https://eventos.utfpr.edu.br//sicite/sicite2020

Estudos e aplicações em estatística espacial

Studies and applications in spatial statistics

João Vitor Magri da Silva
jsilva.2017@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Roberto Molina de Souza
rmolinasouza@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

RESUMO

A Estatística Espacial engloba modelos estatísticos que possibilitam a incorporação no modelo da localização espacial de fenômenos de interesse que estão distribuídos no espaço. Nesse trabalho foram apresentados, e implementados com auxílio do software livre OpenBUGS, alguns modelos do tipo CAR (*Conditional autoregressive*): Intrínseco, Convolução, Cressie e Leroux. Para aplicar a metodologia proposta foram utilizados dados referentes a média do preço de apartamentos por bairro da cidade de Londrina, norte do estado do Paraná. Pode-se notar que o modelo intrínseco, aquele que leva em conta forte correlação espacial, foi o melhor ajustado, indicando assim a necessidade do uso da estatística espacial nesta aplicação em detrimento, por exemplo, de uma análise de regressão linear múltipla apenas. Estes resultados são relevantes para a área de Estatística aplicada, uma vez que contribuem na visibilidade de uma metodologia importante, aplicada a um estudo de caso.

PALAVRAS-CHAVE: Análise espacial. Estatística matemática. Análise bayesiana.

ABSTRACT

Spatial Statistics is called a branch of statistics that allows the incorporation in the model of the spatial location of phenomena of interest that are distributed in space. In this work, some CAR (*Conditional autoregressive*) models were presented - these being models: Intrinsic, Convolution, Cressie and Leroux - and were later applied and implemented with the aid of the OpenBUGS free software. To apply the proposed methodology, data referring to the average price of apartments by neighborhood in the city of Londrina, northern Paraná state, were used. It could be noted that the intrinsic model, the one that takes into account strong spatial correlation, was the best adjusted, thus indicating the need to use spatial statistics in this application to the detriment, for example, of a multiple linear regression analysis only. These results were relevant to the area of applied statistics, since they contribute to the visibility of an important methodology, applied to a case study.

KEYWORDS: Spatial analysis. Mathematical statistics. Bayesian analysis.

INTRODUÇÃO

A Estatística Espacial é intitulada um ramo da estatística que possibilita a incorporação no modelo da localização espacial de fenômenos de interesse que estão distribuídos no espaço. Para Haining e Haining (2003), a análise espacial se faz representada por um conjunto de técnicas e modelos que usam explicitamente a localização espacial; possibilitando a descrição de eventos que ocorrem no espaço e percepção visual sobre a disposição espacial do problema em questão.

Embora o foco deste trabalho esteja na aplicação dos modelos de estatística espacial na precificação de imóveis, existem poucas aplicações com esta especificidade na literatura. Por outro lado, estes modelos têm sido bastante explorados na área da saúde.

Werneck e Struchiner (1997) enfatizaram a importância em confrontar a alocação espacial de doenças com espaço e tempo, sendo esta abordagem um dos temas mais debatidos e estudados na área epidemiológica atualmente, indicando que a distribuição espacial de tais eventos pode ser fator expressivo na conclusão sobre ocorrências de alguns eventos.

Neste sentido, é imprescindível dizer que o mapeamento ligado a propagação de doenças permite analisar espacialmente a origem dos focos em uma cidade, juntamente visualizar a correlação dos casos com variáveis demográficas e auxiliar os órgãos responsáveis a serem incisivos no combate destes focos (BRASIL, 2007).

Rampaso (2014), cita alguns exemplos de aplicação envolvendo o mapeamento de doenças e da criminalidade em uma região de interesse; poluentes do ar em um centro urbano; e quantidade de chuva em uma determinada região. Carvalho et al. (2004), ressalta o caso de epidemiologistas estudando sobre a concentração de alguma doença e a investigação de roubos em uma determinada região como casos de interesse para a investigação utilizando Estatística Espacial.

Este trabalho tem o objetivo de estudar a estatística espacial bem como a aplicação deste ferramental estatístico em um problema aplicado fazendo o uso do Software livre OpenBUGS (LUNN et al., 1997). De forma específica, realizar o estudo da Estatística Espacial bem como seus aspectos computacionais; obter um conjunto de dados para modelagem estatística e aplicação das ferramentas; e analisar os dados, comparar os modelos e concluir a partir dos resultados.

METODOLOGIA

Nesta seção serão apresentados alguns modelos do tipo CAR (*Conditional autoregressive*) propostos na literatura, para serem aplicados posteriormente e implementados com auxílio do software livre OpenBUGS.

A ideia principal de um modelo para análise espacial está em relacionar os possíveis efeitos de vizinhança entre as regiões definidas pelo pesquisador. Neste sentido, a distribuição normal multivariada, a partir de sua matriz de covariâncias, é extremamente útil nesta construção. Logo, denota-se:

$$\emptyset \sim NM\left(\boldsymbol{\mu}, \Sigma(\boldsymbol{\theta})\right) \quad (1)$$

em que $\boldsymbol{\mu}$ é um vetor de médias e $\Sigma(\boldsymbol{\theta})$ é a matriz de covariâncias.

As distribuições condicionais completas de um modelo CAR geral podem ser definidas como:

$$\emptyset_i|\emptyset_{-i} \sim N\left(\mu_i + p\sum_{j\sim i} C_{ij}(\emptyset_j - \mu_j), \sigma^2 m_{ii}\right) \quad (2)$$

em que p captura a dependência espacial ($p = 0$ resulta em independência espacial); $j \sim i$ denota a presença de vizinhança entre as áreas j e i; C_{ij} é a matriz de associação espacial (com os elementos da diagonal igual a 0); σ^2 é a variância geral do modelo; e m_{ii} uma matriz diagonal.

Um grande facilitador na especificação de ϕ se dá com o uso do software livre OpenBUGS, utilizando a distribuição $car.normal$, em que basta especificar apenas a matriz de adjacências, ou seja, qual a relação de vizinhança entre as regiões, uma matriz de pesos (caso seja necessária), a quantidade de vizinhos de cada região e um parâmetro, que aqui podemos denotar por τ, para capturar a precisão (ou inverso da variância) deste efeito espacial detonado por ϕ.

A seguir, serão apresentados quatro modelos espaciais encontrados na literatura e que serão usados na aplicação.

O modelo intrínseco (ICAR) proposto por Besag et al. (1991) pode ser considerado um dos modelos CAR mais simples. Esse modelo é obtido a partir da Equação (1) em que $c_{\text{ij}} = 1/n_i$ se as áreas i e j forem adjacentes e 0 caso contrário; $m_{\text{ii}} = 1/n_i$, em que n_i é o número de vizinhos da área i. Estas implicações, implicam em $p = 1$. A matriz equivale a uma matriz de pesos normalizados. Assume-se também $\mu_i = 0, i = 1,\ldots,n$. Logo, a distribuição condicional completa do modelo ICAR pode ser dada por:

$$\emptyset_i|\emptyset_{-i} \sim N\left(\frac{1}{n_i}\sum_{j\sim i} \emptyset_j, \frac{\sigma^2}{n_i}\right) \quad (3)$$

O modelo ICAR pode apresentar algumas desvantagens. A força da dependência espacial entre os efeitos aleatórios é sempre considerada máxima, uma vez que $p = 1$, sendo o modelo adequado apenas na presença de forte correlação espacial. Outro ponto importante é que o parâmetro de variância σ^2 captura tanto superdispersão quanto dependência espacial, não permitindo a estimação destas fontes de variação separadamente (LEROUX, 2000)

Também proposto por Besag et al (1991), o modelo de convolução adiciona um efeito aleatório sem estrutura espacial ao modelo intrínseco. Esse modelo é dado por:

$$\varphi_i = \theta_i + \psi_i \quad (4)$$

em que $\theta_i \sim N(0, \sigma_\theta^2)$ e $\boldsymbol{\psi} \sim \text{ICAR}(W, \sigma_\psi^2)$.

A ideia da inclusão deste efeito aleatório é justificada por Breslow (1984) com o objetivo de absorver a variação adicional não capturada pelo efeito aleatório espacial intrínseco. Logo, se o termo ψ for dominante, então os riscos são estimados considerando uma maior estrutura espacial. Caso contrário, então a estimação dos riscos será concentrada em torno da média geral μ.

O modelo CAR próprio ou modelo de Cressie (2015) captura diferentes níveis de estatística espacial. Diferente dos modelos anteriores, em que o parâmetro $p = 1$, a ideias deste modelo é estimar o valor de p. Este modelo pode ser definido por:

$$\phi \sim NM(\boldsymbol{\mu}, \sigma^2 \boldsymbol{Q}^{-1}) \quad (5)$$

em que o *ij-ésimo* elemento da matriz $\boldsymbol{Q}$ é definido como:

$$q_{ij} = n_i, \quad \text{se } i = j \quad (6)$$

$$q_{ij} = -p, \quad \text{se } i \sim j$$

$$q_{ij} = 0, \quad \text{caso contrário}$$

A distribuição condicional completa univariada para os efeitos $\emptyset_i$ é dada por:

$$\emptyset_i | \emptyset_{-i} \sim N\left(p \frac{1}{n_i} \sum_{j \sim i} \emptyset_j, \frac{\sigma^2}{n_i}\right) \quad (7)$$

em que $0 \leq p \leq 1$.

Se $p = 0$, tem-se independência espacial, uma vez que os valores irão se concentrar em torno da mesma média zero. Para valores de p próximos a 1, existem indícios de forte correlação espacial. Quando $p = 1$, tem-se o modelo intrínseco.

Um modelo um pouco mais geral, proposto por Leroux et al. (2000) representa o conjunto de efeitos $\emptyset$ por uma distribuição normal multivariada:

$$\phi \sim NM(\mathbf{0}, \sigma^2 [p\boldsymbol{D} + (1-p)\boldsymbol{I}]^{-1}) \quad (8)$$

Detonando-se a matriz de precisão por, $\boldsymbol{L} = p\boldsymbol{D} + (1-p)\boldsymbol{I}$, em que $\boldsymbol{I}$ denota uma matriz identidade de ordem n e a matriz $\boldsymbol{D}$ é definida por:

$$d_{ij} = n_i, \quad \text{se } i = j \quad (9)$$

$$d_{ij} = -1, \quad \text{se } i \sim j$$

$$d_{ij} = 0, \quad \text{caso contrário}$$

Note que, se $p = 0$, tem-se $\boldsymbol{L} = \boldsymbol{I}$ sendo assim um modelo com efeitos aleatórios independentes. Se $p = 1$, tem-se o modelo intrínseco. A distribuição condicional completa univariada para $\emptyset$ é dada por:

$$\emptyset_i | \emptyset_{-i} \sim N\left(\frac{p}{n_i p + 1 - p} \sum_{j \sim i} \emptyset_j, \frac{\sigma^2}{n_i p + 1 - p}\right) \quad (10)$$

Dada a complexidade destes 4 modelos, o uso da Inferência Bayesiana pode contornar diversos problemas que poderiam ser encontrados nos métodos usuais de estimação, como os estimadores de máxima verossimilhança. Muitos dos aspectos teóricos da Inferência Bayesiana podem ser encontrados em Achcar et al. (2019) e todos os passos para a implementação dos modelos aqui apresentados podem ser encontrados em Rampaso (2014).

ESTUDO DE CASO

Para aplicar a metodologia proposta serão utilizados dados referentes a média do preço de apartamentos por bairro da cidade de Londrina. Além da

modelagem espacial, também foram consideradas variáveis explicativas coletadas para compor o modelo e possivelmente explicar a variável resposta (preço do imóvel), sendo: Área total (m^2), Cozinha planejada (Sim ou não), Área de lazer (Sim ou não), Número de dormitórios e Número de vagas na garagem.

A coleta dos dados foi feita a partir do *website* Sub100. Este site é uma ferramenta de busca de imóveis com diversas opções de pesquisa. No caso deste trabalho, considerou-se apenas apartamentos, na cidade de Londrina, no estado do Paraná, nos diversos bairros da cidade, que estivessem à venda, no período de fevereiro a junho de 2018. Não foi utilizado nenhum limitador de preços ou de metragem.

MODELAGEM

Para a incorporação das covariáveis no modelo espacial, o parâmetro μ_i destes modelos receberá uma regressão linear múltipla, ou seja,

$$\mu_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_6 X_{6i} \quad (11)$$

em que α é o intercepto ou média geral; $\beta_1, \ldots, \beta_6$ os parâmetros que mensuram o efeito das seguintes variáveis, para o i-ésimo bairro:

a) é a média do número de dormitórios dos apartamentos;

b) é o proporção de apartamentos com cozinha planejada;

c) é o proporção de apartamentos com área de lazer;

d) é a média da área privada em m^2 dos apartamentos;

e) é a média da área total em m^2 dos apartamentos;

f) é a média do número de vagas na garagem.

Nas Tabelas 1 a 2 são apresentados a estimação dos parâmetros (média a posteriori), desvio padrão e intervalos de credibilidade com 95% para os 4 modelos propostos. Para o ajuste dos modelos foi considerado o valor do imóvel em reais dividido por 10000.

Tabela 1 – Estimação dos parâmetros (Média a posteriori), desvio padrão (D. P.) e intervalos de credibilidade (IC_r) com 95% para os modelos de intrínseco e de convolução.

Parâmetros	Intrínseco		Convolução	
	Média (D.P.)	ICr(95%)	Média (D.P.)	ICr(95%)
α	2,338 (0,285)	(1,777;2,899)	2,298 (0,292)	(1,717;2,877)
β_1	-0,085 (0,127)	(-0,339;0,163)	-0,078 (0,126)	(-0,325;0,163)
β_2	0,037 (0,162)	(-0,278;0,360)	0,060 (0,172)	(-0,273;0,402)
β_3	0,356 (0,163)	(0,036;0,677)	*0,353 (0,167)*	(0,027;0,686)
β_4	0,003 (0,003)	(-0,002;0,009)	0,002 (0,003)	(-0,004;0,008)
β_5	0,000 (0,002)	(-0,005;0,005)	0,000 (0,003)	(-0,005;0,006)
β_6	0,539 (0,153)	(0,235;0,842)	0,557 (0,169)	(0,229;0,902)
σ_ϕ	0,417 (0,080)	(0,277;0,589)	0,316 (0,118)	(0,059;0,537)
σ_ψ	-	-	0,132 (0,069)	(0,010;0,266)
DIC	337,0		348,6	

Fonte: Autoria própria (2020).

Tabela 2 – Estimação dos parâmetros (Média a posteriori), desvio padrão (D. P.) e intervalos de credibilidade (IC_r) com 95% para os modelos de Cressie e Leroux.

Parâmetros	Intrínseco		Convolução	
	Média (D.P.)	ICr(95%)	Média (D.P.)	ICr(95%)
α	2,311 (0,283)	(1,761;2,870)	2,315 (0,299)	(1,719;2,898)
β_1	-0,090 (0,115)	(-0,320;0,135)	-0,085 (0,121)	(-0,330;0,151)
β_2	0,020 (0,162)	(-0,295;0,344)	0,033 (0,164)	(-0,285;0,363)
β_3	0,378 (0,158)	(0,068;0,695)	*0,369 (0,160)*	(0,055;0,687)
β_4	0,003 (0,003)	(-0,003;0,008)	0,003 (0,003)	(-0,003;0,008)
β_5	0,000 (0,003)	(-0,005;0,006)	0,000 (0,003)	(-0,005;0,005)
β_6	0,551 (0,158)	(0,245;0,865)	0,549 (0,159)	(0,239;0,866)
p	0,691 (0,244)	(0,099;0,988)	0,647 (0,242)	(0,128;0,986)
σ	0,453 (0,092)	(0,285;0,643)	0,368 (0,086)	(0,220;0,555)
DIC	384,6		384,8	

Fonte: Autoria própria (2020).

Para decidir qual o melhor modelo entre os 4 propostos, o indicador numérico *Deviance Information Criterion* (DIC) (SPIEGELHALTER et al., 2002), presente nas Tabelas 1 a 2 e calculado para cada modelo é bastante útil.

O critério DIC indica que, quanto menor o valor de DIC de um modelo comparado aos demais, melhor o ajuste deste modelo aos dados. Neste sentido, observa-se que o modelo intrínseco, com menor valor de DIC entre os 4 modelos, foi o que melhor ajustou-se aos dados. Isto indica, por exemplo, que os dados analisados têm forte correlação espacial.

Além da forte correlação espacial entre os bairros, tomando como base para a análise o modelo intrínseco, observa-se que os parâmetros β_3 e β_6 apresentam evidências de relevância no modelo, uma vez que no intervalo de credibilidade construída para os mesmos, o valor 0 não está presente.

Assim, β_3 e β_6 estão relacionados as variáveis área de lazer e vagas na garagem. Como o sinal da estimativa destes parâmetros é positivo, isto indica que possuir área de lazer ou mais vagas de garagem valorizam o apartamento.

Finalmente, para o modelo intrínseco foi gerado o mapa espacial dos valores dos apartamentos (em R$) da cidade de Londrina segundo os bairros. Observa-se na Figura 1 que os apartamentos mais bem valorizados se encontram na parte central da cidade e à Oeste, na região da Gleba Palhano. A região Norte possui os apartamentos de menor valor.

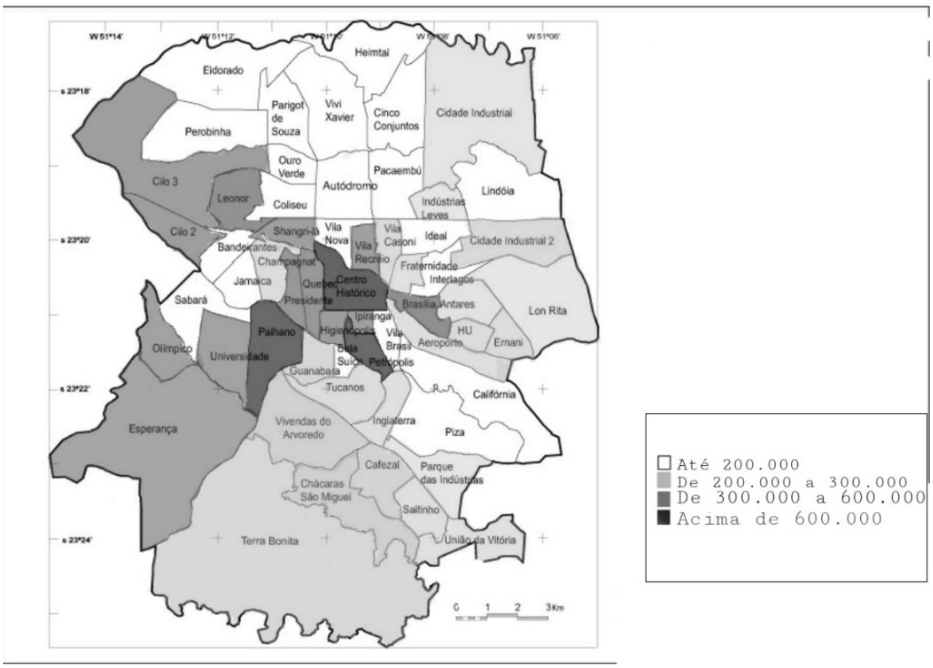
CONCLUSÃO

Os modelos de estatística espacial vêm ganhando especial atenção na modelagem estatística nos últimos anos, uma vez que há uma disponibilidade maior de computadores de alto desempenho bem como a construção de algoritmos que dão melhor desempenho para a estimação de parâmetros.

O uso da metodologia de estatística espacial não pode ser considerado simples, uma vez que necessita de um pouco de conhecimento do usuário em

inferência Bayesiana e estatística computacional, como pré-requisitos. Logo, neste trabalho procurou-se destacar a construção de modelos do ponto de vista metodológico da estatística espacial.

Figura 1 – Mapa do preço médio (R$) dos apartamentos da cidade de Londrina



Fonte: Autoria própria (2020).

No estudo de caso apresentado, pode-se notar que o modelo intrínseco, aquele que leva em conta forte correlação espacial, foi o melhor ajustado, indicando assim a necessidade do uso da estatística espacial nesta aplicação em detrimento, por exemplo, há uma análise de regressão linear múltipla apenas. Em uma análise de regressão linear a correlação espacial não seria capturada.

Logo, estes resultados são relevantes para a área de Estatística aplicada, uma vez que contribuem na visibilidade de uma metodologia importante, em um exemplo com dados reais de interesse local na cidade de Londrina, norte do Paraná.

REFERÊNCIAS

ACHCAR, J. A.; BARROS, E. A. C.; SOUZA, R. M.; MARTINEZ, E. Z. **Uma introdução aos métodos bayesianos aplicados à análise de dados**. 1. ed. Timburi: Cia do E-book, 2019.

BESAG, J.; YORK, J; MOLLIÉ. A. Bayesian image restoration, with two applications in spatial statistics. **Annals of the institute of statistical mathematics**, v. 43, n. 1, p. 1-20. 1991.

BRASIL. Introdução à estatística espacial para a saúde pública. *In*: SANTOS, S. M.; SOUZA, W.V. (org.). **Série**: Capacitação e atualização em geoprocessamento em saúde. Brasília: Ministério da Saúde, 2007.

BRESLOW, N. E. Extra-Poisson Variation in Log-Linear Models. **Journal of the Royal Statistics Society Series C**, v. 33, n. 1, p. 38-44. 1984.

CARVALHO, M. S.; CÂMARA, G.; CRUZ, O. G.; CORREA, V. Análise de Área. *In*: DRUCK, S.; CARVALHO, M.; CÂMARA, G.; MONTEIRO, A. (org.). **Análise espacial de dados geográficos:** Cap. 5. Brasília: EMBRAPA, 2004.

CRESSIE, N. **Statistics for Spatial Data.** 2. ed. Hoboken: Wiley, 2015. (Wiley Series in Probability and Statistics)

HAINING, R. P.; HAINING, R. **Spatial data analysis: theory and practice.** 1. ed. Cambridge: University Press, 2003.

LEROUX, B.; LEI, X.; BRESLOW, N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. *In*: Halloran M.E., Berry D. (org.). **Statistical Models in Epidemiology, the Environment, and Clinical Trials.** The IMA Volumes in Mathematics and its Applications, vol. 116. Springer, New York, NY, 2000.

LUNN, D. J.; THOMAS, A.; BEST. N.; SPIEGELHALTER, D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. **Statistics and computing**, v. 10, n. 4, p. 325-337. 1997.

RAMPASO, R. C. **Análise bayesiana de dados espaciais explorando diferentes estruturas de variância**. 2014. Dissertação (Mestrado em Matemática Aplicada e Computacional) – Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Ciências e Tecnologia, Presidente Prudente, 2014.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 64, n. 4, p. 583-639. 2002.

WERNECK, G. L.; STRUCHINER, C. J. Estudos de agregados de doença no espaço-tempo: conceitos, técnicas e desafios. **Cadernos de Saúde Pública**, v. 13, n. 4, p. 611-624. 1997. Disponível em: https://www.scielo.br/pdf/csp/v13n4/0146.pdf. Acesso em: 01 set. 2020.