1.03.04 - Ciência da Computação / Sistemas de Computação.

UM ESTUDO DE ANÁLISE E PREVISÃO DE SÉRIES TEMPORAIS APLICADO ÀS CURVAS DE INDICES DE ISOLAMENTOS DEVIDOS À PANDEMIA DE COVID-19

Diego Clementino Pompeo[1], Arnaldo R. A. Vallim Fº[2]
1. Estudante da Faculdade Computação e Informatica da Universidade Presbiteriana Mackenzie
2. Professor da Universidade Presbiteriana Mackenzie – Faculdade de Computação e Informatica/Orientador

Resumo

Nesta pesquisa, foi feita uma análise e previsão de séries temporais do índice de isolamento devido a pandemia da COVID-19, durante um período de um ano, fevereiro de 2020 a fevereiro de 2021, das principais cidades das 16 regiões administrativas do estado de São Paulo, divididas em 4 sub-regiões: a Capital, a Região Metropolitana de São Paulo (RMSP), o Litoral e o Interior. Para que fosse possível a análise e previsão do índice de isolamento de cada uma dessas regiões, foram utilizados os dados disponibilizados pelo governo de Estado de São Paulo, que após a limpeza, a transformação e a integração, foi possível o desenvolvimento de quatro tipos de análise: cálculo de média móvel, curva de tendência e curva de sazonalidade (anual e semanal), e ainda a previsão para períodos futuros. As previsões foram feitas para até duas semanas à frente e com uma margem de erro estipulada dentro no próprio algoritmo desenvolvido. Todo o estudo foi desenvolvido com o apoio da linguagem de programação R, e suas bibliotecas, tornando assim, possível a compreensão e análise do comportamento dos índices de isolamento de cada município analisado.

Palavras-chave: Ciência de Dados; Estatística;

Apoio financeiro: PIBIC Mackenzie

Trabalho selecionado para a JNIC: UPM

Introdução

A doença causada pelo coronavírus, a COVID-19, teve seu primeiro caso registrado na China no final de 2019, e em março de 2020, a OMS declarou o surto como pandemia. Após essa declaração da OMS, diversos países adotaram o lockdown (confinamento) como forma de prevenção contra a COVID-19, visto que, por hora, não se tinham vacinas e nem medicamentos 100% eficazes contra a doença.

Em alguns países, mesmo adotando o confinamento como forma de prevenção, o número de pessoas contaminadas ainda continua crescente, devido a população nãorespeitar o confinamento, causando assim, um número crescente também de mortes, como é o caso do Brasil (Observatório COVID-19 BR, 2020)

A evolução da pandemia vem se desenvolvendo de forma acelerada no país, conforme mostram os dados oficiais do Ministério da Saúde e das Secretarias de Saúde dos estados da federação (Ministério da Saúde, 2020).

A previsão do que deve ocorrer nos próximos períodos, considerando-se a evolução das séries temporais de casos registrados, poderia representar um apoio relevante, de forma a propiciar um adequado planejamento das ações dos administradores públicos e privados. E, uma vez que tenham modelos eficazes de análise e previsão, estes poderiam se constituir em ferramentas de apoio a estudos sobre outros tipos de pandemia, que podem vir a ocorrer.

Inúmeros estudos estão sendo desenvolvidos em todo o mundo, e há uma quantidade grande de dados disponíveis, o que é um ponto favorável para o desenvolvimento de pesquisas associadas à Ciência de Dados, que é o caso desta proposta de projeto, que poderá assim, ajudar a compreender melhor esse fenômeno, especialmente pelas técnicas de Previsão de Séries Temporais, que permitem analisar um conjunto de observações ordenadas no tempo, e prever seu comportamento futuro.

Assim, é que surge esta proposta de projeto de Iniciação Científica. No caso específico deste projeto, o trabalho contemplou a análise e previsão de séries temporais de variáveis relevantes associadas à pandemia de COVID-19, estudando séries de tempo de dados internos do país e séries de outros países, particularmente, as séries cronológicas de índices de isolamento das pessoas nos municípios do estado de São Paulo.

Metodologia

Esta foi uma pesquisa de natureza aplicada, com uma abordagem quantitativa, que utilizou como meios bibliografia, bases de dados, ferramentas computacionais e experimentos em ambiente de laboratório. Sua finalidade foi de desenvolvimento de uma metodologia para análise e previsão de séries temporárias.

Após uma revisão bibliográfica do tema, para a análise dos dados do índice de isolamento das principais cidades do estado de São Paulo, foi utilizado como fonte os dados disponibilizados pelo governo do estado de São Paulo, através do site do SIMISP - Sistema de Monitoramento Inteligente de São Paulo (www.saopaulo.sp.gov.br/coronavirus/isolamento/, que é viabilizado por meio de acordo com as operadoras de telefonia Vivo, Claro, Oi e TIM, através da ABR (Associação Brasileira de Recursos em Telecomunicações) e do IPT (Instituto de Pesquisas Tecnológicas), para que o Estado possa consultar informações agregadas e anônimas

sobre deslocamento nos municípios paulistas mapeados.' (ISOLAMENTO, 2021).

Segundo o SIMI, em respeito à proteção de dados, as informações são aglutinadas e anonimizadas, respeitando a privacidade dos usuários. Apresentando dessa forma, dados georreferenciados agrupados para elaborar políticas públicas que aprimorem as medidas de isolamento social para o enfrentamento ao coronavírus." (ISOLAMENTO, 2021)

Para essa análise foram utilizados os dados de índice de isolamento a partir do dia 26/02/2020 até o dia 23/04/2021.

Após a captura dos dados em uma planilha, foi necessário fazer alguns ajustes, a transformação dos dados que estavam em formato de texto para número, e a remoção dos acentos e da cedilha, para que fosse possível desenvolver qualquer análise com os dados, pela tecnologia R.

Além do levantamento dos dados, as seguintes etapas foram desenvolvidas no estudo:

. Revisão Bibliográfica
. Levantamento de Dados
. Pré processamento de Dados
. Aplicação de Algoritmos por meio de desenvolvimento de Scripts em R
. Desenvolvimento de Experimentos
. Análise de Resultados e Documentação

Resultados e Discussão

Antes do desenvolvimento das análises foi necessário um processo de limpeza, transformação e integração dos dados na própria planilha Excel em que são fornecidos pelo SIMI-SP.

A análise foi desenvolvida com o apoio de ferramentas computacionais disponíveis na linguagem R e em suas bibliotecas de algoritmos de séries temporais e de visualização de dados.

Em particular, foi usado o pacote "prophet", que implementa um algoritmo que projeta uma série temporal fazendo uso de um modelo matemático do tipo aditivo, em que tendências não lineares são ajustadas aos dados e a estas são adicionadas sazonalidades, anual, semanal e diária. O pacote é considerado robusto para tratar dados ausentes, variações em tendências e outliers (CRAN.R, 2021a). Além disso, outra biblioteca importante foi aquela do pacote 'forecast, que tem implementados algoritmos para análise e previsões de séries temporais, e visualização de resultados. Inclui o método de suavização exponencial e médias móveis e modelos clássicos de séries temporais (CRAN.R, 2021b). E uma última biblioteca que merece destaque é aquela do pacote 'ggplot2, que possibilita a criação de gráficos, baseados na chamada gramática de gráficos (COX,2007), que trata da questão dos gráficos de forma ampla. No caso deste pacote, dá-se o acesso à base de dados, e informa-se como mapear as variáveis de interesse e definir a estética desejada. Com isto, muitas possibilidades de geração de gráficos estão disponíveis.

Conforme mencionado anteriormente, tinha-se em mãos series temporais do índice de isolamento das pessoas para os municípios do estado de São Paulo, durante um período de um ano entre fevereiro de 2020 e fevereiro de 2021. Para esta análise foram selecionadas as principais cidades das 16 regiões administrativas do estado de São Paulo, divididas em 4 sub-regiões: a capital, a região metropolitana de São Paulo (RMSP), o litoral e o interior do estado.

Em cada cidade escolhida foram gerados cinco tipos de análise:

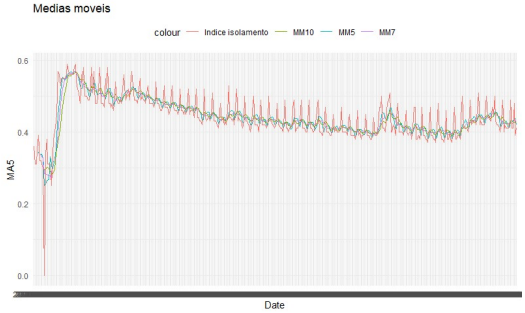
. Diagrama de dispersão dos dados originais do Índice de Isolamento;
. Médias Móveis de 5, 7 e 10 dias;
. Análise de Tendência;
. Análise de Sazonalidade semanal e anual
. Previsões com respectivo intervalo de confiança ao nível de 95%.

A média móvel é uma média dos valores de períodos que vão se sobrepondo. Assim, para uma média móvel de 5 dias, tem-se uma primeira média para os dias 1 a 5. A segunda média considera os dias 2 a 6. A terceira, considera os dias 3 a 7, e assim por diante.

A tendência é uma curva que identifica na curva dos dados originais tendências de crescimento, queda ou estabilização e representa esses comportamentos em diferentes períodos, por meio de uma curva.

Já a sazonalidade identifica movimentos de alta ou baixa para períodos específicos bem definidos no tempo. Uma sazonalidade semanal, por exemplo, identifica esses movimentos por dia da semana.

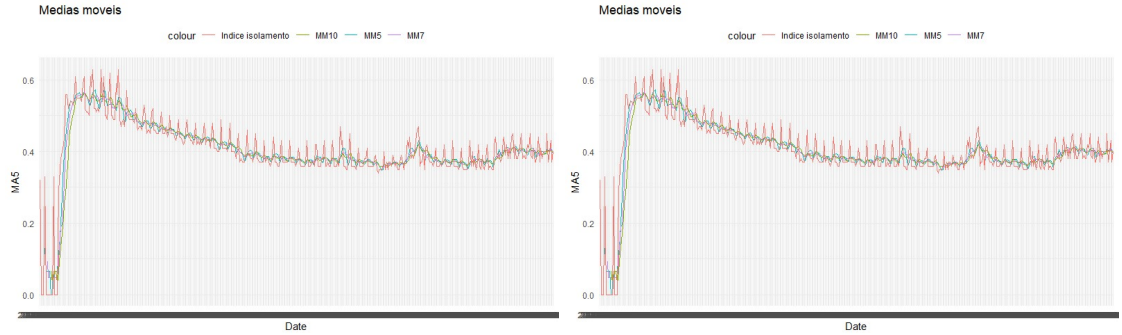
- Análises de São Paulo (Capital)



Série Original e Médias Móveis – São Paulo - SP

Notou-se que ao início da pandemia, muito por falta de informação de como tratá-la, o índice de isolamento era baixo, porém ainda no início, as pessoas foram adquirindo informações de como se portar diante de uma pandemia, fazendo assim, com que o índice de isolamento aumentasse visto que uma das melhores alternativas para diminuição do contágio da COVID-19 era o distanciamento e isolamento social.

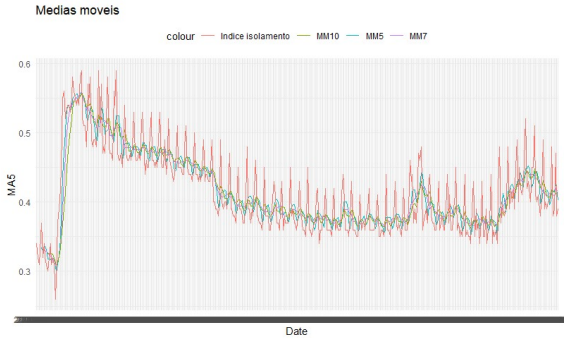
- Análises da Região Metropolitana de São Paulo (Guarulhos e São Bernardo do Campo)



Série Original e Médias Móveis – Guarulhos e São Bernardo de Campos respectivamente

A cidade de Guarulho teve o mesmo comportamento que na Capital, porém a cidade de São Bernardo do Campo, teve um índice de isolamento maior e mais constante.

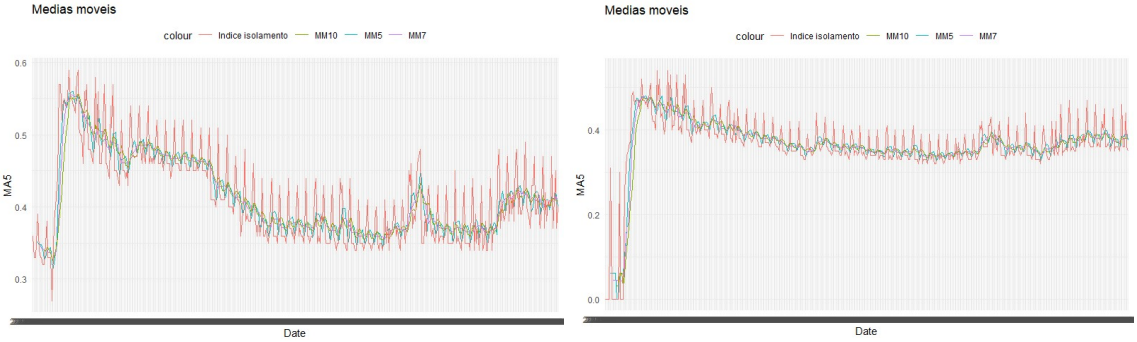
- Análises do Litoral do Estado de São Paulo (Santos)

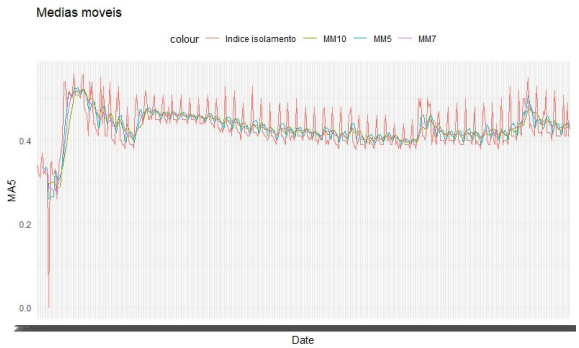


Série Original e Médias Móveis – Santos

Na cidade de Santos, notasse que o comportamento de isolamento demostrado, teve o mesmo o início que nas outras cidades, porém a cidade de Santos teve uma queda mais brusca do índice de isolamento, do que nas cidades da região metropolitana e na capital. Muito se deve a saída de pessoas dessas cidades para a cidade de Santos. Causando uma pequena peculiaridade, já que no início da previsão se tem um aumento considerável do índice de isolamento, e antes da metade do período da previsão analisado esse índice despenca assim como nos dados originais do início da pandemia.

- Análises do Interior do Estado de São Paulo (Campinas, Presidente Prudente e Ribeirão Preto)





Série Original e Médias Móveis – Campinas, Presidente Prudente e Ribeirão Preto respectivamente

Os dados originais da cidade de Campinas tiveram a mesma queda brusca do índice de isolamento, e um baixo valor do índice entre o período de Agosto de 2020 até Dezembro de 2020, que voltaram a cair no mês de Janeiro de 2021.

Diferentemente da cidade de Campinas, as cidades de Presidente Prudente e Ribeirão Preto, que representam o Interior, tiveram, a partir do mês de Abril, somente uma leve queda no índice, porém o índice ainda permanecia alto

Conclusões

Em resumo, conclui-se que através da sazonalidade semanal, todas as cidades tiveram nos dias de final de semana como os dias com maiores índices de isolamento. Além disso, nota-se que na série temporal dos dados originais mais as médias móveis, as cidades da região metropolitana de São Paulo, a capital e as cidades do interior, exceto a cidade de Campinas, tiveram no geral altos índices de isolamento, durante, ao menos um ano, em relação a cidade litorânea de Santos e a cidade de Campinas, que obtiveram baixos índice de isolamento, em comparação desse período.

Outro ponto a ser observado, foi que nas previsões geradas do índice de isolamento de todas as cidades apontadas no estudo, tiveram a mesma característica, um leve aumento do índice, e antes da metade do período ocorria uma queda.

Referências bibliográficas

ALBUQUERQUE, C. G. D.; COSTA, M. A.; CURI, R. L. C.; KRAUSE, C.; LUI, L.; SANTOS, R. M. D. S.; TAVARES, S. R. (2020). Apontamentos sobre a dimensão territorial da pandemia da COVID-19 e os fatores que contribuem para aumentar a vulnerabilidade socioespacial nas unidades e desenvolvimento humano deáreas metropolitanas brasileiras. Diretoria de Estudos e Políticas Regionais, Urbanas e Ambientais [Dirur], nº 15, p. 7.

ALESSI, Gil e ROSSI, Marina (2020). Quantos de seus vizinhos em São Paulo contraíram o coronavírus? Mapa interativo da USP revela. El Pais, 11 de junho de 2020. Disponível em:https://brasil.elpais.com/brasil/2020-06-09/quantos-de-seusvizinhos-em-sao-paulo-contrairam-o-coronavirus-mapa-interativo-da-usprevela.html?ssm=FB_CC&fbclid=IwAR2wakC_oDhPGdGV0y8PJ8Nggwy5GMfrP JFHSA NBGc7w2sXNsRN6PE5PEbg

BOX, G. E. P. e JENKINS, G. M. (1976) Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, Ca.537p.

CRAN.R (2021a). Package prophet: Automatic Forecasting Procedure. The Comprehensive R Archive Network Project. Acesso em abril/2021. Disponível em: https://cran.r-project.org/web/packages/prophet/index.html

CRAN.R (2021b). Package forecast: Forecasting Functions for Time Series and Linear Models. The Comprehensive R Archive Network Project. Acesso em abril/2021. Disponível em: https://cran.r-project.org/web/packages/forecast/index.html

CRAN.R (2021c). Package ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. The Comprehensive R Archive Network Project. Acesso em abril/2021, em: https://cran.r-project.org/web/packages/ggplot2/index.html

COX, N. J. (2007). Journal of Statistical Software. January 2007, Volume 17, Book Review 3. http://www.jstatsoft.org/

HYNDMAN, R.J., & ATHANASOPOULOS, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. Acessado em: janeiro/2021. Disponível em: https://otexts.com/fpp2/

Isolamento | Governo do Estado de São Paulo, Isolamento | Governo do Estado de São Paulo, disponível em: <https://www.saopaulo.sp.gov.br/coronavirus/isolamento/>, acesso em: 2 Aug. 2021

LATORRE, M. D. R. D. D. O.; CARDOSO, M. R. A. (2001). Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. Time series analysis in epidemiology: an introduction to methodological aspects. Revista Brasileira de Epidemiologia. Vol. 4, Nº 3

MINISTÉRIO DA SAÚDE (2020). (2020). Painel Coronavírus. https://covid.saude.gov.br/

MORETTIN, P. A.; TOLOI, C. M. C. (1981). Modelos para Previsão de Séries Temporais. IMPA – Instituto de Matemática Pura e Aplicada. Rio de Janeiro, RJ. 372 p.

OBSERVATÓRIO COVID-19 BR. COVID-19 no Brasil. https://covid19br.github.io/

REIS, M. M. (2000). Análise de séries temporais. Disponivel em: <https://www.inf.ufsc.br/~marcelo.menezes.reis/Cap4.pdf> Acesso em: 10 de junho de 2020. Departamento de Informática e Estatística da Universidade Federal de Santa Catarina Campus Joinville – INE/UFSC.