Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Proposta de um teste contra a hipótese de completa aleatoriedade espacial para configuração pontual construído via Bootstrap não paramétrico

Gilberto Rodrigues Liska[1] [6]
Marcelo Ângelo Cirillo[2]
João Domingos Scalon[3]
Fortunato Silva de Menezes[4]
Guido Gustavo Humada Gonzalez[5] [6]

1 Introdução

Na análise de configurações espaciais de pontos, a variável de interesse a ser analisada é a localização espacial dos eventos (pontos). Essa análise é considerada uma questão importante em diversas áreas do conhecimento tais como: epidemiologia (distribuição de casos de doenças), fisiologia (distribuição de células em um tecido), materiais compósitos (distribuição de partículas em uma matriz de metal), floresta (distribuição de plantas), etc [1, 3, 4]. Um dos primeiros passos realizados na análise de configurações pontuais no espaço é verificar se a distribuição dos eventos apresenta completa aleatoriedade espacial (CAE) [4]. Essa verificação pode ser feita através de diversos procedimentos (quadrantes, distâncias, etc.) utilizando gráficos e testes de hipótese [3, 4]. Os testes baseados em quadrantes devem ser utilizados com certa cautela, pois o número de quadrantes influencia no resultado dos testes e não levam em consideração a localização dos eventos [3]. Nesse sentido, os testes baseados em distâncias se tornam preferidos por utilizarem a distância dos pontos, bem como a distribuição de probabilidade das mesmas [3].

A função G é a função de distribuição dos vizinhos com distância mais próxima (distribuição evento – evento), de uma configuração pontual [4]. Sob a hipótese de CAE, a função *G* segue um processo homogêneo de Poisson e a inferência é feita comparando-se os valores teóricos sob a hipótese de CAE e os valores estimados por essa função [4]. Contudo, a distribuição amostral da função *G* é desconhecida, o que impossibilita a construção de um

[1] DEX-UFLA, e-mail: gilbertoliska@hotmail.com
[2] DEX-UFLA, e-mail: macufla @ dex.ufla.br
[3] DEX-UFLA, e-mail: scalon@dex.ufla.br
[4] DEX-UFLA, e-mail: fmenezes@dex.ufla.br
[5] DEX - UFLA. e-mail: gustavohumad@hotmail.com
[6] Agradecimentos à FAPEMIG pelo apoio financeiro.

teste formal para testar a hipótese de CAE. Uma alternativa seria encontrar a distribuição da função *G* pelo método de Bootstrap.

O bootstrap, desenvolvido por Efron na década de 70, pode ser utilizado em muitas situações. É baseado em uma simples, porém, poderosa ideia de que a amostra representa a população, logo características análogas da amostra devem nos dar informações sobre as características da população. O Bootstrap auxilia o aprendizado sobre essas características da amostra tomando reamostras (amostras com reposição da amostra original) e usamos essa informação para inferir sobre a população [5].

Baddeley [1] utiliza o método bootstrap para testar a hipotese de CAE a partir de reamostras de subáreas dentro da área de estudo. O objetivo do presente trabalho é mostrar que ao aplicar o método bootstrap diretamente nas coordenadas dos eventos e calcular a função *G* para as reamostras também é possível construir um teste contra a hipótese de CAE.

2 Material e métodos

Para ilustrar o teste proposto, serão utilizados os conjuntos de dados *cells, redwood japanesepines* que estão disponíveis em [1] e que foram exaustivamente analisados por Baddeley [1] e Diggle [4] utilizando várias funções baseadas em distâncias tais como as funções *F*, *G* e *K*. Esses conjuntos de dados são paradigmas usuais para distribuições com regularidade, agrupamentos e CAE, respectivamente.

A função G é definida como sendo a probabilidade de encontrar uma distância y_i menor que uma distância *y* em que y_i a distância entre o i-ésimo evento e o seu vizinho (evento) mais próximo. Assim, a função G é dada por $G(y) = P[y_i < y]$. Um estimador para $G(y)$ pode ser obtido a partir da função de distribuição das distâncias de y_i, ou seja,

$$\hat{G}(y) = \frac{\#(y_i < y)}{n}$$

em que *n* é o número de eventos e "#" representa "número de" [4]. A ideia é comparar $\hat{G}(y)$ com uma distribuição acumulada sob a hipótese de CAE.

Sob a suposição de CAE, os eventos seguem uma distribuição Poisson com média μ e, assim, a probabilidade de ter, pelo menos, um evento até a distância *y* é dada pela sua distribuição acumulada [2], ou seja,

$$G(y) = 1 - P[X = 0] = 1 - e^{-\lambda\pi y^2}.$$

Se $\hat{G}(y)$ for maior que $G(y)$, para uma particular distância *y*, então temos um número maior de eventos dentro daquela distância do que seria esperado sob a hipótese de CAE, caracterizando *agrupamento* de eventos naquela distância. Caso contrário, teríamos *regularidade*. Deve-se observar que a função G é válida apenas para áreas infinitas. Na prática temos áreas finitas, logo é necessário que seja feito uma correção para o efeito de borda. Existem vários métodos para fazer a correção de borda [1, 3, 4]. Neste trabalho, utiliza-se a correção em que os eventos próximos à borda são ponderados em relação ao seu vizinho mais próximo [4].

O processo de reamostragem Bootstrap consiste em reamostrar *B* amostras $P^{*(1)}, P^{*(2)}, \ldots, P^{*(B)}$, com reposição, independentes e identicamente distribuídas das *n*. coordenadas da configuração pontual. Podem-se obter as estimativas do parâmetro de interesse, denotado por $\hat{\theta}^*_{(i)}$, para cada amostra, que no caso é $\theta = E\left[G(y)\right]$. Com isso obteremos o vetor $\hat{\boldsymbol{\theta}}^* = \left(\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}\right)$ e a partir do vetor $\hat{\boldsymbol{\theta}}^*$, pode-se obter a distribuição Bootstrap do estimador $\hat{\theta}$.

Uma vez obtido a distribuição empírica do estimador $\hat{\theta}$ pode-se obter intervalos de confiança para θ. O intervalo de confiança Bootstrap baseado nos percentis da distribuição Bootstrap de θ, descrito em [5], é conhecido como intervalo de confiança *p*-Bootstrap. De uma maneira mais formal, o intervalo de confiança pode ser construído seguindo os seguintes passos: (**Passo 1)** Retirar, com reposição, de *P* uma amostra Bootstrap P^*; (**Passo 2)** Da amostra Bootstrap P^*, obter o estimador $\hat{\theta} = E\left[\tilde{G}(y)\right]$; (**Passo 3)** Repetir os passos 1 e 2 B vezes e (P**asso 4)** A partir do vetor $\hat{\boldsymbol{\theta}}^* = \left(\hat{\theta}^*_{(1)} \le \hat{\theta}^*_{(2)} \le \ldots \le \hat{\theta}^*_{(B)}\right)$, para algum nível de significância α, com $0 < \alpha < 1$, o intervalo *p*-Bootstrap com confiança $100 \times (1-\alpha)\%$ é dado por $IC_{(1-\alpha)}(\theta) : \left[\hat{\theta}^*_{(k_1)}; \hat{\theta}^*_{(k_2)}\right]$, em que $k_1 = (B+1)(\alpha/2)$ e $k_2 = (B+1)(1-\alpha/2)$ são os maiores inteiros que não são maiores que $(B+1)(\alpha/2)$ e $(B+1)(1-\alpha/2)$, respectivamente; e $\hat{\theta}^*_{(k_1)}$ é o percentil $100(\alpha/2)\%$ da distribuição Bootstrap empírica; e $\hat{\theta}^*_{(k_2)}$ o percentil $100(1-\alpha/2)\%$ da função de distribuição Bootstrap empírica [5].

Todos os cálculos foram realizados utilizando as bibliotecas *spatstat* e *boot* do *software* R [6].

3 Resultados e discussões

A tabela 1 apresenta os resultados para a média Bootstrap da função *G*, bem como o respectivo erro padrão Bootstrap e o intervalo de confiança de 95% de confiança para $\theta = E\left[G(y)\right]$.

Tabela 1: Estimativa para a média da função *G* dos dados originais ($\hat{\theta}_0$), média bootstrap $E\left[\hat{\theta}^*\right]$, erro-padrão bootstrap $EP\left[\hat{\theta}^*\right]$ e intervalo com 95% de confiança para o teste de completa aleatoriedade espacial $IC_{95\%}(\theta)$ e decisão. Foram utilizadas 10000 amostras Bootstrap.

Dados	$\hat{\theta}_0 = E\left[\tilde{G}_0(y)\right]$	$E\left[\hat{\theta}^*\right]$	$EP\left[\hat{\theta}^*\right]$	$IC_{95\%}(\theta)$	**Decisão**
Japanesepines	0,7336	0,7660	0,0194	[0,7066; 0,7827]	Aceita H0
Redwood	0,8503	0,7235	0,0207	[0,6982; 0,7787]	Rejeita H0
Cells	0,4965	0,7099	0,0319	[0,6198; 0,7441]	Rejeita H0

Observa-se que para os dados *japanesepines,* a estimativa inicial para a média da função *G*, $\hat{\theta}_0 = E\left[\tilde{G}_0(y)\right]$, é de 0,7336, que está contido no intervalo de confiança de 95% para a média Bootstrap da função G, o que sugere que os pontos apresentam uma configuração aleatória. Por outro lado, para os dados *redwood,* a estimativa inicial para a média da função G, $\hat{\theta}_0 = E\left[\tilde{G}_0(y)\right]$, é de 0,8503, que está fora do intervalo de confiança de 95% para a média Bootstrap da função G, o que sugere que os pontos apresentam uma configuração de agrupamento. Como $\hat{\theta}_0$ é maior do que $\hat{\theta}^*_{(9750)} = 0,7787$, indica que em média existe um excesso de pequenas distâncias. Para os dados *cells,* a estimativa inicial para a média da função *G*, $\hat{\theta}_0 = E\left[\tilde{G}_0(y)\right]$ é de 0,4965, que está fora do intervalo de confiança de 95% para a média Bootstrap da função G, o que sugere que os pontos apresentam uma configuração de regularidade. Como $\hat{\theta}_0$ é menor do que $\hat{\theta}^*_{(250)} = 0,6198$, indica que em média existe um excesso de grandes distâncias.

A figura 1 apresenta o histograma e o gráfico quantil-quantil para o conjunto de dados *japanesepines*. O histograma sugere que a distribuição empírica de $\theta = E\left[G(y)\right]$ é uma normal e esse fato é corroborado pelo gráfico quantil-quantil, uma vez que é mantida a proporcionalidade de um para um considerando-se os quantis da normal padrão versus os quantis observados. Resultados idênticos foram observados para os dados redwood e cells,

porém os resultados não são apresentamos. Os resultados obtidos neste trabalho corroboram os resultados obtidos por Diggle [4], que utilizou as funções F, G e K, conjuntamente com métodos de Monte Carlos. Os resultados também corroboram os obtidos por Badedely [1] que utilizou o método bootstrap em subáreas dentro da área de estudo.

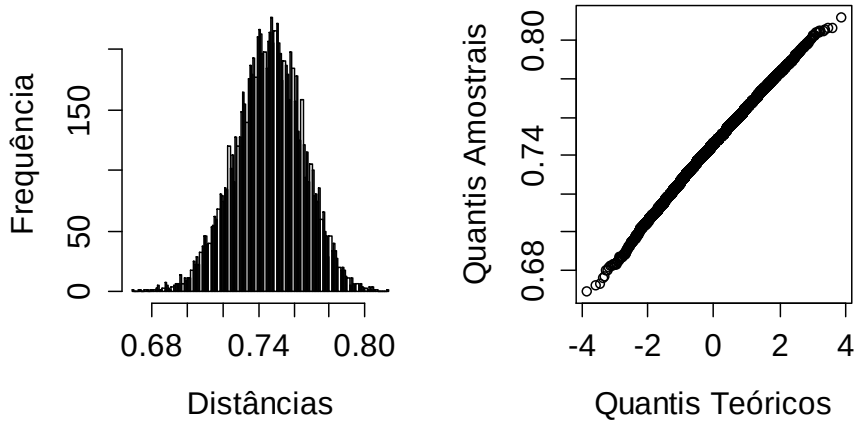

Figura 1: Histograma para os 10000 valores da média bootstrap da função *G* e o respectivo gráfico quantil-quantil normal.

Apesar dos resultados obtidos pela aplicação do método proposto neste trabalho corroborar os resultados obtidos por Badedely [1] e Diggle [4], deve-se observar que o método apresentado possibilita que, eventualmente, dois eventos possam ocorrer no mesmo local, o que para alguns pesquisadores é uma suposição padrão nesse tipo de análise.

4 Conclusões

O método apresenta uma forma intuitiva e inovadora que, além de decidir sobre a hipótese de CAE, traz conclusões sobre a regularidade e o agrupamento de configurações pontuais. Contudo, é necessário realizar simulações para verificar o poder e as taxas de erro do método, o que constitui a próxima etapa do trabalho.

5 Bibliografia

[1] BADDELEY, A. **Analysing spatial point patterns in R**. Workshop notes, version 4.1, CSIRO and University of Western Australia, 2010.
[2] CASELLA G.; BERGER, R. **Statistical Inference**, 2º ed., Duxbury Advanced Series, 660 p., 2002.
[3] CRESSIE, N.A.C. **Statistics for spatial data**. John Wiley and Sons, 1991.
[4] DIGGLE, P.J. **Statistical analysis of spatial point patterns**. Academic Press, 1983.
[5] EFRON, B.; TIBSHIRANI, R. J. **An introduction to the Bootstrap**. Chapman and Hall, New York, 1993.
[6] R DEVELOPMENT CORE TEAM (2013). **R: a language and environment for statistical computing.** Vienna: R Foundation for Statistical Computing. www.r-project.org.