VIII Brazilian Symposium on GeoInformatics, Campos do Jordão, Brazil, November 19-22, 2006, INPE, p. 315-320.

Medida de correlação entre padrões pontuais de origem-destino

Renato M. Assunção, Danilo L. Lopes

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

Abstract. *When we use origin-destination survey data, we can be interested in verifying if the link between the two point patterns is or is not a result of randomness. In this paper we present a new measure of bivariate processes linking strength, find out its probability distribution and apply the method to data of car thefts location and its retrieving locations in Belo Horizonte.*

Resumo. *Quando se utilizam dados pontuais do tipo origem-destino é interessante verificar se a ligação entre os dois padrões pontuais é resultado de uma aleatoriedade ou não. Para tanto o presente artigo apresenta uma nova medida para expressar a força da ligação entre dois processos pontuais, determina a sua distribuição de probabilidade sob hipótese de aleatoriedade na ligação e apresenta um exemplo de aplicação da metodologia em furtos e localização de veículos em Belo Horizonte.*

1. Introdução

Em diversas aplicações, como biologia, criminologia ou sociologia, dados de padrões pontuais podem apresentar a característica de expressarem eventos de fluxo no espaço, correspondendo, para cada evento, a uma localização inicial e a uma localização final. Esses padrões pontuais podem ser observados a partir de exemplos como dados de fluxos migratórios, dados de furto e localização de objetos, dados de local de nascimento e de morte de seres vivos ou mesmo os locais de assalto e de residência do assaltante. Todos esses dados possuem em comum o fato de apresentarem dois padrões pontuais cujos eventos estão ligados entre si par a par. Ou seja, para cada um dos eventos em um dos padrões existe somente um evento correspondente no outro padrão. Como em grande parte desses tipos de dados está evolvida uma sucessão temporal, será utilizada uma notação neste artigo: um dos processos pontuais será chamado de processo de origem e o outro de processo de destino. Para o estudo da ligação entre os dois processos, deverá ser assumido que cada origem possui um destino correspondente, apesar de isso não corresponder a alguns tipos de dados (nem todo objeto furtado, por exemplo, será localizado, mas para o presente estudo deveriam ser utilizados apenas os objetos furtados que foram localizados).

Analisando-se isoladamente os processos de origem e de destino é possível identificar áreas de maior ocorrência de pontos de origem e áreas de maior ocorrência de pontos de destino. Pode existir, porém, uma dúvida quanto à ligação entre os dois processos: dada a existência de um ponto de origem em uma determinada região, que

tipo de inferência pode ser realizada sobre a localização do ponto de destino correspondente.

O desejável seria apontar aquelas áreas que são mais prováveis de se observar o ponto de destino a partir da observação de um ponto de origem. Essa informação, porém, é muito difícil de ser obtida baseando-se em uma única realização dos padrões pontuais de origem e destino. Alternativamente o presente artigo propõe o uso de uma medida sintética, a qual chamaremos de função M, para expressar a força da ligação entre dois processos pontuais. A partir dela é possível verificar se, em geral, pontos de origem que ocorrem próximos entre si geram pontos de destino que também são próximos entre si. Assim, apesar de ainda não ser possível apontar áreas mais prováveis para a ocorrência do ponto de destino dada a localização do ponto de origem, será possível inferir se, para o processo como um todo, essas áreas são bem definidas ou não.

2. Metodologia

2.1. Função K: uma breve descrição

A criação da função M está fundamentada na teoria existente sobre a função K. A função K ou "medida reduzida de segundo momento", introduzida originalmente em [Ripley 1976], é um método que proporciona uma descrição efetiva da dependência espacial dentro de um padrão pontual em determinada área, ou seja, verifica se os eventos estão distribuídos ao acaso, se eles se encontram em padrão de regularidade, ou se existem conglomerados específicos, sendo tal análise possível para uma vasta gama de escalas espaciais. A função K é definida como o número esperado de eventos dentro de uma distância *h* de um evento arbitrário divido pela intensidade de pontos na região. A sua estimação baseia-se na variação do número médio de outros eventos em torno de um evento qualquer, à medida que variamos o raio de consideração.

A popularidade da função K encontra-se primeiramente na facilidade de sua estimação se comparada com outras medidas de segunda ordem para processos pontuais como a função intensidade de segunda ordem. Ela apresenta uma propriedade muito útil, a invariância sob redução aleatória, o que permite que o tamanho de amostra não interfira na sua estimação. Sob a hipótese de Completa Aleatoriedade Espacial (CAE), que reflete uma disposição casual dos eventos, sem interações entre eles, a função K é expressa como: $K(h) = \pi h^2$. O teste de Monte Carlo de CAE a partir da função K utiliza-se da geração de diversas simulações sob CAE, estimando a função K para cada realização. Como se está interessado em verificar o comportamento da função K através de um intervalo de distâncias, constrói-se envelopes em volta do valor esperado sob hipótese nula. Para maiores informações sobre função K, veja Diggle (2003) ou Waller and Gotway (2004).

2.2. A função M

Seja (N_1, N_2) um processo pontual bivariado em um polígono $A \subset R^2$ formado pelos eventos $x_1, \cdots, x_n$ de N_1 e $y_1, \cdots, y_n$ de N_2 onde $x_i = (x_{1i}, x_{2i})$ e $y_i = (y_{1i}, y_{2i})$. Os processos N_1 e N_2 são ligados de forma que todo evento em N_1 causa um evento em N_2 a ser observado dentro de A. N_1 e N_2 serão aqui chamados, respectivamente, de processo de origem e de destino. Um exemplo principal seria *n* pares de locais de carros furtados

em x_i e localizados mais tarde em y_i, $i = 1,...,n$. Outro exemplo de dados de origem-destino seria o local de assassinato e a residência da vítima.

A função M(k1, k2) será definida como o número esperado de eventos que estão entre os k1 pontos de origem mais próximos de um ponto de origem arbitrário cujos pontos de destino estão entre os k2 vizinhos mais próximos do ponto de destino correspondente a esse ponto de origem arbitrário. Aqui os k eventos vizinhos mais próximos de um determinado evento correspondem àqueles eventos cuja distância euclidiana em relação ao evento de referência está entre as k menores dentro da realização do processo pontual. A função pode ser estimada como o número médio observado dos eventos na realização que pertençam à vizinhança (cujo tamanho é definido por k1 e k2) do evento arbitrário tanto na origem quanto no destino. Fixando-se os valores de k1 e k2, para cada evento de origem x_i, seja m_i o número de eventos que estão entre os k1 vizinhos mais próximos de x_i cujos destinos estão entre os k2 eventos mais próximos de y_i, destino de x_i. A estimativa da função M será dada por: $\hat{M}(k1,k2) = \frac{1}{n}\sum_i m_i$. A utilização do número de vizinhos mais próximos como parâmetro da função em vez da distância como na função K tem por objetivo retirar o efeito que as características de cada processo pontual poderiam gerar sobre a medida, permitindo que ela seja influenciada somente pelas características da ligação entre os processos de origem-destino. Assim, ao se testar a aleatoriedade da ligação entre os processos, nenhuma suposição sobre o modelo de cada processo pontual precisa ser assumida.

Sob a hipótese de aleatoriedade da ligação entre os processos, para um evento de origem arbitrário dentre os n eventos o número dos k1 pontos de origem vizinhos mais próximos cujos pontos de destino estiveram entre os k2 vizinhos mais próximos do ponto de destino correspondente a esse evento segue uma distribuição hipergeométrica (n-1, k1, k2), pois tanto os k1 pontos de origem quanto os demais possuem igual probabilidade de estarem ligados a algum daqueles k2 pontos de destino vizinhos mais próximos do ponto de destino em questão. Como a estimativa da função M(k1, k2) é uma média desses valores para todos os pontos de origem, ela possui a distribuição da média entre n variáveis hipergeométricas (n-1, k1, k2) com relações de dependência muito complicadas entre elas. Sob a hipótese de aleatoriedade da ligação entre os processos a esperança da função M é dada por:

$$E\left[\hat{M}(k1,k2)\right] = \frac{k1 * k2}{n-1}$$

Para a realização do teste, mantêm-se as posições de ambas as realizações, mas permutam-se as ligações entre os processos e para cada uma dessas permutações computa-se o valor de M(k1, k2). Por fim, um intervalo de 95% percentílico de confiança é criado. Como a função M varia conforme os valores de k1 e k2, é possível criar envelopes bidimensionais de 95% para testar a hipótese nula. A análise é feita de forma semelhante àquela realizada nos testes de função K.

3. Aplicações do método

Como exemplo de aplicação do método da função M, foram utilizados dados de furto e localização de 5217 veículos no município de Belo Horizonte, fornecido pela Polícia Militar de Minas Gerais. A Figura 1 apresenta os mapas de Kernel para os dois padrões pontuais em estudo com a divisão do município em bairros. Para o padrão pontual de furtos de veículos em Belo Horizonte, verifica-se que a área central da cidade possui um grande pico de intensidade de furtos, seguido por outros picos na em direção ao Oeste da cidade; esse padrão deve ser devido à proximidade dessas regiões de alta intensidade em relação à rodovia BR-262, que dá acesso ao Triângulo Mineiro e é uma rota para o Paraguai. Para o padrão de localização de veículos furtados em Belo Horizonte, verifica-se que os maiores picos de intensidade correspondem a picos de intensidade para o padrão de furto de veículos, sendo que a intensidade de localizações é maior quanto mais próximo se está da área central da cidade. Essas observações sobre a intensidade da localização de veículos furtados são um bom indício em favor da hipótese de que um carro furtado possui grandes chances de ser localizado na proximidade do local de furto ou na proximidade da área central da cidade.

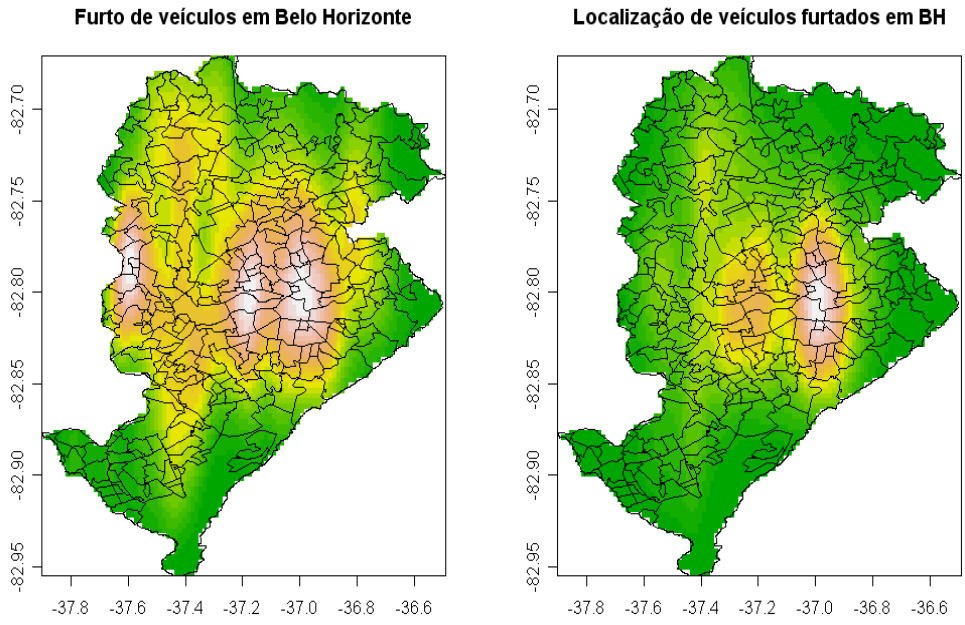

Figura 1: Mapas de Kernel das intensidades estimadas de furto e localização de veículos com a divisão de Belo Horizonte por bairros

Infelizmente a utilização de todos os 5217 veículos do banco de dados para o cálculo da função M é computacionalmente trabalhosa, pois este envolve para cada um dos pares de eventos um cálculo de distância, além de ser necessária a comparação de todas as distâncias relativas a um determinado evento para determinar aqueles que são seus vizinhos mais próximos em cada um dos padrões pontuais. Por isso, nesta aplicação da metodologia foi utilizada apenas uma amostra aleatória de tamanho 2000 destes veículos. Foram calculados valores da função M para valores de k1 e k2 iguais a

250, 500, 750 e 1000. Os valores observados encontram-se dispostos na Tabela 1, enquanto na Tabela 2 são apresentados os valores esperados sobre hipótese de aleatoriedade na ligação entre os processos de origem-destino:

Tabela 1. Valores observados da estimativa para a função M para uma amostra aleatória de 2000 veículos furtados e localizados em Belo Horizonte

k1\k2	250	500	750	1000
250	66.302	108.668	142.433	171.359
500	109.282	202.549	278.999	339.745
750	141.913	276.801	402.331	504.115
1000	170.871	337.262	500.339	649.798

Tabela 2. Valores esperados para a estimativa da função M para uma amostra aleatória de 2000 veículos furtados e localizados em Belo Horizonte sob a suposição de ligação aleatória entre os padrões pontuais

k1\k2	250	500	750	1000
250	31.266	62.531	93.797	125.063
500	62.531	125.063	187.594	250.125
750	93.797	187.594	281.391	375.187
1000	125.063	250.125	375.187	500.250

Como se pode perceber a partir da comparação entre as tabelas, os valores observados para a estatística encontram-se consideravelmente distantes de seus valores esperados sob hipótese nula de aleatoriedade na ligação entre os padrões. Existe um forte indício de que a ligação entre os dois processos não seja aleatória: na verdade suspeita-se que, como dito anteriormente, em Belo Horizonte, carros furtados possuem grande tendência de serem localizados ou no centro da cidade ou nas proximidades da localidade do furto. Seria necessário, porém, um teste não-paramétrico para verificação da hipótese, no caso o uso dos envelopes bidimensionais. Os próximos passos da pesquisa se concentrarão no desenvolvimento do teste de envelopes para a função M.

4. Discussões e Conclusões

A função M tem se demonstrado computacionalmente árdua, devido ao grande número de operações e comparações envolvidas em seu cálculo. A utilização das quantidades de vizinhos mais próximos como parâmetros para a função é vantajosa por retirar possíveis efeitos que os modelos dos padrões pontuais pudessem acrescentar na medida, mas cria

dificuldades para se controlar a criação de padrões pontuais para o teste dos envelopes sob hipótese alternativa. Um próximo passo seria, ainda, obter quais faixas de k1 e k2 seriam ideais para a realização do teste; por enquanto estão sendo utilizados valores arbitrários para k1 e k2; sabe-se que para valores muito altos (próximos de n-1) ou muito baixos de k1 e k2 não há muita variação dos valores estimados para a função, o que prejudica a criação de intervalos de confiança. Resta também verificar se os valores de k1 e k2 causam ou não impactos diferentes na estimativa da função M.

A função M tem demonstrado, porém, bastante eficácia ao apontar casos de aleatoriedade ou não da ligação entre os processos de origem-destino em uma visão preliminar e apresenta a grande vantagem de possuir uma estimativa de fácil compreensão e de distribuição relativamente simples.

Bibliografia

Diggle, P. (2003) "Statistical analysis of spatial point patterns". 2nd ed. London: Arnold; New York: Oxford University Press.

Gelfand A. E., Kim H. J., Sirmans C. F. (2001). "Spatial Modeling with Spatially Varying Coefficient Processes". Technical Report 2001-18, Department of Statistics, University of Connecticut.

Ripley, B. D. (1976) "The second-order analysis of stationary point patterns", In *Journal of Applied Probability* 13, 255-266.

Waller L. A., Gotway C. A. (2004) "Applied Spatial Statistics for Public Health Data". John Wiley & Sons, Inc., Hoboken, New Jersey, p. 137-141.