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

Fábio Rocha da Silva[1,2]

Resumo: *Existem varias técnicas estatísticas para testar a hipotese de que os tempos e posições de eventos pontuais em R^3 são independentes. Isto, testar se casos que estão próximos no tempo tendem a estar próximos no espaço também. Estes testes sofrem de um problema típico dos testes de hipotese: se existirem muitos eventos, o teste ser significativo mesmo se a associação entre tempo e espaço seja fraca. Existem propostas de medidas de associação para tabelas de contingência que procuram corrigir este problema. Neste trabalho, tentaremos adaptar estas idéias e introduzir uma nova medida de associação para dados contínuos. Esta medida pode ser usada em estudos de processos pontuais espaço-temporais. O objetivo deste trabalho é o desenvolvimento de metodologia estatística para medir o grau de associação entre as coordenadas espaciais e as coordenadas temporais de eventos pontuais.*

Palavras-chave: *Medida de associação, espaço-tempo, τ de Goodman e Kruskal.*

Abstract: *There are several statistics techniques to test the hypothesis that the times and positions of events in R^3 are independent. That is, test whether cases that are close in time tend to be close in space . These tests suffer from a typical problem of hypothesis tests: if there are many events, the test can be significant even if the association between time and space is poor. There are proposals for measures of association for contingency tables that try to correct this problem. In this work, we try to adapt these ideas and introduce a new measure of association for continuous data. This measure can be used in studies of specific spatial-time processes. This study is one development of statistical methodology for measuring the degree of association between the spatial coordinates and temporal coordinates of point events.*

Keywords: *measure association, space-time, τ Goodman and Kruskal.*

1 Introdução

A consideração simultânea dos padrões espaciais e temporais da ocorrência dos eventos é importante para identificar clusters ou conglomerados espaços-temporais. Definimos o cluster espaço-temporal como uma região geograficamente pequena em relação à região em estudo e que concentra um número excessivo de eventos durante um período limitado de tempo.

O teste de detecção de conglomerados espaços-temporais mais popular foi desenvolvido por Knox (1964). Especificando-se distâncias críticas temporais e espaciais é possível determinar se um par de eventos está próximo no tempo e no espaço. O teste baseia-se no número X de pares

[1]DEST - UFMG. e-mail: *fabiorochadasilva@yahoo.com.br*
[2]Agradeço a Fapemig, ao CNpq e a Universidade Federal de Minas Gerais pelo apoio financeiro a esta pesquisa.

de eventos que estão simultaneamente próximos no espaço e no tempo. Um alto valor X seria uma indicação de que há uma tendência de casos próximos no tempo serem também próximos no espaço, retratando a interação espaço-tempo.

O teste de Knox, assim como as outras técnicas para testar a hipótese de independência entre espaço e tempo, sofre de um problema típico dos testes de hipóteses: se existirem muitos eventos, o teste pode ser significativo mesmo se a associação entre espaço tempo for fraca.

Seria de grande valia termos uma medida de associação que possa ser usada em conjunto com o teste de hipóteses e que mensure a magnitude da possível relação entre as variáveis. Existem diversas propostas de medidas de associação para tabelas de contingência que procuram complementar o teste qui-quadrado de Pearson. Uma das quais foi proposta por Goodman e Kruskal (1964) denominada por tau de Goodman e Kruskal.

O tau de Goodman e Kruskal (notação τ_{GK}) é obtido usando o principio da redução proporcional dos erros.Isto é, o coeficiente tem por objetivo responder à questão: Em que medida o fato de conhecermos a classificação de uma das variáveis (por exemplo, a linha da tabela em que a observação se encontra) nos torna mais hábeis para prevermos a classificação da outra variável (a coluna na qual cai a observação)?

Esta estatística tem algumas propriedades desejáveis, tais como ser uma medida na direção da associação das variáveis com limites zero (nenhuma associação) e um (completa associação) e não mudar o seu valor com a permutação de linhas e colunas.

Além disso, τ_{GK} tem uma interpretação muito clara: mede o decréscimo relativo na probabilidade de errar a previsão da variável linha ao conhecer a variável coluna (ou vice versa). Por exemplo, se $\tau_{GK} = 0.8$, isto significa que temos uma redução de 80% na probabilidade de errar a previsão de uma das variáveis, quando se usa a informação sobre a outra variável.

Neste trabalho foram usadas as idéias de Goodman e Kruskal para construir uma medida de associação entre espaço e tempo para processos pontuais, bem como variáveis aleatórias (X, Y) quaisquer. Esta medida tem boas propriedades tais como: ter os seus valores entre 0 e 1; se as variáveis são independentes, o índice é zero; tem uma interpretação no sentido do quanto o conhecimento de uma das variáveis nos torna aptos para prever os valores da outra.

2 Material e métodos

Sabemos que o tipo de relação que pode existir entre espaço e tempo não é monótona. Isto quer dizer que não poderemos ter uma interpretação direta sobre a associação como temos como coeficiente de correlação de Pearson, por exemplo. Seria interessante obtermos uma medida de associação que conseguisse captar relacionamentos não monótonos como na Figura 1 e, que tenha boas bases teóricas e que seja interpretável.

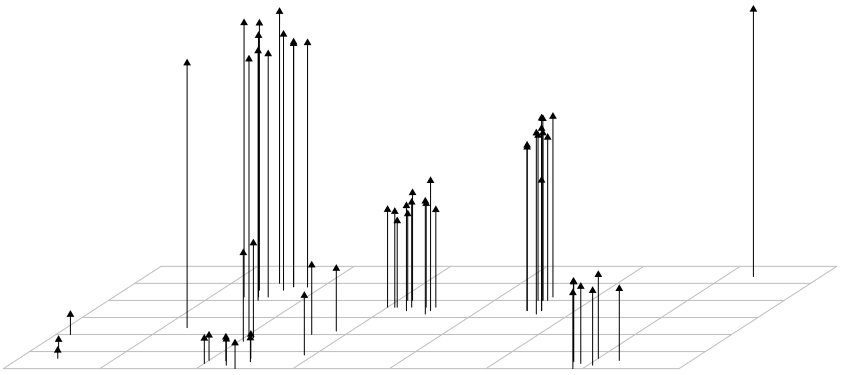

Figura 1: Gráfico de coordenadas espaciais e temporais em R^3

Neste trabalho usamos as idéias de Goodman e Kruskal para construir uma medida de associação entre espaço e tempo. Esta medida usa duas quantidades em sua construção:

O tau de Goodman e Kruskal (notação τ_{GK}) é uma medida de redução proporcional do erro de predição na análise de uma tabela de contigência. Esta medida é obtida usando o principio da redução proporcional dos erros na predição de uma das variáveis. São calculados dois erros de predição, um é calculado o erro ao se tentar alocar os elementos de uma variavel ao seu respectivo nível na ausência de informações sobre de outra variável, e o segundo tipo é calculado erro ao se tentar alocar os elementos de uma variavel ao seu respectivo nível mas dessa vez com o conhecimento prévio do valor da outra variável.

Isto é, este coeficiente tem por objetivo responder à questão: Em que medida o fato de conhecermos a classificação de uma das variáveis (seja ela linha ou coluna) nos torna mais hábeis para prevermos a classificação da outra variável?

O método consiste em apagar a informação de que linha e de que coluna um elemento pertence, e, logo depois , tentar recolocar este elemento a sua respectiva linha seguindo duas regras:

A regra 1 consiste em tentar realocar o elemento a sua verdadeira linha usando apenas a informação do total marginal das linhas. Melhor dizendo, vamos supor que se o elemento pertence à linha i ele recebe uma cor. Suponha agora que a cor de todos os elementos foram apagadas e que a única informação de que dispomos para recolocar estes indivíduos a sua verdadeira linha seja o número total de indivíduos pertencentes a cada linha.

É claro que ao tentarmos realocar os elementos a sua respectiva linha cometermos erros. Denotamos este primeiro conjunto de erros por A, sendo calculado pela seguinte fórmula:

$$A = N \sum_i n_l p_{i.}(1 - p_{i.}),$$

em que N é o total de elementos na tabela de contingência, n_l é o total níveis da variável linha, e $p_{i.}$ é a probabilidade marginal da linha i.

A regra 2 consiste em usar a informação sobre os totais marginais da variável coluna para tentar realocar cada elemento a sua verdadeira linha. Melhor explicando, suponhamos novamente que se o elemento pertence à linha i ele recebe uma cor e que por algum motivo a cor de todos os elementos foram apagadas . A diferença é que agora para realocá-los à sua verdadeira linha ,temos além da informação do número total de indivíduos pertencentes a cada linha, temos também a informação do número total de indivíduos pertencentes a cada coluna.

O número de erros obtidos por esta regra é denotado por B e é calculado da seguinte forma:

$$B = \sum_j n_{.j} \sum_i \frac{p_{ij}}{p_{j.}} \left(1 - \frac{p_{ij}}{p_{j.}}\right),$$

na qual $n_{.j}$ representa o total marginal da coluna j, p_{ij} é a probabilidade conjunta , e $p_{.j}$ é a probabilidade marginal da coluna j.

Então a estatística de Goodman e Kruskal é definida como:

$$\tau_{GK} = \frac{A - B}{A} \tag{1}$$

Esta estatística tem algumas propriedades desejáveis, tais como ser uma medida com limites zero (nenhuma associação) e um (completa associação) e não mudar com a permutação de linhas e colunas.

Além disso τ_{GK} tem uma interpretação muito clara: mede o decréscimo relativo na probabilidade de errar a previsão da variável linha ao conhecer a variável coluna (ou vice versa). Por exemplo, se $\tau_{GK} = 0.8$ uma redução de 80% na probabilidade de errar a previsão de uma das variáveis, quando se usa a informação sobre a outra variável.

3 Resultados e discussões

Encontramos na literatura vários aplicações do teste de Knox em que o teste sugeriu que existiam conglomerados espaço-temporais, porém estes conglomerados tinham pouca significância prática.

Esta situação ocorre porque o teste de knox, assim como os outros métodos utilizados para testar a hipótese de interação espaço-tempo, se limitam a verificar a existência ou não de uma interação espaço-tempo, sem identificar a magnitude desta interação.

Além do problema descrito acima, sabemos que o tipo de relação que pode existir entre espaço e tempo não é monótona. Isto quer dizer que não poderemos ter uma interpretação direta sobre a associação como temos como coeficiente de correlação de Pearson, por exemplo.

O que se busca é obter é uma medida de associação que consiga captar relacionamentos não monótonos e, que tenha boas bases teóricas.

Neste trabalho usamos as idéias de Goodman e Kruskal para construir uma medida de associação entre espaço e tempo. Inicialmente iremos mostrar como a construção da medida de associação para duas variáveis aleatórias contínuas X e Y quaisquer e, logo depois, estender para as variáveis espaço tempo.

Gráficos de Dispersão são comumente usados para exibir e comparar valores numéricos, como dados científicos, estatísticos e de engenharia. Gráficos de Dispersão têm dois eixos de valores, mostrando um conjunto de dados numéricos ao longo do eixo horizontal e outro ao longo do eixo vertical.

Suponha que iremos traçar uma grade, com quadrados de lado de lados Δ_i e Δ_j. Observe que agora temos uma estrutura semelhante a uma tabela de contingência, quer dizer cada celular seria correspondente a um quadrado de lados Δ_i e Δ_j e os elementos desta célula seria o número de pontos dentro deste quadrado.

Ao encararmos um gráfico de dispersão em R^2 como uma tabela de contingência iremos adaptar a medida de Goodman e Kruskal para dados contínuos. Esta medida usa duas quantidades em sua construção:

A primeira é **A** e é a taxa de acertos na predição de cada valor da variável X (que são as coordenadas no espaço R^2) . Ou seja, ao invés de procurar obter a taxa esperada de erros, como fizeram Goodman e Kruskal, calcularemos a taxa esperada de acertos na predição de cada valor da variável X . Fazendo Δ_i e Δ_j tender a zeero pode ser mostrado que :

$$A = E\left[f(X)\right]$$

Ou seja, A é o valor esperado da densidade de X, em um ponto X que segue a distribuição f. Para a quantidade **B** é usada a informação sobre a variável tempo para predizer os valores, em cada quadrado de tamanho Δ, da variável espacial. O valor esperado da taxa de acertos da variável espacial X dado o valor da variável tempo é o que chamaremos de B. E, depois de simplificações matemáticas obtemos:

$$B = E_T\left[E_{X|T=t}\left(f(X|T=t)\right)\right]$$

Onde E_T é o valor esperado com relação a variável tempo; $E_{X|T=t}$ é o valor esperado da variável espaço dado o tempo t; $f(X|T=t)$ é a densidade condicional da variável espaço dado o tempo t.

A nossa medida de associação se propõe a responder a seguinte pergunta: "Em que proporção a informação de uma das variáveis nos ajuda a acertar a predição de cada valor da outra variável?"Sendo assim definimos a nossa medida de associação como sendo a razão:

$$\Psi = \frac{B - A}{B} \tag{2}$$

Além da interpretação de aumento da capacidade preditiva, esta medida de associação possui três propriedades muito claras:

1. Os valores do índice estão entre 0 e 1.

2. Independência:
   Se as variáveis são independentes, o índice é zero.

3. Coerência:
   O índice aumenta com o aumento da dependência, sendo igual 1, quando uma variável é totamente dependente da outra.

Como podemos notar Ψ depende de densidades de probailidade e ao fazermos inferência de um modelo específico é possível obter um ganho muito grande em eficiência, mas somente se o modelo de probabilidade assumido for pelo menos aproximadamente verdadeiro. Com o objetivo de estimar estas densidades utilizamos o método de Estimação de Densidades via núcleo estimador (Kernell density) e fornecemos um método bootstrap para a estimação de Ψ.

Baseado na igualdade $E[f(X)] = f_Z(0)$, (onde $Z = X_1 - X_2$, e X_1 e X_2 variáveis aleatórias independentes e identicamente distribuídas com função de distribuição acumulada F_x) . e usando o método de estimação de densidade via nucleo estimador, temos o seguinte procedimento bootstrap (Efron (1984)):
Passo 1: obtenção de A:

1. Serão selecionadas duas amostras de forma independente, com reposição, cada uma com tamanho m (suficientemente grande) da variável aleatória X

2. posteriormente será construída um vetor composta pelas diferenças entre cada um dos elementos da primeira amostra com os elementos da segunda amostra.

3. Estimaremos a densidade deste vetor usando o método de Kernel para estimativas de densidade.

4. Avaliaremos esta densidade estimada no ponto zero, obtendo assim, uma estimativa para A.

Passo 2 : Estimação de B:

1. Inicialmente iremos dividir os possíveis valores de Y em k intervalos.

2. Para cada um destes k intervalos faça:

   (a) selecionamos duas amostras e calcularemos as diferenças entre os valores das duas amostras conforme o Passo 1 .

   (b) Estimaremos a densidade deste vetor usando o método de Kernel .

   (c) Avaliaremos esta densidade estimada no ponto zero.

   (d) Ainda neste intervalo, multiplicaremos o valor obtido da densidade do vetor de diferenças no ponto zero pelo valor da probabilidade de selecionarmos um valor de Y neste intervalo considerado.

3. Após repetir este procedimento para os k intervalos teremos que o valor de B,que é a soma dos k valores obtidos por o procedimento mencionado acima

Passo 3 : Calcularemos a nossa medida:

$$\frac{B - A}{B}$$

Nos investigamos a aproximação do calculo de nossa medida pelo bootstrap citado acima. Geramos 20 pares X e Y de uma normal bivariada com os seguintes parâmetros:

$\mu_X = \mu_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 5$ e $\rho = 0.5$

Nesta investigação reamostramos 200 vezes, a distribuição das quantidades A, B e da nossa medida. Na figura 2, temos o gráfico de uma normal bivariada em uma destas gerações:

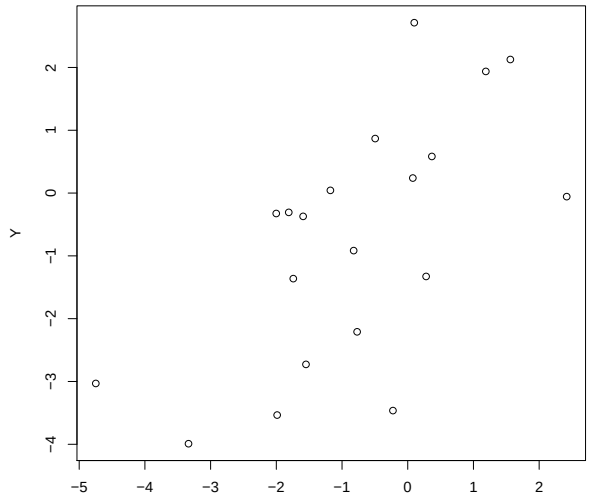

Figura 2: Gráfico dos 20 pontos gerados de uma normal bivariada com os parâmetros:$\mu_X = \mu_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 5$ e $\rho = 0.5$.

Para uma normal bivariada com os parâmetros do item anterior temos que a nossa medida assume o valor:

$$\Psi = GK = \frac{B - A}{B} \quad = \quad 0.134$$

Sendo

$$A = \frac{1}{2\sigma_X\sqrt{\pi}} = \frac{1}{2\sqrt{5\pi}} = 0.126, \text{ e:}$$

$$B = \frac{1}{2\sigma_X\sqrt{\pi(1-\rho^2)}} = \frac{1}{2\sqrt{5\pi(1-0.5^2)}} = 0.146$$

Os resultados da simulação estão sintetizados na figura 3.Observando esta figura podemos dizer que o procedimento bootstrap nos fornece uma boa ferramenta de estimação para a nossa medida de associação. Isto porque é perceptível, no histograma à esquerda, que as medidas obtidas via simulação bootstrap ficaram em torno do valor calculado no exemplo(0.134).

3.1 Aplicação

Uma aplicação real foi feita na análise espaço-temporal dos arrombamentos a residências em Belo Horizonte. Os dados consistem dos tempos de ocorrência e das coordenadas dos locais onde houve roubos à residência em Belo Horizonte no período de Janeiro de 1995 a Dezembro de 2005. Ao todo foram registrados 2688 casos nesse período considerado. Os dados ao longo dos anos foram os seguintes: 89, 82, 87, 137, 119, 248, 180, 260, 463, 508, 515.

Para verificar se existe independência entre as váriaveis tempo e espaço deste problema aplicamos o teste de Knox . Considerando uma distância crítica de 1000m e um tempo crítico de 15 dias observa-se que o índice de Knox foi significativo , a 5% de significância, para quatro anos: 1999, 2001, 2004 e 2005, ou seja, temos evidência as variáveis tempo e espaço não são independentes para estes anos.

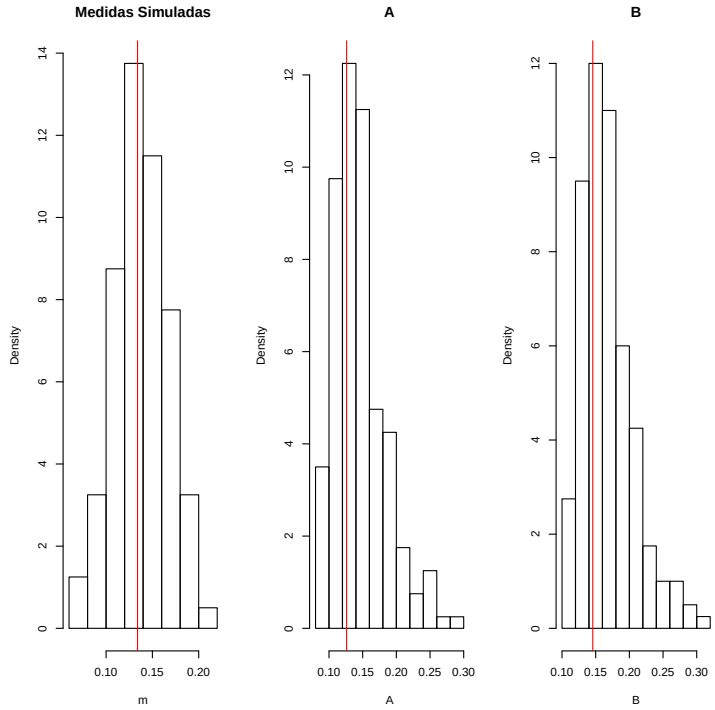

Figura 3: Gráfico das medidas de associação simuladas com o procedimento bootstrap, nos quais a linha vermelha representa a quantidade real: Gráfico à esquerda: medidas de associação obtidas via bootstrap, gráfico do centro: medidas obtidas com o bootstrap da quantidade A e o gráfico à direita são as medidas da quantidade B.

No entanto, na prática, não encontrados um número grande de vizinhos no espaço e no tempo a ponto de ter a dependencia entre espaço e tempo tem pouco significado real. Utilizaremos aqui a nossa medida de associação (Ψ) para tentar mensurar a associação entre espaço tempo para os dados de arrombamento de Belo Horizonte. Com este intuito , usamos o procedimento Bootstrap proposto com uma reamostragem de tamanho 200.

Os resultados mostram valores de Φ baixo(muito próximos de 0.1) para os anos de 1999, 2001, 2004 e 2005, o que mostra que a medida de associação Ψ consegue mostra que apesar das variáveis tempo e espaço não serem indendepentes a associação destas variáveis pode não ser tão significativa do ponto de vista prático.

4 Conclusões

A medida de associação Ψ tem boas propriedades tais como: ter os seus valores entre 0 e 1; se as variáveis são independentes, o índice é zero; tem uma interpretação no sentido do quanto o conhecimento de uma das variáveis nos torna aptos para prever os valores da outra. Além de ser interpretação de ser a proporção que a informação de uma das variáveis nos ajuda a acertar a predição de cada valor da outra variável.

A medida de associação entre espaço e tempo proposta neste trabalho tem o inconveniente de usar densidades de probabilidade em sua construçãoo que conseguimos superar com um procedimento bootstrap que estima as quantidades necessárias através da estimação de densidades via núcleo estimadores. Para dados simulados e dados reais o Ψ conseguiu bons resultados, o que nos da uma boa evidência da eficácia da medida proposta.

Referências

[1] DETTE, H. e NEUMEYER, N. Nonparametric analysis of covariance. *The Annals of Statistics*, 29(5):1361-1400, 2001.

[2] BOWMAN, A. W. e AZZALINI,A. *Applied Smoothing Techiniques for Data Analysis.Oxford Statistical Science Series*, 18, 1997.

[3] EFRON, B. *The Jacknife, the Bootstrap and Other Resampling Plans*. Department of Statistics Stanford University, Philadelphia, 1982.

[4] GOODMAN, L. A. *Some alternatives to ecological correlation, The American Journal of Sociology* 6: 610-625, 1964.

[5] GOODMAN, L. A. e KRUSKAL, W. H. *Measures of association for cross classifications, American Statistical Association* 49: 732-764, 1954.

[6] KNOX, E. *The detection of space-time interactions, Applied Statistics* 13: 25-29, 1964.

[7] MILES, M. B. e HUBERMAN, A. M. *Qualitative Data Analysis: A Sourcebook of New Methods, Sage Publ.*, 1984

[8] SHERVISH, M. J., SEIDENFELD, T. e KADANE, J. B. *Proper scoring rules, dominated forecasts and coherence, Decision Analysis* 6: 202-221, 2005.

[9] SIMONOFF, J. S. *Smoothing Methods in Statistics*. Springer, New York, 1996.

[10] WEI, B.C., HU, Y.Q. and FUNG, W.K. Generalized leverage and its applications. *Scandinavian Journal of Statistics* **25**, 25-37, 1998.

[11] WU, C. F. J. Jackkinfe, bootstrap and other resampling methods in regression analysis. *Ann. Stat.,* **14**, 1261-1295, 1986.

[12] YOUNG, S. G. e BOWMAN, A. W. Nonparametric analysis of covariance. *Biometrics*, 51:920-931, 1995.