de Identificação Correta, **2,86%** de Identificação Incorreta e **8,22%** de Identificação Nula. Esta alta taxa de *Identificação Correta* mostra que, de forma geral, o método tem um bom desempenho. As configurações 1 a 3 são exceções a esta regra. Dentre os conglomerados avaliados, o menor e mais fraco (menor ρ e menor ε) está extamente nestas três configurações, o que ajuda a explicar a alta taxa de *Identificação Nula*.

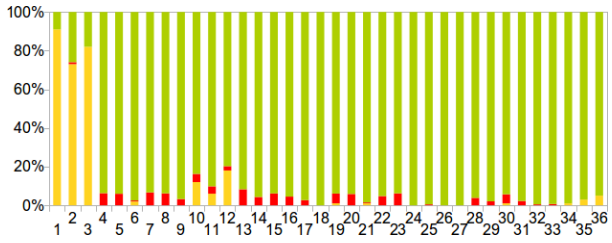

Figura 3: Percentual de identificações do STCD em bases com um conglomerado. Legenda por identificação: **Nula**, **Incorreta** e **Correta**. A ordem das configurações segue a numeração descrita na seção 3.1

A Tabela 1 mostra um resumo dos valores obtidos para o *delay* (em unidades de tempo),

desconsiderando-se as três primeiras configurações em que a taxa de *Identificação Nula* é muito alta.

Tabela 1: Resumo dos resultados do atraso de detecção do STCD em bases com um conglomerado, desconsiderando-se as configurações 1 a 3.

Δt		Atraso Médio	2,5 Percentil	97,5 Percentil
[5 - 10]	Mínimo:	0,018	0,004	0,048
	Máximo:	1,266	0,265	4,200
	Média:	**0,301**	**0,066**	**0,996**
	Mediana:	0,095	0,018	0,258
[7 - 10]	Mínimo:	0,018	0,004	0,045
	Máximo:	0,945	0,178	2,300
	Média:	**0,238**	**0,058**	**0,626**
	Mediana:	0,081	0,024	0,179
[8 - 10]	Mínimo:	0,016	0,005	0,051
	Máximo:	1,266	0,263	1,542
	Média:	**0,223**	**0,061**	**0,517**
	Mediana:	0,080	0,026	0,210

Pode-se observar que, em média, o tempo de atraso do STCD é pequeno. De modo geral, quanto mais tarde o conglomerado emerge, mais rápida é a sua identificação.

Visto que o *delay* do método é bem pequeno, utilizou-se a métrica F1 Parcial para verificar o quão próximo o conglomerado encontrado pelo método estava do conglomerado verdadeiro. Um resumo dos valores obtidos para o F1 Parcial, assim como para a Revocação Parcial e para a Precisão, é apresentado na Tabela 2, também desconsiderando-se as três primeiras configurações.

Tabela 2: Resumo dos resultados obtidos para F1 Parcial do STCD em bases com um conglomerado, desconsiderando-se as configurações 1 a 3. O p no cabeçalho significa percentil.

	Precisão			Revocação Parcial			F1 Parcial		
	média	2,5 p	97,5 p	média	2,5 p	97,5 p	média	2,5 p	97,5 p
Mínimo	0,77	0,30	1,00	0,67	0,09	1,00	0,70	0,16	1,00
Máximo	1,00	1,00	1,00	0,99	0,87	1,00	0,99	0,87	1,00
Média	**0,94**	**0,67**	**1,00**	**0,90**	**0,46**	**1,00**	**0,91**	**0,54**	**1,00**
Mediana	0,96	0,68	1,00	0,93	0,48	1,00	0,94	0,56	1,00

Os resultados acima mostram que o método identifica bem o conglomerado verdadeiro, na média o F1 parcial é de 90%. Além disso, o STCD é mais preciso do que completo: os eventos que o STCD identifica como pertencentes ao conglomerado são normalmente pertencentes ao verdadeiro, porém menos eventos são identificados do que se poderia.

3.3 Detecção de conglomerados simultâneos

A Figura 4 mostra as porcentagens de Identificação *Nula*, *Incorreta*, *Incompleta* e *Completa*, obtidas aplicando-se a extensão proposta em bases de dados contendo dois conglomerados simultâneos.

Em 73,67% das bases analisadas, o STCD-Sim chegou a um alarme Duplo Correto. Em relação ao total de alarmes, os percentuais médios foram: **28,06%** de Identificação Completa, **55,55%** de Identificação Incompleta, **1,47%** de Identificação Incorreta e **14,92%** de Identificação Nula. Estes resultados indicam que o método soa muitos alarmes simples correto e duplo semi-correto antes de soar um alarme duplo correto. A taxa de *Identificação Nula* é muito alta

nas configurações 1 a 3 e 34 a 36. As três primeiras configurações correspondem aos conglomerados menores e mais fracos dentre os considerados, e têm altas taxas de *Identificação Nula* mesmo quando existe apenas um conglomerado. As configurações 34 a 36 também apresentam altas taxas de *Identificação Nula*, diferentemente do caso de um único conglomerado. Um estudo mais detalhado é necessário para entender o motivo destas altas taxas. Desconsiderando-se estas seis configurações, em que o nível de identificação nula é alto, o percentual médio de bases com identificação completa sobe para 88,23%.

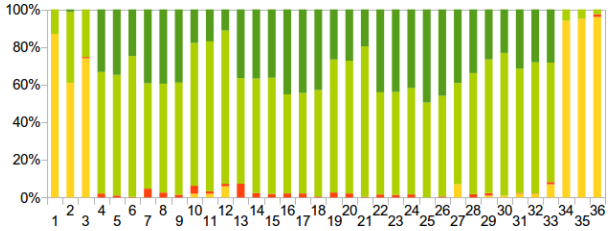

Figura 4: Percentual de identificações do STCD-Sim em bases com 2 conglomerados. Legenda por identificação: **Nula**, **Incorreta** e **Correta**. A ordem das configurações segue a numeração descrita na seção3.1

A Figura 5 mostra um comparativo dos atrasos do STCD original (com um conglomerado) e do STCD-Sim (com dois conglomerados). Em todos os gráficos, a primeira barra representa o atraso médio até um *Alarme Correto* em bases com um conglomerado (*Delay 1*). As demais barras correspondem aos atrasos médios em bases com dois conglomerados. Estas barras representam, nesta ordem, os seguintes atrasos médios: *Delay Min*, *Delay C1*, *Delay C2* e *Delay Duplo*. O segmento em cada barra é o intervalo $[p_{2,5}; p_{97,5}]$, onde $p_{2,5}$ é o percentil 2,5 e $p_{97,5}$ é o percentil 97,5. Os resultados obtidos para as configurações $\rho \times \varepsilon = (0,5;\ 1)$ e $(2,0;\ 10)$ não são apresentados nesta figura, uma vez que o método STCD-Sim se mostrou ineficiente nestes extremos, com altas taxas de *Identificação Nula*.

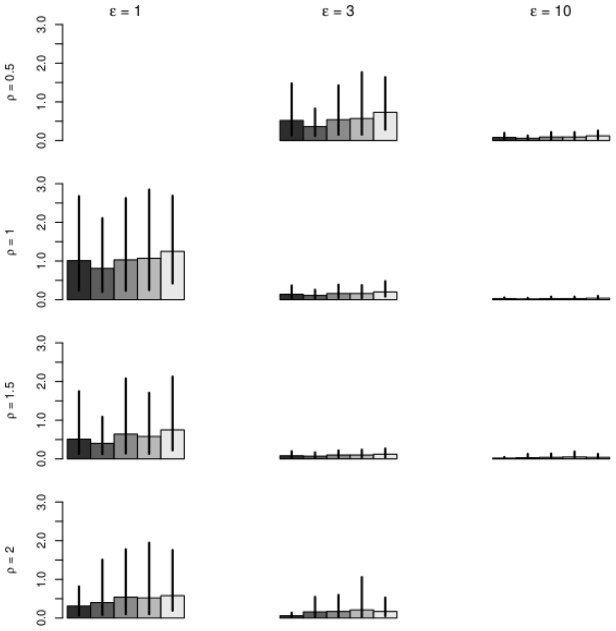

Figura 5: Atrasos de detecção, comparação entre STCD original e STCD-Sim. Em todos os gráficos, a primeira barra representa o atraso até um *Alarme Correto* no caso de um conglomerado (*Delay 1*). As demais barra são relativas ao caso de dois conglomerados e representam, nesta ordem: *Delay Min*, *Delay C1*, *Delay C2* e *Delay Completo*.

Os resultados acima indicam que a extensão para múltiplos conglomerados detecta um único

conglomerado tão rápido quanto o método original. A diferença entre o atraso *Delay Duplo* e os atrasos *Delay C1* e *Delay C2* é bem pequena. Isto significa que esperando-se um pouco mais, é possível detectar corretamente os dois conglomerados em um único alarme. Em geral, mesmo antes de um *Alarme Duplo Correto*, a extensão STCD-Sim detecta os dois conglomerados através de alarmes diferentes.

4 Conclusões

Uma das principais vantagens do sistema de vigilância espaço-tempo estudado neste trabalho é sua robustez a diferenças de concentração do processo no tempo e no espaço, dispensando o uso de população para normalização de dados. Como não é feita nenhuma suposição sobre não estacionariedade do processo, o método também se adapta bem às prováveis diferenças de volume de dados ao longo do tempo.

Os resultados apresentados mostram que a extensão STCD-Sim proposta é satisfatória para detecção de conglomerados simultâneos separados espacialmente. Em cerca de 73% das bases de dados utilizadas chegou-se a um alarme duplo correto, com um atraso razoavelmente pequeno. O STCD-Sim parece ser eficaz no sentido de detectar o número correto de conglomerados, visto que nas simulações realizadas com dois conglomerados não houve nenhum alarme triplo.

O desempenho, tanto do método STCD quanto da extensão proposta, deixa a desejar em algumas configurações analisadas. Estes casos são a minoria, mas precisam ser analisados com mais detalhes em trabalhos futuros. Cenários não adotados neste trabalho - com mais de dois conglomerados, com conglomerados que se intersectam no espaço - devem ainda ser analisados posteriormente. A calibração automática dos parâmetro de entrada ρ e ε e a outras formas geométricas de conglomerado também são desafios futuros que podem tornar o método ainda mais eficaz.

Referências

[1] Assunção, R.; Correa, T. Surveillance to detect emerging space-time clusters. **Computational Statistics & Data Analysis**. Elselvier. v. 53, p. 2817-2830, 2009.

[2] Assunção, R. Iabrudi, A.; Kulldorff, M.; Correa, T. Space-time cluster identification in point processes. **The Canadian Journal of Statistics**, v. 35, p. 1-17, 2007.

[3] Baeza-Yates, R.; Ribeiro-Neto, B. **Modern Information Retrieval**. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[4] Demattei, C.; Cucala, L. Multiple spatio-temporal cluster detection for case event data: an ordering-based approach. **Communications in Statistics-Theory and Methods**, v. 2, p. 358-372, 2010.

[5] Diggle, P.; Rowlingson, B.; Su, T. Point process methodology for on-line spatio-temporal disease surveillance. **Environmetrics**, v. 16, p. 423-434, 2005.

[6] Eckley, D.; Curtin, K. Evaluating the spatiotemporal clustering of traffic incidents. **Computers, Environment and Urban Systems**, v. 37, p. 70-81, 2013.

[7] Kenett, R.; Pollak, M. Data-analytic aspects of the Shiryayev-Roberts control chart: surveillance of a non-homogeneos Poisson process. **Journal of Applied Statistics**, v. 23, p. 125-137, 1996.

[8] Kulldorff, M. A spatial scan statistic. **Communications in Statistics - Theory and Methods**, v. 26, p. 1481-1496, 1997.

[9] Kulldorff, M. Prospective time periodic geographical disease surveillance using a scan statistic. **Journal of the Royal Statistical Society, Series A**, v. 164, p. 61-72, 2001.

[10] Kulldorff, M.; Heffernan, R.; Hartman, J.; Assunção, R.; Mostashari, F. A space-time permutation scan statistic for disease outbreak detection. **PLoS Medicine**, v. 2, p. e59, 2005.

[11] Lee, R.; Wakamiya, S.; Sumiya, K. Discovery of unusual regional social activities using geo-tagged microblogs. **World Wide Web**, v. 14, p. 321-349, 2011.

[12] Li, Xiao-Zhou; Wang, Jin-Feng; Yang, Wei-Zhong; Li, Zhong-Jie; Lai, Sheng-Jie. A spatial scan statistic for multiple clusters. **Mathematical biosciences**, v. 12, p. 135-142, 2011.

[13] Marshall, J. B.; Spitzner, B. D.; Woodall, W. H. Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time. **Statistics in Medicine**, v.26, p. 1579-1593, 2007.

[14] Oliveira, M.; Baptista, C. GeoSTAT: A system for visualization, analysis and clustering of distributed spatiotemporal data. **Proceedings XIII GEOINFO**, p. 108-119, 2012.

[15] Richardson, D. B. Real-Time Space–Time Integration in GIScience and Geography. **Annals of the Association of American Geographers**, v. 5, p. 1062-1071, 2013.

[16] Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. **World Wide Web**, p. 851-860, 2010.

[17] Sonesson, C.; Bock, D. A review and discussion of prospective statistical surveillance in public health. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, v. 166, p. 5-21, 2003.

[18] Takahashi, K.; Kulldorff, M.; Tango, T.; Yih, K. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. **International Journal of Health Geographics**, v.7, p. 14, 2008.

[19] Tango, T.; Takahashi, K.; Kohriyamma, K. A space-time scan statistic for detecting emerging outbreaks. **Biometrics**, v. 67, p. 106-115, 2011.

[20] Zhang, Z.; Assunção, R.; Kulldorff, M. Spatial scan statistics adjusted for multiple clusters. **Journal of Probability and Statistics**, 2010.

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