ISSN: 2317-0840

Utilização de matrizes de vizinhança socioeconômicas em modelos da classe STARMA aplicados a dados epidemiológicos

Matheus F. Freitas[1†], Haiany Aparecida Ferreira[2], Thelma Sáfadi[3], Daniella F. Freitas[4]

[1]*Departamento de estatística- UFLA.*
[2]*Departamento de estatística- UFLA. email: haianyferreira@yahoo.com.br.*
[3]*Departamento de estatística- UFLA.. email: safadi@ufla.br.*
[4]*Departamento de saúde- UFLA. email: daniella.feres@hotmail.com.*

***Resumo**: Neste trabalho estudou-se a utilização de matrizes de vizinhança socioeconômicas em modelos espaço temporais da classe auto regressivo e de médias móveis (STARMA). O conjunto de dados escolhido é composto por nove séries temporais que quantificam a taxa de incidência de Tuberculose, observadas entre 2002 e 2017, nas seguintes cidades mineiras: Belo Horizonte, Betim, Contagem, Governador Valadares, Juiz de Fora, Lavras, Montes Claros, Pouso Alegre e Uberlândia. Uma vez que a maior parte das cidades encontram-se geograficamente distantes, foi necessária a utilização de matrizes de vizinhança socioeconômicas. As matrizes foram obtidas por meio de duas variáveis socioeconômicas: o IDH municipal e o investimento anual médio na saúde básica. Foram ajustados modelos da classe STARMA considerando-se o conjunto de dados e as duas matrizes de vizinhança obtidas. A obtenção do modelo foi feita computacionalmente e consistiu de três etapas: Identificação, estimação e diagnóstico do modelo. Concluiu-se que, as matrizes de vizinhança socioeconômicas em modelos STARMA aplicados ao conjunto de dados escolhido, foi apropriada uma vez que estas matrizes podem ser utilizadas em séries espaço-temporais nas quais os locais de interesse encontram-se geograficamente distantes.*

***Palavras-chave**: Matriz de vizinhança socioeconômica; STARMA; Tuberculose.*

***Abstract**: In this work the use of socioeconomic neighborhood matrices was studied in space-time models of the autoregressive and moving averages class (STARMA). The selected data set is composed of nine time series that quantify the incidence rate of Tuberculosis observed between 2002 and 2017 in the following cities: Belo Horizonte, Betim, Contagem, Governador Valadares, Juiz de Fora, Lavras, Montes Claros, Pouso Alegre and Uberlândia. Since most cities are geographically distant, the use of socioeconomic neighborhood matrices was necessary. The matrices were obtained through two socioeconomic variables: the municipal IDH and the average annual investment in basic health. STARMA class models were adjusted considering the data set and the two neighborhood matrices obtained. The model was obtained computationally and consisted of three stages: Identification, estimation and diagnosis of the model. It was concluded that the socioeconomic neighborhood matrices in STARMA models applied to the data set chosen were appropriate since these matrices can be used in space-time series in which the places of interest are geographically distant.*

***Keywords**: Socioeconomic neighborhood matrix; STARMA; Tuberculosis.*

† Autor correspondente: matheus712@hotmail.com.

Introdução

Os modelos espaço-temporais se caracterizam por ajustar conjuntos de dados que apresentem além da correlação temporal a correlação espacial. Entende-se que um conjunto de dados apresenta correlação quando existe dependência entre seus elementos. De maneira que, um conjunto de dados apresenta correlação temporal quando os elementos deste conjunto de dados se distribuírem ao longo do tempo de maneira dependente. De forma análoga, dizer que um conjunto de dados apresenta correlação espacial significa dizer que existe entre os elementos deste conjunto de dados um grau de associação espacial, de forma que suas disposições espaciais e suas interações com seus vizinhos devem ser consideradas durante o ajuste do modelo (PFEIFER; DEUTRCH, 1980; JIN, 2017).

Considere que determinada variável de interesse será monitorada simultaneamente em N diferentes locais durante T instantes de tempo, de forma que para todo instante $t \in T$, tem-se N observações desta mesma variável. Entende-se que este conjunto de dados apresentará correlação espaço-temporal se, em um dado instante de tempo tomar-se aleatoriamente uma dentre as N observações, e tal observação relacionar-se simultaneamente com observações obtidas em sua vizinhança e com as observações obtidas anteriormente e/ou posteriormente naquele local.

A relação existente entre os locais de interesse é explicitada por meio dos pesos espaciais $w_{i,j}$, contidos na matriz de ponderação espacial (**W**). Esta matriz, que busca indicar o grau de similaridade entre as regiões de interesse, pode ser construída de diferentes maneiras. Para os modelos espaço-temporais da classe auto regressiva e de médias móveis, classe de interesse deste trabalho, a obtenção de uma matriz **W** apropriada é de extrema importância e impactará diretamente no processo de ajuste do modelo (ALMEIDA, 2012).

Usualmente as matrizes de vizinhança são construídas em função de critérios geográficos como, por exemplo, a existência ou inexistência de fronteiras, o tamanho de fronteira compartilhada ou ainda em distâncias observadas entre duas áreas da região de interesse. Embora seja menos comum, é possível também obter matrizes **W** baseando-se em critérios socioeconômicos.

Matrizes de vizinhança socioeconômicas podem ser construídas baseando-se no grau de similaridade, entre as subáreas da região de interesse, dada uma mesma variável socioeconômica mensurada nestas áreas. Este conceito fundamenta-se no pensamento de que áreas com características iguais têm maior proximidade, ainda que geograficamente distantes (ALMEIDA, 2012). De maneira que a disposição e as iterações espaciais podem ser melhores representadas por meio de critérios socioeconômicos que a partir de critérios geográficos. Na Equação 1, tem-se a equação de um modelos espaço-temporal da autorregressiva e de medias móveis (STARMA).

$$\boldsymbol{Z}(t) = \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k} \emptyset_{kl}\boldsymbol{WZ}(t-k) - \sum_{k=1}^{q}\sum_{l=0}^{m_k} \theta_{kl}\boldsymbol{W\epsilon}(t-k) + \boldsymbol{\epsilon}(t), \quad (1)$$

sendo que **Z**(t) representa uma observação no local i realizada no instante t; p a ordem auto regressiva e q a ordem de médias móveis do modelo, λ_k a ordem máxima de vizinhança do k-ésimo termo auto regressivo e m_k a ordem máxima espacial do k-ésimo termo de médias móveis, $\emptyset_{kl}$ e θ_{kl} são parâmetros do modelo, **W** a matriz de vizinhança e e $\epsilon_i(t)$ erros normais e aleatórios associado ao modelo, com $E[\epsilon_i(t)] = 0$ e $E[\epsilon_i(t)\epsilon_j(t+s)'] = \begin{cases} \sigma^2, i = j \; e \; s = 0 \\ 0, caso\ contrário \end{cases}$.

Materiais e métodos

Os dados foram obtidos do site do Departamento de Informática do Sistema Único de Saúde (DATASUS), e quantificam casos confirmados de tuberculose em nove cidades mineiras, sendo elas: Belo Horizonte, Betim, Contagem, Divinópolis, Juiz de Fora, Montes Claros, Pouso Alegre, Uberlândia e Lavras. O conjunto de dados trata-se de observações mensais coletadas entre janeiro de 2002 e dezembro de 2017.

Foram construídas duas matrizes de vizinhança socioeconômicas(**W**). Na primeira matriz de vizinhança considerou-se a média de investimento anual em saúde básica das nove cidades de interesse. Na segunda matriz de ponderação socioeconômica considerou-se o último IDH das cidades mineiras, uma vez que se sabe que a tuberculose é fortemente relacionada a questões econômicas, sociais e ambientais (GUIMARÃES et al., 2012). Para a obtenção dos pesos espaciais de tais matrizes de vizinhança utilizaram-se as Equações 2 e 3.

$$w_{i,j} = {}^{1}\Big/_{\left|IDH_i - IDH_j\right|} \qquad (2)$$

$$w_{i,j} = {}^{1}\Big/_{\left|In_i - In_j\right|} \qquad (3)$$

em que IDH_i é o IDH da i-ésima cidade; IDH_j é o IDH da j-ésima cidade; In_i é o investimento médio da i-ésima cidade em saúde básica e In_j é o investimento médio da j-ésima cidade em saúde básica.

Para as duas matrizes de vizinhança construídas foi ajustado um modelo da classe STARMA. O ajuste dos modelos foi feito através do software estatístico livre R (R CORE TEAM, 2018), utilizando-se um pacote denominado ‘starma’ de autoria de Felix Cheysson (CHEYSSON, 2016).

O pacote ‘starma’ fornece todas as ferramentas necessárias para identificar, estimar e diagnosticar modelos da classe STARMA para séries espaço-temporais. O pacote utiliza o procedimento de construção do modelo iterativo de três estágios desenvolvido por Box e Jenkins (1970) e estendido para a modelagem espaço-temporal proposta por Pfeifer e Deutsch (1980). (CHEYSSON, 2016).

É importante salientar que o pacote ‘starma’, assim como no ajuste de séries temporais, se aplica somente a dados estacionários. Por este motivo todas as funções que serão apresentados em seguida, foram aplicadas após a eliminação da tendência das séries temporais que compõem a série espaço-temporal de interesse.

Resultados e discussões

Para a obtenção das matrizes de vizinhança foram consideradas duas variáveis socioeconômicas, o IDH municipal e o investimento direto na saúde básica municipal. Por meio das Equações 2 e 3, foram obtidas as matrizes exibidas a seguir:

$$Inv_{norm.} = \begin{bmatrix} 0{,}0000 & 0{,}1232 & 0{,}1324 & 0{,}1231 & 0{,}1278 & 0{,}1074 & 0{,}1409 & 0{,}1089 & 0{,}1361 \\ 0{,}0005 & 0{,}0000 & 0{,}0080 & 0{,}9576 & 0{,}0154 & 0{,}0038 & 0{,}0044 & 0{,}0042 & 0{,}0059 \\ 0{,}1807 & 0{,}0938 & 0{,}0000 & 0{,}0936 & 0{,}1973 & 0{,}0302 & 0{,}1168 & 0{,}0326 & 0{,}2548 \\ 0{,}0547 & 0{,}2425 & 0{,}2193 & 0{,}0000 & 0{,}2190 & 0{,}0550 & 0{,}0639 & 0{,}0614 & 0{,}0844 \\ 0{,}0099 & 0{,}0649 & 0{,}2875 & 0{,}2599 & 0{,}0000 & 0{,}0521 & 0{,}1069 & 0{,}0568 & 0{,}1620 \\ 0{,}0088 & 0{,}0809 & 0{,}0468 & 0{,}0692 & 0{,}0554 & 0{,}0000 & 0{,}0371 & 0{,}6598 & 0{,}0419 \\ 0{,}0116 & 0{,}0770 & 0{,}1814 & 0{,}0804 & 0{,}1135 & 0{,}0370 & 0{,}0000 & 0{,}0451 & 0{,}3818 \\ 0{,}0077 & 0{,}0915 & 0{,}0433 & 0{,}0663 & 0{,}0518 & 0{,}6671 & 0{,}0338 & 0{,}0000 & 0{,}0384 \\ 0{,}0092 & 0{,}0881 & 0{,}3258 & 0{,}0876 & 0{,}1421 & 0{,}0346 & 0{,}2755 & 0{,}0363 & 0{,}0000 \end{bmatrix}$$

$$IDH_{norm.} = \begin{bmatrix} 0{,}0000 & 0{,}0765 & 0{,}0864 & 0{,}0562 & 0{,}1458 & 0{,}1666 & 0{,}1166 & 0{,}1296 & 0{,}2222 \\ 0{,}0429 & 0{,}0000 & 0{,}3739 & 0{,}1189 & 0{,}0902 & 0{,}0793 & 0{,}1246 & 0{,}1047 & 0{,}0654 \\ 0{,}0424 & 0{,}3268 & 0{,}0000 & 0{,}0789 & 0{,}1040 & 0{,}0880 & 0{,}1634 & 0{,}1271 & 0{,}0693 \\ 0{,}0633 & 0{,}2387 & 0{,}1811 & 0{,}0000 & 0{,}1030 & 0{,}0955 & 0{,}1221 & 0{,}1117 & 0{,}0847 \\ 0{,}0369 & 0{,}0407 & 0{,}0537 & 0{,}0231 & 0{,}0000 & 0{,}2953 & 0{,}1476 & 0{,}2953 & 0{,}1074 \\ 0{,}0493 & 0{,}0419 & 0{,}0531 & 0{,}0251 & 0{,}3454 & 0{,}0000 & 0{,}1151 & 0{,}1727 & 0{,}1973 \\ 0{,}0369 & 0{,}0702 & 0{,}1053 & 0{,}0343 & 0{,}1843 & 0{,}1229 & 0{,}0000 & 0{,}3686 & 0{,}0776 \\ 0{,}0332 & 0{,}0478 & 0{,}0664 & 0{,}0375 & 0{,}3613 & 0{,}1807 & 0{,}3613 & 0{,}0000 & 0{,}0964 \\ 0{,}1009 & 0{,}0529 & 0{,}0642 & 0{,}0342 & 0{,}1925 & 0{,}3026 & 0{,}1115 & 0{,}1412 & 0{,}0000 \end{bmatrix}$$

Na Figura 1 são exibidos os gráficos da função de auto correlação espaço-temporal (stacf) e função de auto correlação parcial espaço-temporal (stpacf). Por meio da análise do gráfico da stacf, para o caso de defasagem no especial nula (slag = 0), é possível constatar a existência de auto correlação nas series temporais. Uma vez que a stacf assume valores significativos (valores que excedem o intervalo delimitado em azul).

De maneira análoga, é possível oberservar a existência de auto correlação espaço-temporal no conjunto de dados, considerando-se a primeira defasagem especial (slag = 1). Isto evidencia a pertinência da utilização de modelos da classe STARMA.

O decaimento exponencial observado no gráfico da stpacf (slag = 0) indica que modelos espaço-temporais de medias móveis (STMA) podem ser apropriado (JIN, 2017). Por este motivo ajustou-se modelos STMA (5_1).

Uma vez que os modelos da classe STMA não foram capazes de ajustar completamente a estrutura de auto correlação espaço-temporal presente nos dados, e verificando-se que o decaimento exponencial não é observado no gráfico stpacf (slag = 1), optou-se por ajustar modelos da classe STARMA. Foram ajustados modelos da STARMA(6_1,6_1), novamente considerando-se as duas matrizes de vizinhança obtidas. Os resultados do ajuste do modelo STARMA são exibidos nas Tabelas 1 e 2.

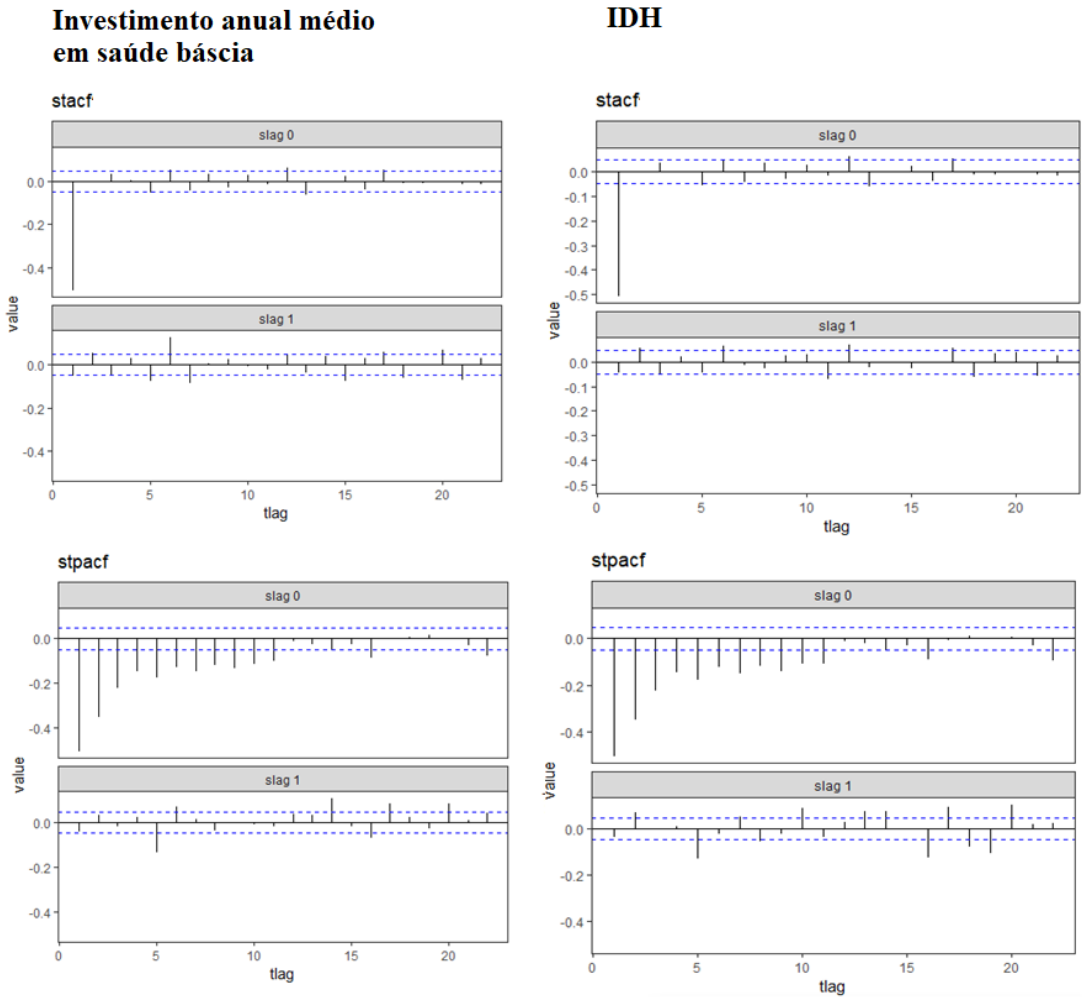

Figura1- Gráfico das função de auto correlação e auto correlação espaço-temporal, considerando-se ambas matrizes de vizinhança obtidas.

Tabela 1- Resultado do ajuste para o modelo STARMA $(6_1, 6_1)$ considerando-se **W** baseada no investimento anual médio em saúde básica.

PARÂMETRO	ESTIMATIVA (ERRO PADRÃO)	p-valor
$\emptyset_{5,0}$	-0,068844 (0,018883)	0,000275
$\emptyset_{6,1}$	0,241055 (0,041939)	$1{,}07\text{x}10^{-8}$
$\theta_{1,0}$	-0,830716 (0,025199)	$<2{,}20\text{x}10^{-16}$
$\theta_{6,1}$	-0,119665 (0,055938)	0,032558

Tabela 2- Resultado do ajuste para o modelo STARMA ($\mathbf{6_1}, \mathbf{1_1}$) considerando-se **W** baseada no IDH municipal.

PARÂMETRO	ESTIMATIVA (ERRO PADRÃO)	p-valor
$\emptyset_{5,0}$	-0,067049 (0,018669)	0,000338
$\emptyset_{2,1}$	0,172787 (0,036590)	$2{,}52\text{x}10^{-6}$
$\emptyset_{6,1}$	0,132327 (0,036384)	0,000284
$\theta_{1,0}$	-0,893149 (0,025600)	$<2{,}20\text{x}10^{-16}$

Os modelos da classe STARMA apresentaram resultados ligeiramente superiores aos modelos STAR (5_1), no que se refere a auto correlação espaço-temporal dos resíduos. No entanto, mesmo após o ajuste dos modelos STARMA foi observado a existência de correlação espaço-temporal. Este fato verifica-se pelos gráficos da função de auto correlação espaço-temporal residual, apresentados na Figura 2.

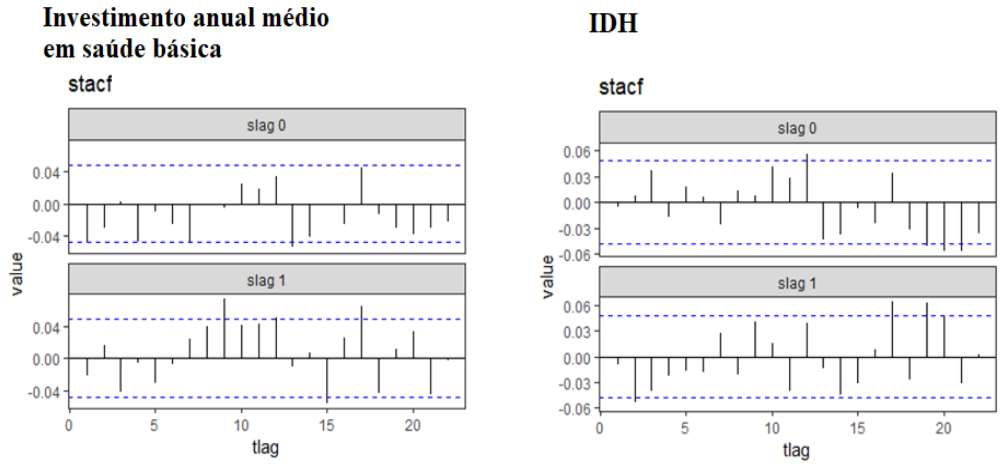

Figura2- Gráfico da função de auto correlação espaço-temporal após ajuste dos modelos.

Foram ajustados outros modelos, uns mais parcimoniosos tais como STARMA (2_1,2_1) e STMA(5_1) e outros menos parciomoniosos tais como STARMA (10_1,10_1). No entanto, nenhum dos modelos foi capaz de ajustar completamente a estrutura de correlação espaço-temporal existente no conjunto de dados.

Considerações finais

Uma vez que nenhum dos modelos ajustados foram capazes de ajustar completamente a estrutura de correlação espaço-temporal presente no conjunto de dados, é pertinente a proposição de matrizes de vizinhança diferentes.

Não é possível a utilização de matrizes de vizinhança baseada em critérios geográficos. No entanto é interessante pesquisar a existência de outras variáveis socioeconômicas que apresentam relação com a taxa de incidência de tuberculose.

Também é pertinente a proposição de outras maneiras de ponderação dos pesos espaciais. Considerando novas variáveis socioeconômicas ou as mesmas já utilizadas no trabalho.

Agradecimentos

Agradeço o apoio financeiro da Fapemig e da Capes.

Referências Bibliográficas

ALMEIDA, E. Econometria espacial aplicada. Campinas–SP. Alínea, 2012.

CHEYSSON F. (2016). starma: Modelling Space Time Auto Regressive Moving Average (STARM A). Processes. R package version 1.3. https://CRAN.R-project.org/package=starma

GUIMARÃES, R. M. et al. Tuberculose, HIV e pobreza: tendência temporal no Brasil, Américas e mundo. Jornal Brasileiro de Pneumologia, v. 38, n. 4, p. 518-525, 2012.

JIN, E. Y. Estrutura de vizinhanças espaciais nos modelos autorregressivos e de médias móveis espaço-temporais STARMA. Dissertação de Mestrado. Universidade de São Paulo, 2017.

PFEIFER, P. E.; DEUTRCH, S. J. A three-stage iterative procedure for space-time modeling phillip. Technometrics, v. 22, n. 1, p. 35-47, 1980.

R CORE TEAM (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.