SMNE UFPR
Simpósio de Métodos Numéricos em Engenharia
25 a 27 de outubro, 2017

Estruturas de dependência definidas no espaço de covariáveis

Fernando Gomes Moro
Programa de Pós-Graduação em Métodos Numéricos (PPGMNE)
Universidade Federal do Paraná (UFPR)
Curitiba, Brasil

Paulo Justiniano Ribeiro Jr.
Departamento de Estatística
Universidade Federal do Paraná (UFPR)
Curitiba, Brasil

***Resumo*—Modelos espaciais modelam a dependência a partir de estruturas de distância entre observações espacialmente indexadas. Tal procedimento segue a intuição de que observações geograficamente mais próximas tendem a ser mais parecidas entre si. Por outro lado, informações auxiliares, na forma de covariáveis podem ser incluídas na modelagem, tipicamente na estrutura de médias do modelo como em problemas usuais de regressão. Neste trabalho explora-se a ideia de que a proximidade pode ser definida não apenas no espaço geográfico mas também em um espaço definido por covariáveis. Diversas estratégias para definir tal espaço são apresentadas e avaliadas. São mostrados e comparados resultados obtidos com estruturas espaciais e estruturas definidas sobre as covariáveis individualmente e combinações entre as mesmas. A proposta expande a possibilidade de definição de modelos utilizados na modelagem de dados espaciais e não espaciais.**

***Palavras-chave*—estrutura de dependência; inferência Bayesiana; espaço de covariáveis**

I. Introdução

Estatística espacial é um conjunto de técnicas estatísticas que descrevem ou resumem a dispersão de variáveis aleatórias no espaço, e está ramificada em três subáreas: processos pontuais, geoestatística e dados de área.

Segundo [3] dados de área é quando o espaço geográfico é fragmentado em n subpartições ou áreas, por exemplo, a divisão político administrativa do município de Curitiba em seus 75 bairros. As coordenadas geográficas das observações individuais das variáveis não são conhecidas, mas sim observações associadas aos totais dos domínios das áreas do espaço geográfico.

A dependência espacial em dados por área é representada por uma matriz de vizinhança, cujas áreas do espaço geográfico são indexadas nas linhas e colunas da mesma. A matriz é preenchida por linha, de forma a receber 1 na casela ij caso a área da coluna j é vizinha da correspondente área i e 0 caso contrário. Informações adicionais podem ser encontradas em [2].

Segundo [5], a forma mais conhecida de vizinhança é a vizinhança por adjacência de áreas. Esta é definida como sendo vizinhas as áreas que compartilham a mesma divisão geográfica.

Existem algumas alternativas propostas na literatura, em [5], é proposta uma extensão da vizinhança espacial de primeira ordem, na qual, a construção da matriz de vizinhança do modelo depende de um parâmetro desconhecido, responsável por controlar o alcance da dependência espacial. Em [1], é proposta uma vizinhança a partir de um *cluster* de variáveis com restrição espacial.

A proposta do trabalho é construir estruturas de vizinhanças segundo a similaridade no espaço de variáveis que exprimam fatores relacionados à variável sob estudo. Tais estruturas

podem ser definidas de várias formas, das quais três foram testadas:

- Algoritmos de agrupamento.
- Função de covariância geoestatística.
- Partição do espaço de covariáveis por tecelagem de Voronoi.

As estruturas propostas foram comparadas com a estrutura de vizinhança por adjacência e a não estruturada (iid), através de informações de qualidade de ajuste dos modelos.

II. Materiais e Métodos

O estudo utilizou dados das contagens de óbitos infantis nos municípios do Paraná acumuladas do período de 2008 a 2012. A base de contagem ou população sob risco em cada área foi dada pelo número de nascidos vivos no mesmo período de referência definindo assim o termo *offset* dos modelos. Além de dez variáveis de condições socioeconômicas.

Os ajustes foram feitos por meio de um modelo Bayesiano hierárquico com resposta Poisson. E através dos critérios de qualidade de ajuste foram encontrados os melhores modelos tanto em ajuste global quanto preditivo.

O modelo CAR via inferência Bayesiana se enquadra em uma classe de modelos, conhecida como modelos hierárquicos Bayesianos. O termo hierárquico se deve ao fato da hierarquia das distribuições a priori, efeitos latentes e verossimilhança.

Por suposição, as observações da variável resposta são independentes condicionadas aos parâmetros e as variáveis latentes. Desta forma, a verossimilhança fica expressa por um produtório da distribuição avaliada na resposta observada condicionada aos demais argumentos. O segundo nível da hierarquia é o efeito aleatório, e no terceiro nível da hierarquia, as distribuições a prioris.

Assim, o modelo hierárquico Bayesiano com resposta Poisson, pode ser escrito genericamente como:

$$Y_i|\boldsymbol{\theta} \sim Poisson(\lambda_i E_i) \tag{1}$$

$$\log(\lambda_i) = \sum_{j=1}^{k} \mathbf{X}_{ij}\beta_j + \phi_i \tag{2}$$

$$\boldsymbol{\phi} \sim N(\boldsymbol{\psi}, \mathbf{Q}_\tau^{-1}) \tag{3}$$

$$\beta_j \sim N(0, 1000) \tag{4}$$

$$\tau \sim Gamma(1, 0.01) \tag{5}$$

No presente estudo, à resposta foi atribuída distribuição de Poisson, com o *offset* representado por E_i. O que difere os modelos ajustados é a presença ou não das covariáveis na estrutura de média $\sum_{j=1}^{k} \mathbf{X}_{ij}\beta_j$ na equação 2, e a forma como é preenchida a matriz de precisão $\mathbf{Q}_\tau$ na equação 3, no caso, as estruturas de dependência.

Foram ajustados modelos para cada estrutura de dependência com ou sem covariável no preditor linear, e os mesmos comparados através de critérios de qualidade de ajustes globais e preditivos.

O DIC(*Deviance Information Criterion*) e o WAIC(*Watanabe-Akaike Information Criterion*) são critérios de informação de qualidade de ajuste globais, quando menor o valor melhor o ajuste. Já o CPO(*Conditional preditive ordinates*) e o PIT(*Probability Inverse Transform*) são critérios de informação de qualidade de ajuste preditivas, o valor negativo do logaritmo do CPO quanto menor melhor, e o PIT varia de 0 a 1, acima de $0,90$ significa bom ajuste preditivo.

As análises foram realizadas com auxílio do *software* R [4], e com o pacote R-INLA [6], que implementa a metodologia INLA (*Integrated Nested Laplace Aproximation*) no R.

III. Resultados e Discussão

Tabela I: Medidas de qualidade de ajuste dos modelos ajustados, ordenadas segundo o critério DIC, com a inclusão de um efeito aleatório e de um modelo ajustado sem.

Estrutura	ICMN	CPO	PIT	WAIC	DIC
Matriz Cofenética	sim	1061.70	0.96	2081.70	2082.60
Voronoi3	sim	1063.00	0.95	2090.50	2085.60
Não estruturado(iid)	sim	1063.30	0.96	2085.20	2085.70
Exponencial(CP)	sim	1063.30	0.96	2085.20	2085.70
Matriz Cofenética	não	1065.50	0.95	2087.10	2088.30
Não estruturado(iid)	não	1068.90	0.95	2094.90	2094.40
Exponencial(CP)	não	1068.90	0.95	2094.80	2094.40
Espacial	sim	1067.60	0.94	2104.10	2095.00
Voronoi3	não	1070.40	0.95	2103.40	2095.80
Voronoi2	sim	1074.40	0.94	2105.60	2097.30
Espacial	não	1071.20	0.94	2106.50	2098.20
Voronoi1	sim	1073.70	0.94	2108.70	2099.00
Voronoi1	não	1076.70	0.94	2115.30	2104.80
Voronoi2	não	1079.60	0.94	2115.20	2106.40
K-means	sim	1081.80	0.94	2152.10	2127.80
CLARA	sim	1084.20	0.92	2148.40	2129.40
K-means	não	1088.60	0.92	2165.10	2142.90
CLARA	não	1097.10	0.92	2179.50	2157.70
Ag. hierárquico	sim	1109.50	0.91	2203.30	2170.30
Ag. hierárquico	não	1109.70	0.91	2203.80	2175.90
SKATER	sim	1127.40	0.90	2248.80	2223.50
GLM	sim	1127.40	0.90	2248.80	2223.50
SKATER	não	1170.90	0.90	2340.70	2320.40

[1]FONTE: Elaborado pelos autores.

NOTA: ICMN = inclusão de covariáveis na média.

Na tabela I, encontram-se as informações resumo referentes aos modelos ajustados com apenas um efeito aleatório e de um modelo sem efeito aleatório. Nota-se que os modelos

que produziram melhores ajustes, em geral nos critérios de informações preditivas quanto em ajuste global, foram aqueles em que as estruturas de dependência eram definidas pela similaridade no conjunto de variáveis. Isto implica que estes se ajustam melhor que os modelos cuja estrutura de dependência era definida pela vizinhança espacial. Entre os modelos das estruturas propostas, os melhores foram aqueles com estrutura de covariância geoestatística e estrutura de adjacência nos polígonos da tecelagem de Voronoi de variáveis. Já os modelos com vizinhança definida pelos algorítmicos de agrupamento tiveram os piores resultados, tanto em critério de informação global quanto preditivo, exceto para o modelo GLM sem a inclusão de efeitos aleatórios.

Tabela II: Medidas de qualidade de ajuste dos modelos ajustados, ordenadas segundo o critério DIC, com combinação de um efeito aleatório estruturado e outro não estruturado (iid).

Estrutura	ICNM	CPO	PIT	WAIC	DIC
K-means	não	1059.20	0.95	2076.50	2077.90
K-means	sim	1059.40	0.96	2077.00	2078.90
CLARA+iid	sim	1061.10	0.96	2077.40	2079.90
Exponencial(CP)	sim	1060.10	0.95	2079.70	2081.00
Voronoi3	sim	1060.70	0.96	2081.80	2082.60
Matriz Cofenética	sim	1061.90	0.96	2081.30	2083.40
Ag. hierárquico	sim	1062.80	0.96	2085.00	2085.40
SKATER	sim	1063.10	0.96	2085.00	2085.60
Espacial	sim	1062.90	0.95	2088.00	2086.90
1CLARA	não	1065.40	0.96	2085.00	2087.00
1Voronoi1	sim	1064.50	0.96	2086.40	2087.00
1Voronoi2	sim	1064.70	0.96	2085.80	2087.00
1Exponencial(CP)	não	1065.30	0.95	2086.90	2087.90
1Matriz Cofenética	não	1065.80	0.95	2087.60	2088.80
1Ag. hierárquico	não	1066.20	0.95	2088.40	2089.10
1Voronoi3	não	1066.20	0.95	2089.60	2089.60
1Espacial	não	1066.70	0.95	2092.00	2091.40
1Voronoi1	não	1067.80	0.95	2091.90	2092.20
1SKATER	não	1068.90	0.95	2094.60	2094.30
Voronoi2	não	1069.90	0.95	2094.60	2095.00

[1]FONTE: Elaborado pelos autores.

NOTA: ICMN = inclusão de covariáveis na média.

Pela tabela II, nota-se que, em geral a adição do efeito aleatório iid trouxe ganhos para o modelo em relação aqueles com apenas um efeito aleatório estruturado (I), tanto em ajuste global quanto em qualidade preditiva. Sendo que, dentre os modelos com efeito aleatório estruturado adicionado do efeito aleatório iid, destacam-se aqueles cuja estrutura de dependência é definida pelos agrupamentos de variáveis (SKATER, *k-means*, CLARA, Ag. Hierárquico).

O efeito aleatório iid nos modelos da tabela II, absorve a maior parte da variabilidade captada pelos efeitos aleatórios, na maioria dos casos. Isto também pode ser verificado nas proporções de cada componente aleatório da variabilidade total explicada por estes, conforme tabela III. Também, a inclusão das covariáveis na estrutura de média, faz com que a variabilidade total explicada pelos dois componentes aleatórios de cada modelo diminua.

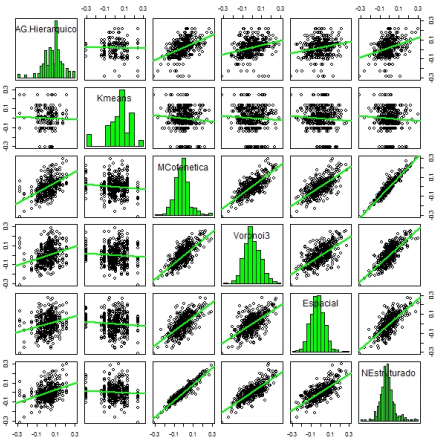
Tabela III: Proporção da variabilidade explicada por cada efeito aleatório e a variância total dos modelos com combinação de um efeito aleatório estruturado e outro não estruturado (iid).

Estrutura	ICMN	Estruturado	IID	Variância total
K-means	não	0.383	0.617	0.038
K-means	sim	0.337	0.663	0.034
CLARA	sim	0.267	0.733	0.034
Exponencial(CP)	sim	0.421	0.579	0.035
Voronoi3	sim	0.487	0.513	0.033
Hierárquico	sim	0.573	0.427	0.030
MCofenetica	sim	0.225	0.775	0.031
SKATER	sim	0.689	0.311	0.084
Espacial	sim	0.334	0.666	0.030
CLARA	não	0.236	0.764	0.036
Voronoi1	sim	0.256	0.744	0.031
Voronoi2	sim	0.260	0.740	0.031
Exponencial(CP)	não	0.376	0.624	0.042
Hierárquico	não	0.592	0.408	0.037
MCofenetica	não	0.251	0.749	0.036
Voronoi3	não	0.371	0.629	0.036
Espacial	não	0.332	0.668	0.035
Voronoi1	não	0.267	0.733	0.034
SKATER	não	0.670	0.330	0.092
Voronoi2	não	0.234	0.766	0.034

[1]FONTE: Elaborado pelos autores.

NOTA: ICMN = inclusão de covariáveis na média.

Figura 1: Dispersão entre os efeitos aleatórios dos modelos com estruturas de dependências definidas pelas covariáveis e estrutura de média constante.

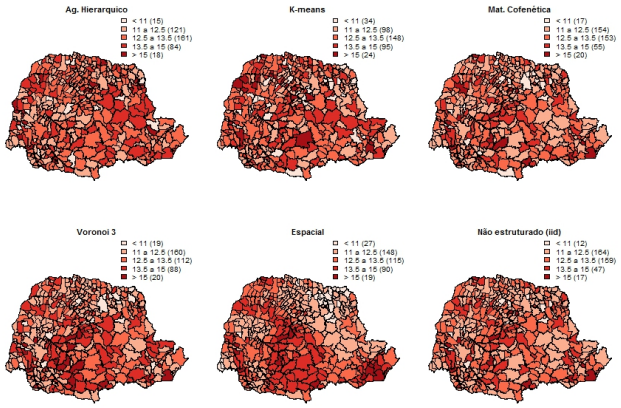


[1]Fonte: Elaboradas pelos autores.

A fim de explorar a relação entre as variáveis latentes dos modelos, foram selecionados seis ao total, sem covariáveis na média e com um efeito aleatório, em que cada um destes modelos possui uma estrutura de covariância preenchida de uma forma em particular. No gráfico da figura 1, observa-se que os valores dos efeitos aleatórios dos modelos definidos com uma estrutura de covariância não iid, isto é, com valores não nulos não apenas nas diagonais, apresentaram correlação com os valores do efeito aleatório com estrutura iid e individual

para cada observação. Já no caso dos modelos cujas estruturas de covariâncias eram iid, e valores dos efeitos aleatórios para cada grupo definido pela similaridade das variáveis, mostra-se não haver uma correlação forte com os valores de efeitos aleatórios dos outros modelos.

Figura 2: Mapas dos valores preditos pelos modelos com um efeito aleatório e sem covariáveis na estrutura de média.



[1]Fonte: Elaboradas pelos autores.

Pode-se ainda verificar pelos gráficos da figura 2, que os modelos cujas estruturas foram definidas pela similaridade do conjunto de variáveis socioeconômicas fornecem um padrão dos valores preditos que não necessariamente é espacial, de certa forma isso pode representar uma maior flexibilidade em relação a estrutura de vizinhança espacial. Também é possível notar que os valores preditos dos modelos cujas estruturas são definidas através dos agrupamentos das variáveis socioeconômicas, são os que apresentaram padrões menos espacializados em relação aos demais métodos de definição de similaridade no espaço de variáveis.

IV. Conclusões

Dentre os modelos com uma variável latente, obteve-se modelos com estruturas definidas pela similaridade no espaço de variáveis, com melhores ajustes que os modelos com estrutura de vizinhança espacial. Especificamente, nestes modelos de melhores ajustes, as similaridades no espaço de covariáveis eram definidas de forma que tivesse conexão entre todas as áreas do espaço geográfico, porém, ao combinar os efeitos aleatórios estruturados com o efeito aleatório não estruturado (iid) com valores estimados para cada observação, os modelos com combinações entre o efeito aleatório não estruturado e um efeito aleatório definido pelo agrupamento das variáveis produziram melhores resultados, tanto em ajuste global quanto qualidade preditiva.

Também, as variáveis latentes com estruturas definidas pelo agrupamento de variáveis, apresentaram um padrão menos espacializado do que aquelas com estruturas definidas pela função exponencial geoestatística ou vizinhança espacial no polígono da tecelagem de Voronoi. Porém, os modelos com estruturas de dependência definidas pela função exponencial geoestatística ou vizinhança espacial no polígono da tecelagem de Voronoi, apresentaram em geral, melhores resultados comparados com aqueles com estrutura iid, devido ao fato de os efeitos latentes com estas estruturas induzirem a um padrão específico, sendo mais informativos.

Pretende-se futuramente, estudar melhores formas para escolha do do número de grupos dos algoritmos de agrupamentos, entre eles, a inclusão do mesmo como parâmetro a ser estimado pelo modelo.

Referências

[1] R.M. Assunção, J.P. Laje, and E.A Reis. Análise de conglomerados espaciais via árvore geradora mínima. *Revista Brasileira de Estatística*, 62:1–23, 2002.

[2] T.C. Bailey and A.C. Gatrell. *Interactive Spatial Data Analysis*. Harlow:Logman, 1996.

[3] R.X. Cortes. Estimando modelos dinâmicos utilizando o inla para campos aleatórios markovianos não gaussianos. Master's thesis, UFMG, Janeiro 2014.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[5] E.C. Rodrigues and R.M. Assunção. Bayesian spatial models with mixture neighbourhood structure. *Revista da Estatística da UFOP*, 1:92–108, 2011.

[6] H. Rue, S. Martinho, and N. Chopin. Approximate bayesian inference for latent gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society*, 71(2):319–392, 2009.