Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

Trabalho apresentado no XL CNMAC, Evento Virtual - Co-organizado pela Universidade do Mato Grosso do Sul (UFMS).

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Correlação cruzada em séries temporais de casos de dengue - em busca do atraso que maximiza a correlação

Cátia S. N. Sepetauskas[1]
INPE, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP
Lívia R. Tomás[2]
Cemaden, São José dos Campos, SP
Davi Sanches[3]
Cemaden/UNIFESP, São José dos Campos, SP
Luciana R. Londe[4]
Cemaden, São José dos Campos, SP
Leonardo B. L. Santos[5]
Cemaden, São José dos Campos, SP

Resumo. Medidas de correlação entre séries temporais tem sido aplicadas em diversos temas de pesquisa, dentre eles epidemiologia matemática. A correlação cruzada permite identificar o atraso que maximiza a correlação entre séries temporais, por exemplo, de casos de dengue em diferentes regiões de uma cidade. A distribuição desses atrasos é ainda uma questão em aberto, especialmente na escala intra-urbana, assim como sua relação com questões ambientais e sociais. Neste trabalho, confrontamos tais atrasos com dados de mobilidade urbana para a cidade de São José dos Campos/SP. Os resultados mostram que a média para as correlações decresce frente à distância geográfica e cresce com a mobilidade de pessoas entre as regiões, e que o atraso nulo (correlação convencional) é o mais frequente entre os que maximizam a correlação.

Palavras-chave. Epidemiologia Matemática, Correlação cruzada, Mobilidade, Dengue.

1 Introdução

Medidas de correlação entre séries temporais tem sido aplicadas em diversos temas de pesquisa, em ordem de mensurar similaridades entre diferentes padrões [5]. Em particular, a área de epidemiologia matemática tem vários exemplos de pesquisas envolvendo tal abordagem [4, 9]. Dentre as diferentes métricas de correlação está a correlação cruzada, que permite identificar o atraso que maximiza a correlação entre as séries temporais.

Nos últimos anos, doenças como Dengue, Zika e Chikungunha tem causado diversos impactos negativos em muitos municípios brasileiros, com aumento do número de internações e ausências no trabalho, além da dificuldade individual para realização das atividades cotidianas. Preocupações com propagação de epidemias vão desde doenças "recorrentes", a pandemias mundiais, como a ameaça pelo novo Coronavírus, tema de interesse global [6, 9].

[1]souz.kti@gmail.com
[2]liviatomas@gmail.com
[3]davisanches04@gmail.com
[4]luciana.londe@cemaden.gov.br
[5]santoslbl@gmail.com

Cardoso *et al.* (2013) e Saba *et al.* (2018) geraram grafos para representar a rede de transportes no estado da Bahia. Ambos trabalhos mostraram uma grande correlação entre o expoente de criticalidade e o tráfego entre cidades, e concluíram ilustrando a possibilidade de prever e diminuir o número de casos de dengue, agindo preventivamente com base na topologia da rede [3, 9]. Em um trabalho com aplicação em meteorologia, Ceron *et al.* (2019) analisaram como os valores de correlação entre séries temporais de diferentes pares de regiões decaem frente à distância geográfica entre as regiões [4].

A distribuição dos atrasos obtidos via análise da correlação cruzada é ainda uma questão em estudo na literatura, especialmente na escala intra-urbana, assim como sua relação com questões ambientais e sociais. Neste artigo confrontamos tais atrasos com dados de mobilidade urbana para a cidade de São José dos Campos/SP (SJC). Vale ressaltar que a região Sudeste do Brasil é responsável pelo registro de incidência de 46,56 casos de dengue/100 mil habitantes, sendo no estado de São Paulo a incidência de 67,71 casos/100 mil habitantes.

Este artigo está assim organizado: a Seção 2 apresenta informações sobre a área de estudo, a cidade de São José dos Campos/SP, além da descrição dos dados epidemiológicos utilizados nesse trabalho. Em seguida, na Seção 3 são mostrados, por meio de fluxogramas, a estrutura dos códigos desenvolvidos e os produtos de cada processamento. A Seção 4 apresenta os principais resultados da investigação e, finalmente, a Seção 5 sumariza as principais contribuições e apresenta possíveis encaminhamentos.

2 Método

2.1 São José dos Campos

A cidade de São José dos Campos localiza-se na Região do Vale do Paraíba, a uma distância de cerca de 90 km de São Paulo e 320 km do Rio de Janeiro. Possui uma população de 721.944 habitantes (IBGE, 2019) e área territorial de 1.099,4 km^2, dos quais 356 km (33%) são de área urbana e 743,4 km (67%) de área rural [8]. Em relação à mobilidade urbana, os habitantes de SJC realizam em média 2,57 viagens por dia. Esse valor é considerado alto se comparado à média de municípios de mesmo porte, que é de 1,90 viagens/dia [1].

No ano de 2011, foi realizada uma pesquisa de mobilidade urbana, chamada Pesquisa Origem-Destino (OD). Para fins de pesquisa e análise dos dados coletados, a cidade foi dividida em 55 áreas chamadas Zonas de Tráfego (ZT). Essas zonas foram obtidas por meio da divisão do território municipal em polígonos menores utilizando-se alguns critérios para agrupamento de áreas com características similares, tal como setores censitários do IBGE, malha viária, barreiras naturais e uso e ocupação do solo [8].

2.2 Dados Epidemiológicos

Os registros de casos de dengue foram fornecidos em formato CSV pelo Centro de Controle de Zoonoses(CCZ) do Departamento de Saúde da cidade de São José dos Campos e compreendem o período de 2013 a 2019. O arquivo recebido possui os seguintes campos: identificador, classificação dos casos, data, latitude, longitude e bairro. Foram desconsiderados da análise os registros classificados como "DESCARTADOS". Os demais casos formam uma tabela contendo coordenadas geográficas (Latitude e Longitude). Posteriormente, foi utilizado um dado complementar (*shapefile*) com a delimitação das Zonas de Tráfego - ZT para cruzamento com os casos de dengue. Cada caso de dengue foi então atribuído à uma ZT. Por meio desse *shapefile* final uma nova tabela foi gerada.

No arquivo que contém os casos de dengue, os casos estão organizados de forma sequencial em relação à data do registro, gerando uma série temporal, casos por semana, para cada uma dessas zonas. Cada coluna representa uma ZT e cada linha, o número de casos registrados para uma semana.

3 Desenvolvimento computacional

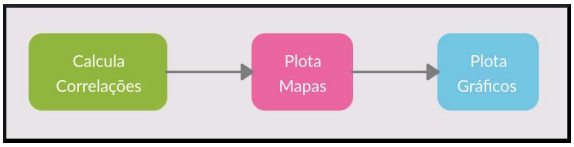


Figura 1: *Notebooks* desenvolvidos no trabalho e disponibilizados no github.

O código computacional foi desenvolvido utilizando linguagem Python e algumas bibliotecas específicas como Pandas, Numpy, Geopandas e Matplotlib. Foram gerados três *notebooks*, como pode ser visto na Figura 1, um para cada fim. Todos os códigos estão disponíveis no github e podem ser acessados em: https://github.com/catianascimento/ProjetoMobilidade.

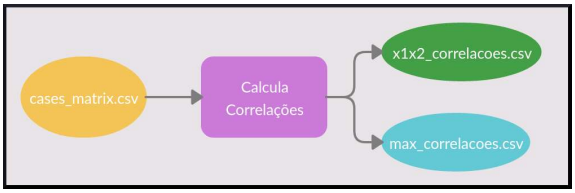


Figura 2: *Notebook* que calcula as correlações.

A Figura 2 mostra os arquivos usados como entrada e gerados por esse *notebook*. A partir do arquivo com registro de dados são gerados outros dois com o cálculo da correlação de Pearson [7], mostrada na equação (1) para cada par de zonas, onde x e y representam duas variáveis de séries temporais as quais desejamos calcular correlação. O valor de 'r' varia entre -1 e 1 da seguinte forma:

- r = 1 - correlação perfeita e positiva entre as duas séries;
- r = 0 - as duas séries não dependem linearmente uma da outra;
- r= -1 - correlação perfeita e negativa entre as duas séries.

O primeiro arquivo, "x1x2_correlations.csv", contém as correlações calculadas sobre os dados originais, utilizando a fórmula (1).

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{1}$$

O arquivo "max_correlations.csv"possui o valor de máxima correlação para um dado intervalo de atraso variando de -11 a 11 semanas. Nesse arquivo, fica armazenado qual o valor de k que

maximiza e qual o valor de máxima correlação.

$$r = \frac{\sum_{i=1}^{n+k}(x_i - \overline{x'})(y_{i+k} - \overline{y'})}{\sqrt{\sum_{i=1}^{n+k}(x_i - \overline{x'})^2(y_{i+k} - \overline{y'})^2}} \tag{2}$$

A fórmula (2) apresenta a alteração feita para calcular os atrasos nas séries temporais. Para um dado k, relacionamos x_i com y_{i+k}, por exemplo, x_1 com y_2, x_2 com y_2, x_2 com y_4, e assim por diante. As variáveis $\overline{x'}$ e $\overline{y'}$ representam as médias sobre os novos conjuntos de dados, uma vez que a cada k, as séries temporais possuem um elemento a menos para o cálculo. Depois disso, é feita a mesma coisa trocando a ordem dos conjuntos e, em seguida, é feita a união dos conjuntos dos resultados, como mostra a Figura 3.

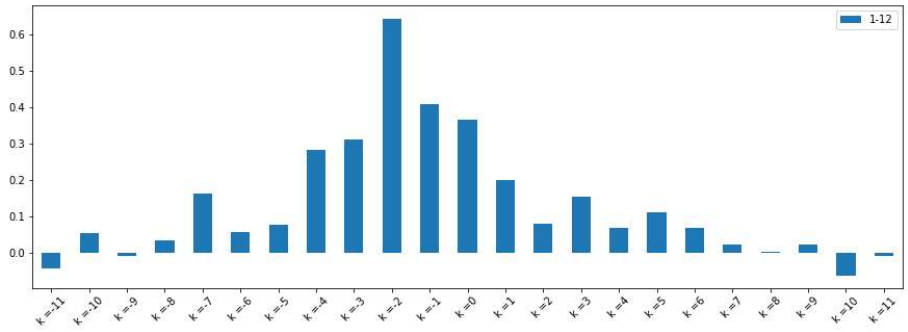


Figura 3: Um exemplo de correlações calculadas para cada atraso sobre as semanas, considerando um atraso entre -11 e 11 semanas.

A ideia é verificar se existe um atraso k em semanas que representa o tempo que a doença levaria pra se propagar e chegar a um outro bairro.

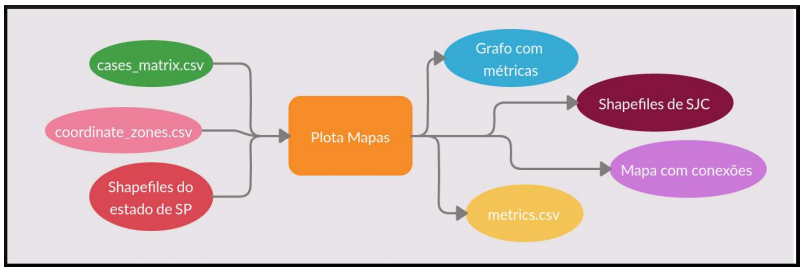


Figura 4: *Notebook* que gera mapas, shapefiles e métricas.

O segundo *notebook* "PlottingMapsWithEdges.ipynb", Figura 4 manipula diversos arquivos. Como entrada, ele recebe o arquivo "coordinates_zones.csv", que contém a LATITUDE e LONGITUDE do centróide de cada ZT, o arquivo "x1x2_correlations.csv", gerado pelo primeiro *notebook*, e shapefiles do estado de São Paulo, obtidos no site `http://geoftp.ibge.gov.br/`, fornecido pelo IBGE (Instituto Brasileiro de Geografia e Estatística).

A biblioteca Geopandas foi usada para manipular e gerar mapas da cidade de São José dos Campos, além de shapefiles específicos para a região. Um grafo foi gerado, usando a biblioteca igraph. As arestas do grafo são criadas baseadas no limiar de correlação entre as zonas. A partir deste grafo, são calculadas métricas que são guardadas dentro do arquivo "metrics.csv".

O último *notebook*, Figura 5, recebe como entrada quatro arquivos "distances_matrix.csv", contendo as distâncias entre as zonas, "mob_matrix.csv", que possui a mobilidade entre cada par de zonas e os arquivos "x1x2_correlations.csv"e "max_correlations.csv". A partir desses arquivos são gerados seis gráficos relacionando distâncias e correlações, mobilidade e correlações, distâncias e k(atraso) que maximiza correlação, mobilidade e k, distância e máxima correlação, e por fim, mobilidade e máxima correlação.

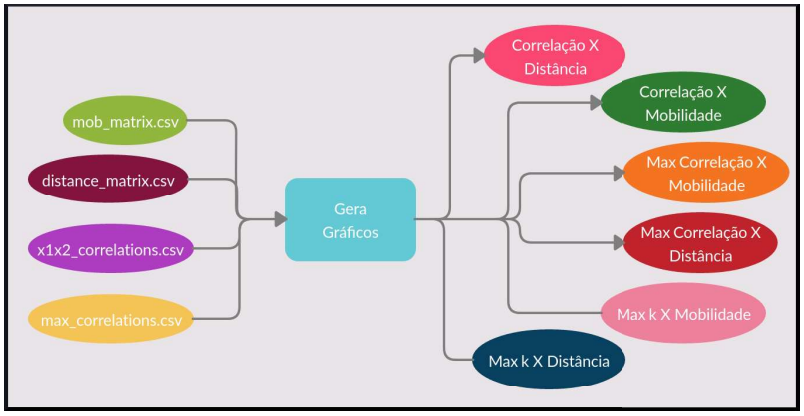
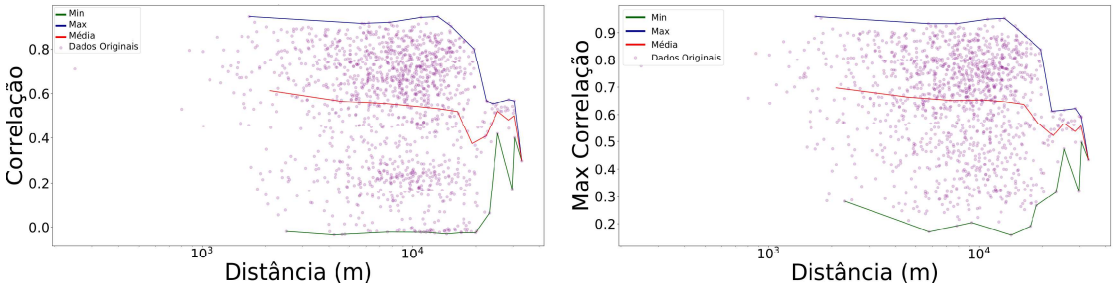


Figura 5: *Notebook* que gera gráficos de relações.

4 Resultados

A Figura 6(a) mostra uma leve declinação na curva, demonstrando que quanto maior a distância euclidiana (distância calculada a partir de uma linha reta traçada entre cada par de centróides), menor é a correlação entre séries temporais de casos de dengue registrados nas diferentes zonas. Para máximas correlações, Figura 6(b), a curva possui uma inclinação semelhante, porém deslocada para cima, uma vez que considera maiores valores de correlação.



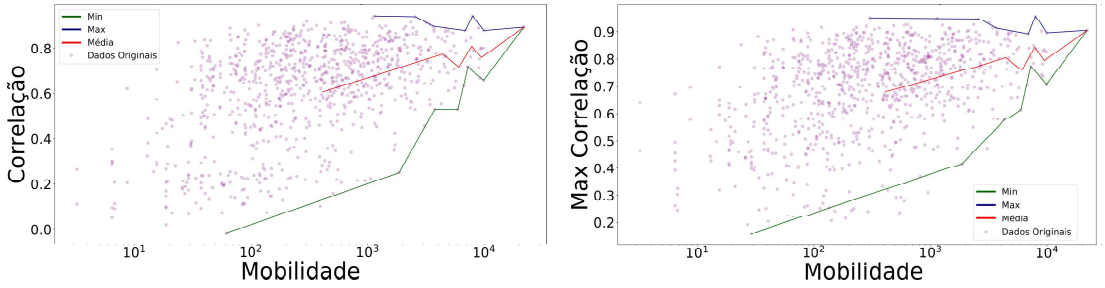
(a) Gráfico Correlação X Distância.

(b) Gráfico Máxima Correlação X Distância.

Figura 6: Comportamento da correlação entre séries temporais frente à distância entre as zonas relativas a cada série.

Os gráficos representados nas Figuras 7(a) e 7(b) mostram boa relação entre mobilidade e correlação de Pearson entre as séries temporais. À medida que a mobilidade aumenta, a correlação

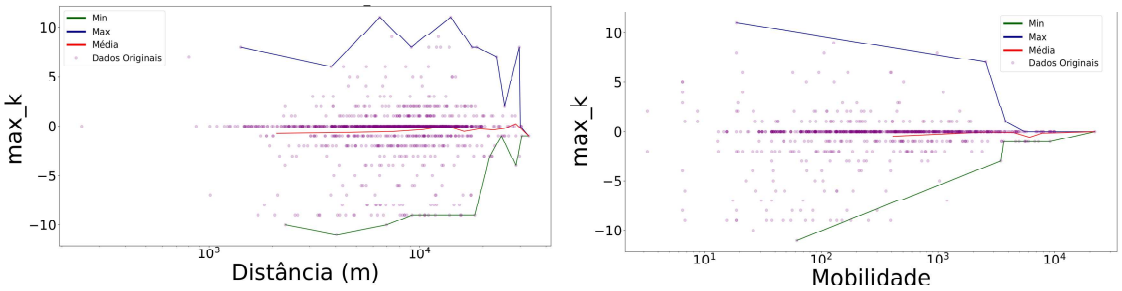
aumenta de forma considerável. Na segunda figura, essa relação fica mais aparente, passando de 0.9.



(a) Gráfico Correlação X Mobilidade. (b) Gráfico Máxima Correlação X Mobilidade.

Figura 7: Comportamento da correlação entre séries temporais frente à mobilidade de pessoas entre as zonas relativas a cada série.

Considerando o atraso que maximiza a correlação entre cada par de séries temporais, as Figuras 8(a) e 8(b) mostram maior concentração de pontos (valores de atraso) quando o valor de k é igual a 0, independentemente da distância ou mesmo da mobilidade entre as zonas.



(a) Gráfico Máximo K X Distância. (b) Gráfico Máximo K X Mobilidade.

Figura 8: Gráficos para Máximo K(atraso em semanas).

5 Conclusões

Neste trabalho, analisamos a correlação entre séries temporais de dengue em escala intra-urbana, considerando dados de distância geográfica e mobilidade de pessoas entre as regiões. O estudo de caso foi feito com dados reais da cidade de São José dos Campos/SP. Os resultados mostram que a média para as correlações decresce frente à distância geográfica e cresce com a mobilidade de pessoas entre as regiões, e que o atraso nulo (correlação convencional) é o mais frequente entre os que maximizam a correlação. Além disso, não foi encontrada dependência entre o valor do atraso e as informações de distância ou mobilidade.

Como trabalho futuro os dados de mobilidade e de casos serão confrontados sob uma outra abordagem, baseada em redes complexas, com o objetivo de captar não apenas as interações imediatas entre as zonas mas também informações de fluxo região [2, 10].

Referências

[1] ANTP. Sistema de Informações da Mobilidade Urbana - Relatório Geral 2012, `http://www.antp.org.br`, Último acesso em 03-03-2020, Associação Nacional de Transportes Públicos, 2012.

[2] Brockmann, D. e Helbing, D., *The hidden geometry of complex, network-driven contagion phenomena*, American Association for the Advancement of Science, volume 342, number 6164, pages 1337–1342, 2013.

[3] Cardoso, H. S. P. e Miranda, J. G. V. e Jorge, E. M. de F. e Moret, M. A., *Correlation between transport and occurrence of dengue cases in Bahia*, 2013.

[4] Ceron, W. e Santos, L. B. L. e Neto, G. D. e Quiles, M. G. e Candido, O. A., *Community Detection in Very High-Resolution Meteorological Networks*, IEEE Geoscience and Remote Sensing Letters, pages 1–4, 2019. DOI: 10.1109/LGRS.2019.2955508.

[5] Çınlar, E. Probability and Stochastics, *Springer-Verlag New York*, volume 261, 2011. DOI:10.1007/978-0-387-87859-1.

[6] Jucá, B. Dengue coloca o Brasil na mira de um novo surto em meio a preocupação com o coronavírus, `https://brasil.elpais.com/brasil/2020-02-13/dengue-coloca-o-brasil-na-mira-de-um-novo-surto-em-meio-a-preocupacao-com-o-coronavirus.html`, Último acesso em 20-02-2020, El Pais, 2020.

[7] Mukaka, M. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research, *Malawi medical journal : the journal of Medical Association of Malawi*, volume 24, pages 69–71, month 09, 2012.

[8] PMJSC e IPPLAN. Prefeitura de São José dos Campos - PMSJC E Instituto de Pesquisa, Administração e Planejameto - IPPLAN. - Atlas da Pesquisa Origem e Destino - Panorama da Mobilidade em São José dos Campos., `https://www.sjc.sp.gov.br/media/56152/atlas_origem_destino_baixa_res.pdf`, Último acesso em 01-03-2020, 2014.

[9] Saba, H. e Moret, M. A. e Barreto, F. R. e Araújo, M. L. V. e Jorge, E. M. F. e Nascimento Filho, A. S. e Miranda, J. G. V., *Relevance of transportation to correlations among criticality, physical means of propagation, and distribution of dengue fever cases in the state of Bahia*, Science of the Total Environment, volume 618, chapter 23, pages 971–976, 2018.

[10] Santos, L. B. L. e Carvalho, L. M. e Seron, W. e Coelho, F. C. e Macau, E. E. e Quiles, M. G. e Monteiro, A. M. V., *How do urban mobility (geo) graph's topological properties fill a map?*, Applied Network Science, volume 4, number 1, pages 1–14, 2019.

[11] Secretaria de Vigilância Sanitária - Ministério da Saúde. Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika), Semanas Epidemiológicas 1 a 4, 2020, volume 51, 2020.