REVISTA DO SEMINÁRIO INTERNACIONAL DE ESTATÍSTICA COM R | **ISSN: 2526-7299**
VOL 5, Nº 2, JULHO DE 2024
VIII SEMINÁRIO INTERNACIONAL DE ESTATÍSTICA COM R - AI IN DATA SCIENCE

MODELO ESPACIAL BINOMIAL NEGATIVO INFLACIONADO DE ZEROS

Daniel Claudiano Cabral Pinto[1] e Patrícia Lusié Velozo da Costa[2]

Resumo

Em muitas aplicações reais, dados de contagem são utilizados para descrever a frequência do fenômeno estudado em uma determinada região em um recorte de tempo de uma população. Em muitos casos, o fenômeno pode ser majoritariamente ausente na população, consequentemente a contagem deste sendo inflada por zeros em suas observações, além de possuir uma dependência espacial entre regiões vizinhas. A proposta deste artigo é comparar a performance do modelo linear generalizado binomial negativo usual com um modelo análogo, porém adaptado à inflação de zeros com autocorrelação espacial condicional (CAR), quando aplicados a um conjunto de dados espaciais inflacionados de zero. Utilizou-se a linguagem *R*. A inferência foi realizada sob a ótica bayesiana e a performance do modelo foi analisada pelas medidas de qualidade de ajuste DIC e MAPE. Os resultados apontam que o modelo adaptado à inflação de zeros e com estrutura espacial possui significativos ganhos de performance comparados ao modelo sem estrutura espacial e sem inflação de zeros.

Palavras-chave: Inflação de zeros**,** modelo linear generalizado, inferência bayesiana, estatística espacial.

Abstract

In many real-world applications, count data are used to describe the frequency of the studied phenomenon in a specific region within a population over a period of time. In many cases, the phenomenon may be largely absent in the population, leading to the inflation of zeros in its observations, as well as having spatial dependence among neighboring regions. The aim of this study is to compare the performance of the usual negative binomial generalized linear model with an analogous model, but adapted to zero inflation with conditional spatial autocorrelation (CAR) when applied to a set of spatial data inflated with zeros. The R language was used. Inference was conducted from a Bayesian perspective, and the model's performance was analyzed using goodness-of-fit measures DIC, MAPE, and MSE. The results indicate that the model adapted to zero inflation and with spatial structure shows significant performance gains compared to the unadapted model.

Keywords: Zero-inflated data, linear generalized model, Bayesian inference, spatial statistics.

[1] Universidade Federal Fluminense (UFF), danielclaudiano@id.uff.br

[2] Universidade Federal Fluminense (UFF), patricialusie@id.uff.br

Introdução

Em muitos casos o fenômeno a ser estudado pode ser majoritariamente ausente na população, consequentemente a contagem deste sendo inflada por zeros em suas observações. Por exemplo, número de pessoas infectadas por uma doença, número de pessoas mortas por conta de acidente de trânsito, mortalidade materna, ocorrência de incêndios florestais, quantidade de produtos de luxo vendidos, quantidade de infecções hospitalares, entre outros.

Quando há um excesso de zeros, a moda dos valores observados costuma ser zero e a média pode resultar em um valor bem distante desse número. Quando o modelo Poisson e binomial negativo são ajustados em conjunto de dados inflacionados de zero, o parâmetro de média estimado assume valor próximo ao da moda (Agresti, 2015). Neste caso, modelos como o Binomial Negativo Inflacionado de Zeros (MBNIZ) e o Poisson Inflacionado de Zeros (MPIZ) são mais apropriados.

Em estatística espacial, dados de área referem-se a dados que estão agregados em unidades de área geográfica, como municípios, bairros, regiões censitárias ou células de uma grade espacial. Esses dados podem incluir uma variedade de informações, como contagens de eventos (por exemplo, número de crimes, número de habitantes), médias de variáveis contínuas (por exemplo, renda média, temperatura média), proporções (por exemplo, percentual da população com ensino superior), entre outros. Eles são importantes para entender como diferentes variáveis estão distribuídas no espaço e como elas interagem entre si em diferentes áreas geográficas. Além disso, os dados de área são essenciais para a tomada de decisões e o planejamento de políticas públicas em níveis regionais ou locais.

A estrutura espacial condicional autoregressiva (CAR) em modelos espaciais oferece uma série de vantagens e benefícios significativos. Primeiramente, a incorporação de uma estrutura CAR permite capturar a dependência espacial entre as unidades geográficas vizinhas, levando em consideração a correlação espacial entre os dados. Isso é crucial em muitas áreas de pesquisa, como epidemiologia, ciências ambientais e economia regional, onde fenômenos adjacentes frequentemente apresentam comportamentos similares ou interagem de maneira significativa. Além disso, a estrutura CAR permite modelar padrões de autocorrelação espacial não explicados por variáveis observadas, ajudando a capturar e explicar a variabilidade residual nos dados. Essa abordagem proporciona uma representação mais completa e precisa da realidade subjacente, resultando em estimativas mais robustas e inferências mais confiáveis. Em resumo, a inclusão de uma estrutura espacial condicional autoregressiva em modelos espaciais aprimora nossa capacidade de entender e interpretar padrões espaciais nos dados.

Objetivo

O objetivo deste artigo é comparar os modelos binomial negativo adaptado a inflação de zeros e com estrutura espacial CAR com o modelo de regressão binomial negativo usual, aplicados a dados espaciais inflacionados de zeros.

Material e Método

Inspirado no modelo de regressão Poisson inflacionado a zeros proposto em Lambert (1992), considere o seguinte modelo Binomial Negativo Inflacionado de Zeros (MBNIZ):

$$M_i \sim Bern(p_i), \quad M_i = 0,\ 1, \quad 0 < p_i < 1.$$

Se $M_i = 1$ então a variável resposta Y_i assume o valor zero. Caso $M_i = 0$ então $Y_i \sim BinNeg(\mu_i, k)$. Ou seja, $Y_i = 0 \cdot I(M_i = 1) + y_i \cdot I(M_i = 0)$, com $y_i = 0, 1, 2, \ldots$, sendo $I(A)$ a função indicadora que resulta em 1 quando a condição *A* é atendida e 0, caso contrário. Dessa forma, tem-se que a distribuição de Y_i pode ser descrita da seguinte forma:

$$P(\mu_i, p_i, k) = \{p_i + (1 - p_i)\left(\frac{k}{\mu_i + k}\right)^k, \ para \ \ y_i = 0, \ (1 - p_i)\left(\frac{k}{\mu_i + k}\right)^k \left(\frac{\mu_i}{\mu_i + k}\right)^{y_i} \binom{y_i + k - 1}{k - 1}, \ para \ y_i$$

Covariáveis são usadas neste modelo para explicar a probabilidade p_i que parametriza a distribuição Bernoulli e o parâmetro de média μ_i da distribuição Binomial Negativa da seguinte forma:

$$ln\left(\frac{p_i}{1 - p_i}\right) = x_i^{E^T} \psi \quad e \quad ln(\mu_i) = x_i^T \beta + b_i,$$

sendo $x_i^{E^T}$ um vetor com p covariáveis associadas à ausência do fenômeno, x_i^T um vetor com q covariáveis associadas a ocorrência do fenômeno, ψ e β vetores de ordem p e q, respectivamente, com os efeitos dos preditores lineares, e b_i representando o efeito aleatório espacial da *i*-ésima unidade.

Considere que o vetor de efeitos aleatórios espaciais $b = (b_1, b_2, \ldots, b_n)^T$ seja distribuído por uma normal multivariada de média zero e matriz de covariâncias proposta por Simpson (2015) e dada pela seguinte forma:

$$\Sigma_b = \left(\tau_b^2\right)^{-1/2} \left[(1 - \phi) I_n + \phi (Q)^{-1}\right]$$

sendo τ_b^2 e ϕ parâmetros de variância e de correlação espacial, respectivamente, I_n a matriz identidade de ordem n e Q uma matriz de vizinhança transformada, na qual a diagonal contém o número total de vizinhos de primeira ordem. Sendo assim, o vetor paramétrico

desconhecido é $\left(\psi, \beta, k, \phi, \tau_b^2\right)$. Considerou-se *a priori* que $\beta \sim NMV\left(0, \Sigma_\beta\right)$, $\psi \sim NMV\left(0, \Sigma_\psi\right)$, $k \sim Gama(1,1)$, $\tau_b^2 \sim Gama(1,1)$ e $\phi \sim Beta(1,1)$, sendo $\Sigma_\beta = \Sigma_\psi$ uma matriz diagonal com todos os elementos da diagonal principal iguais a 1000.

Para avaliar o ganho do modelo espacial inflacionado de zeros comparado ao modelo sem estrutura espacial e sem a estrutura para acomodar a inflação de zeros, considerando em ambos os casos a distribuição binomial negativa, utilizou-se dados simulados.

O conjunto de dados foi gerado utilizando o software *R* considerando que a região geográfica é composta pelos municípios do estado de São Paulo, menos Ilhabela, na construção da matriz de vizinhança. Sendo assim, considerou-se um tamanho de amostra $n = 644$. Utilizou-se o critério *Rainha* para definir a vizinhança dos municípios. Os vetores, $x_i^{E^T}$ e x_i^T, foram geradas de uma normal padrão. Arbitrariamente, fixou-se $\beta^T = (2, 3)$ e $\psi^T = \left(\frac{1}{10}, -6\right)$. Usou-se k = 10, ϕ= 0,7 e τ_b^2=0,5.

Resultados e Discussão

Após a obtenção dos dados simulados, foram comparados o modelo linear generalizado binomial negativo usual com o Binomial Negativo Inflado de Zeros (MBNIZ) com estrutura de dependência espacial condicional autorregressiva (CAR). A inferência dos parâmetros foi realizada sob a abordagem bayesiana, e a estimação foi realizada utilizando Métodos de Monte Carlo via Cadeia de Markov (MCMC), em especial o algoritmo de Monte Carlo Hamiltoniano, fornecido pelo pacote *brms*. A convergência das cadeias foi obtida após 50 mil iterações, sendo 10 mil iterações de aquecimento, com 50 iterações de espaçamento, gerando amostras de tamanho 800 para cada parâmetro. As estimativas pontuais e intervalares estão apresentadas nas tabelas abaixo:

Tabela 1– Estimativas pontuais e intervalares dos parâmetros do modelo Binomial Negativo.

Parâmetro	Média *a posteriori*	Intervalo de Credibilidade de 95%
β_0	3,30	[3,03 ; 3,58]
β_1	3,38	[3,03 ; 3,70]
k	0,8	[0,07 ; 0,09]

Tabela 2– Estimativas pontuais e intervalares dos parâmetros do modelo Espacial Binomial Negativo Inflado a Zeros.

Parâmetro	Média *a posteriori*	Intervalo de Credibilidade de 95%

β_0	2,09	[1,85 ; 2,35]
ψ_0	0,61	[-0,12 ; 1,36]
β_1	3,17	[2,94 ; 3,42]
ψ_1	-5,94	[-9,25 ; -3,68]
k	2,71	[1,23 ; 5,60]
ϕ	0,9	[0,72 ; 1,00]
τ_b^{-2}	1,95	[1,65 ; 2,25]

As medidas de qualidade de ajuste e previsão para comparação dos modelos foram: o critério de informação da Deviance (DIC) e o erro percentual absoluto médio (MAPE).

A Tabela 3 apresenta os resultados das medidas de qualidade de ajuste, que mostram que o modelo espacial binomial negativo inflacionado de zeros apresentou uma performance significativamente maior do que o modelo linear generalizado binomial negativo, pois todas as suas medidas foram menores que o modelo binomial negativo, reforçando a ideia de que vale o custo de ser estimado ao lidar com dados desta natureza.

Tabela 3 – Medidas de Qualidade de Ajuste.

Modelo	DIC	MAPE
MBNIZ CAR	2.554	977
Binomial Negativo usual	3.324	6.678

Conclusão

O modelo Binomial Negativo adaptado a inflação de zeros com estrutura espacial apresentou resultados melhores na qualidade do ajuste do que o modelo binomial negativo tradicional. O próximo passo deste estudo é utilizar dados reais inflados de zeros com dependência espacial.

Referências

AGRESTI, A. **Foundations of linear and generalized linear models**. 1.ed. Estados Unidos: John Wiley & Sons, 2015.

BÜRKNER, P.C.. **brms**: An R Package for Bayesian Multilevel Models Using Stan. Áustria: Journal of Statistical Software, 2017.

LAMBERT, D. **Zero-inflated poisson regression, with an application to defects in manufacturing**. Estados Unidos: Technometrics, Taylor Francis, v. 34, n. 1, p. 1–14, 1992.

R Core Team. **R:** A Language and Environment for Statistical Computing**.** R Foundation for Statistical Computing, Vienna, Austria. 2023.

SIMPSON, D. P., RUE, H., MARTINS, T. G., RIEBLER, A., SØRBYE, S. H.. **Penalising Model Component Complexity**: A Principled, Practical Approach to Constructing Priors. Statistical Science, Estados Unidos: Institute of Mathematical Statistics, v. 32, n. 1, p. 1 – 28, 2017.