Identificação de períodos de instabilidade de coeficientes da série temporal do COVID-19

Identification of periods of instability in COVID-19 timeseries coefficients

Leonardo Luís Schneider Simon*, Daniel Cavalcanti Jeronymo†

RESUMO

Com as primeiras manifestações do vírus SARS-CoV-2 em dezembro de 2019 e seu alastramento pelo globo, diversos estudos que buscam prever futuros cenários de transmissão e das mortes causadas por ele têm surgido no meio científico. Pesquisas baseadas na avaliação da série temporal em diversos países têm mostrado precisão em determinar estes cenários baseando-se em modelos como de NAÏVE, Holt-Winters e ARIMA. Neste trabalho, buscou-se identificar os períodos de instabilidade de coeficientes da série temporal do COVID-19 no Brasil, utilizando para isso o método dos mínimos quadrados recursivos para encontrar os erros de predição e, para a detecção dos outliers, o método da floresta de isolamento. Os resultados obtidos indicaram de forma satisfatória o que era esperado para a quantidade de dias analisados, sendo possível identificar no gráfico, através dos outliers e das datas onde aparecem, períodos em que houve aumento ou diminuição abrupta do número diário de mortes pelo vírus no país e quais eventos reais poderiam estar associados à estas variações.

Palavras-chave: COVID-19, SARS-CoV-2, Mínimos Quadrados Recursivos, Identificação de Sistemas, Detecção de Outliers.

ABSTRACT

Since the first appearance of SARS-CoV-2 virus in December 2019 and its spread across the globe, several studies that seek to predict future scenarios of transmission and deaths have emerged in the scientific community. Surveys based on the evaluation of timeseries in several countries have demonstrated accuracy in determining these scenarios based on models such as NAÏVE, Holt-Winters and ARIMA. In this work, periods of instability COVID-19 timeseries coefficients are identified in Brazil, using the recursive least squares method to find the prediction errors and, for the detection of outliers, the isolation forest method. The results obtained indicated what was expected for the number of days analyzed, and it was possible to identify in the graph, through the outliers and the dates where they appear, periods in which there was an abrupt increase or decrease in the daily number of deaths from the virus in the country and what real events could be associated with these variations.

Keywords: COVID-19, SARS-CoV-2, Recursive Least Squares, System Identification, Outliers Detection.

1 INTRODUÇÃO

* Engenharia de Computação, Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil; leonardo.luisimon@hotmail.com
† Universidade Tecnológica Federal do Paraná, Campus Toledo; danielc@utfpr.edu.br

Desde o aparecimento dos primeiros casos de COVID-19 em humanos em dezembro de 2019 até se tornar uma pandemia, a humanidade tem buscado entender as origens deste vírus e como amenizar seus impactos da melhor forma possível (WHO, 2020). No Brasil, os primeiros casos da doença foram confirmados em fevereiro de 2020 e, desde então, diversas medidas têm sido tomadas buscando amenizar os impactos desta pandemia (FONGARO et al, 2021).

Acompanhando a disseminação e contágio da doença, muito estudos têm surgido com o objetivo de prever possíveis cenários em relação ao aumento de casos e de mortes, como no trabalho de Barria-Sandoval et al. (2021), onde os autores modelaram casos de contágios e mortes por COVID-19 no Chile usando modelos auto-regressivos integrados de médias móveis (ARIMA - *Autoregressive Integrated Moving Average*), técnicas de suavização exponencial e modelos de Poisson para dados dependentes do tempo, avaliando a precisão destas previsões através de um conjunto de treinamento e de testes. Após a análise destes modelos, ou autores tiveram como resultado que o melhor método para prever o número de novos casos foi o modelo ARIMA, enquanto que para prever o número de mortos foi o método não sazonal de suavização de tendência amortecida, a qual se trata de uma técnica de suavização exponencial. No entanto, apesar dos resultados, os autores concluíram que outros métodos podem prever melhor o comportamento dos dados registrados, sendo que isto dependerá da base de dados utilizada.

Já no estudo de Maurya e Singh (2020), os autores fizeram uso de quatro modelos diferentes para analisar a série temporal da COVID-19, com o objetivo de prever futuros cenários em relação à pandemia. Os modelos utilizados foram o de NAÏVE, a tendência linear de Holt, Holt-Winters e ARIMA. Através dos resultados obtidos, os autores puderam concluir que o modelo NAÏVE foi o melhor dentre os 4 analisados, uma vez que possuía o menor erro quando comparado aos outros. Contudo, os autores destacam que este modelo era o melhor para o momento em que a pesquisa foi realizada, pois a quantidade de dados era menor, destacando ainda que, futuramente, havendo uma base de dados maior, os outros modelos também podem vir a apresentar melhores resultados.

Buscando realizar uma abordagem diferente do que tem sido proposto em outros artigos, este trabalho visa explorar a identificação de um modelo linear de ordem 8 por mínimos quadrados recursivos para identificar a série temporal de mortes causadas por COVID-19 no Brasil. Os parâmetros da série temporal são rastreados ao longo do tempo e são identificadas datas com variações bruscas dos parâmetros identificados, sendo traçados paralelos com eventos próximos às datas encontradas a fim de responder a questão: é possível explicar as variações bruscas na série temporal de mortes por COVID-19 através de eventos reais próximos às datas encontradas?

Este trabalho está organizado da seguinte maneira. A seção 2 apresenta a metodologia utilizada no trabalho e os fundamentos da técnica de mínimos quadrados recursivos. A seção 3 apresenta os resultados obtidos e, por fim, na seção 4 há a conclusão do artigo.

2 MÉTODO

Para o desenvolvimento do trabalho, optou-se por utilizar a linguagem Python, uma vez que esta é familiar aos autores, e por possuir bibliotecas específicas para o tratamento e análise de dados.

Com a linguagem definida, pesquisou-se uma base de dados que fornecesse a série histórica do total de óbitos causados pelo vírus SARS-CoV-2 no Brasil. A base escolhida foi a de POMBO (2021), pois nela temos os dados fornecidos publicamente pelo Centro de Ciência e Engenharia de Sistemas da Universidade Johns Hopkins (JSU CCSE) (CSSEGISandData, 2021) transformados no formato JSON (*JavaScript Object Notation*). Esta base foi escolhida pois os dados fornecidos provém de uma das principais referências

mundiais no monitoramento da COVID-19, que é a Universidade Johns Hopkins, e por estar com os dados em formato JSON.

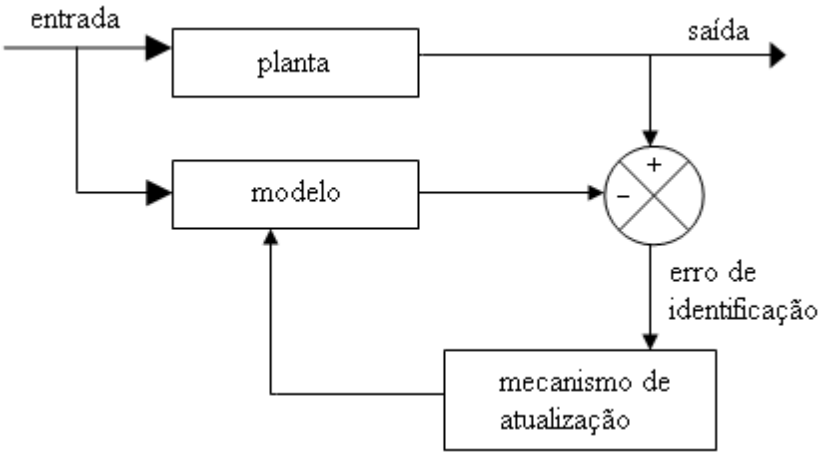
Inicialmente os dados são tratados de forma a desconsiderar os dias em que ainda não haviam mortes registradas, isto é, anteriores ao dia 17 de março de 2020 (data da primeira morte registrada no Brasil por COVID-19). Em seguida, é utilizado o método dos Mínimos Quadrados Recursivos utilizando para o número de amostras o valor de 8. Este valor foi definido após analisar a série temporal através dos Mínimos Quadrados Recursivos para os primeiros 150 valores de números de amostras, sendo que o 8º valor é o de menor erro, portanto, a melhor opção para a análise dos dados. Por fim, são encontrados os outliers com base nos erros de predição da análise temporal. Para gerar os outliers, escolheu-se um limiar de contaminação de 5%, pois este resulta em 28 dias a serem analisados, sendo estes identificados pela Floresta de Isolamento.

2.1 Mínimos quadrados recursivos

O método dos Mínimos Quadrados Recursivos é a essência da análise da série temporal das mortes diárias por COVID-19, a qual depende principalmente da eficiência do estimador. A estimação dos parâmetros é realizada baseando-se nas medidas obtidas da entrada e saída do processo. Os parâmetros estimados são, em geral, variantes no tempo e o modelo estimado é uma "simplificação" do sistema real. Isto permite a atualização dos parâmetros, em um modelo linear, que pode estar representando uma planta não-linear (COELHO, 2004).

Em cada período de amostragem novas medidas de entrada e de saída tornam-se disponíveis e são utilizadas com o modelo atual para gerar um novo erro de identificação. Pode-se visualizar o processo de estimação em termos de um modelo paralelo conforme mostrado na Fig. 1 (COELHO, 2004).

Figura 1 – Esquema de operação do estimador



Fonte: Autoria própria (2021).

Para determinar a estimativa ótima é necessário definir uma função custo que minimize o índice de desempenho (J). Este índice é uma medida quantitativa do desempenho do estimador, medindo o desvio do valor estimado em relação ao valor real da saída da planta. O objetivo é encontrar a estimativa dos parâmetros desconhecidos que minimize a função custo representada na Eq. (1), onde y(k) é a variável do processo, ŷ(k) é a estimação da saída e N é o número de amostras da experimentação (COELHO, 2004).

$$J = \frac{1}{N}\sum_{k=1}^{N}[y(k) - \hat{y}(k)]^2 \quad (1)$$

2.2 Detecção de outliers

Para melhor visualização dos resultados, aplicou-se o algoritmo da floresta de isolamento para encontrar os outliers dos resultados. Estes outliers, também conhecidos por anomalias, são padrões de dados que apresentam características distintas daqueles dados normais da série. A detecção destas anomalias é bastante importante, pois elas são capazes de fornecer informações úteis em diversas situações, como por exemplo, anomalias em transações com cartões de crédito, que poderiam indicar algum tipo de atividade fraudulenta, ou ainda, anomalias no padrão de tráfego de uma rede de informática, as quais poderiam indicar acessos não autorizados na rede. A maior parte das abordagens existentes para detecção de anomalias gera um perfil de dados normais e os dados que estiverem fora destes são considerados anomalias. Alguns métodos, como os estatísticos e os de agrupamento, fazem uso deste tipo de abordagem, no entanto, ela apresenta alguns problemas, como o fato de ser otimizada para detectar o perfil de dados normais, mas não para detectar as anomalias, logo, os resultados da detecção muitas vezes não são tão bons quanto o esperado, podendo causar a identificação de anomalias incorretas (que deveriam ser identificados como dados normais) ou detectando poucas anomalias. Além disso, outro inconveniente dos métodos baseados neste tipo de abordagem é que, muitas vezes, são limitados a dados de baixa dimensão e pouco tamanho, por causa da sua elevada complexidade computacional (LIU et al., 2008).

Para evitar estes problemas, neste artigo foi utilizado o método da floresta de isolamento para realizar a identificação dos outliers. Este método se baseia em construir um conjunto de árvores de isolamento para um conjunto de dados. Uma árvore de isolamento tem como função isolar os outliers para perto das raízes das árvores ao mesmo tempo em que leva os pontos normais para as extremidades mais próximas das folhas. Logo, a floresta de isolamento nada mais é que uma um agrupamento de árvores de isolamento capaz de verificar todo nosso conjunto de dados, sendo que os outliers são os pontos que têm menor distância das raízes nas árvores. O desempenho deste tipo de análise converge rapidamente para uma pequena quantidade de árvores, sendo necessária uma menor quantidade de dados para detectar os outliers e ter desempenho satisfatório (LIU et al., 2008).

3 RESULTADOS

Aplicando o método dos mínimos quadrados recursivos para a série temporal das mortes por COVID-19 no Brasil e estimando os erros de predição, gerou-se o gráfico da Fig. 2. Neste é possível observar a variação dos erros de predição (linha azul) em função do tempo em dias, sendo que as maiores variações nos erros, tanto positivamente quanto negativamente, representam dias em que foram registrados um grande aumento ou diminuição de mortos em relação aos dias anteriores.

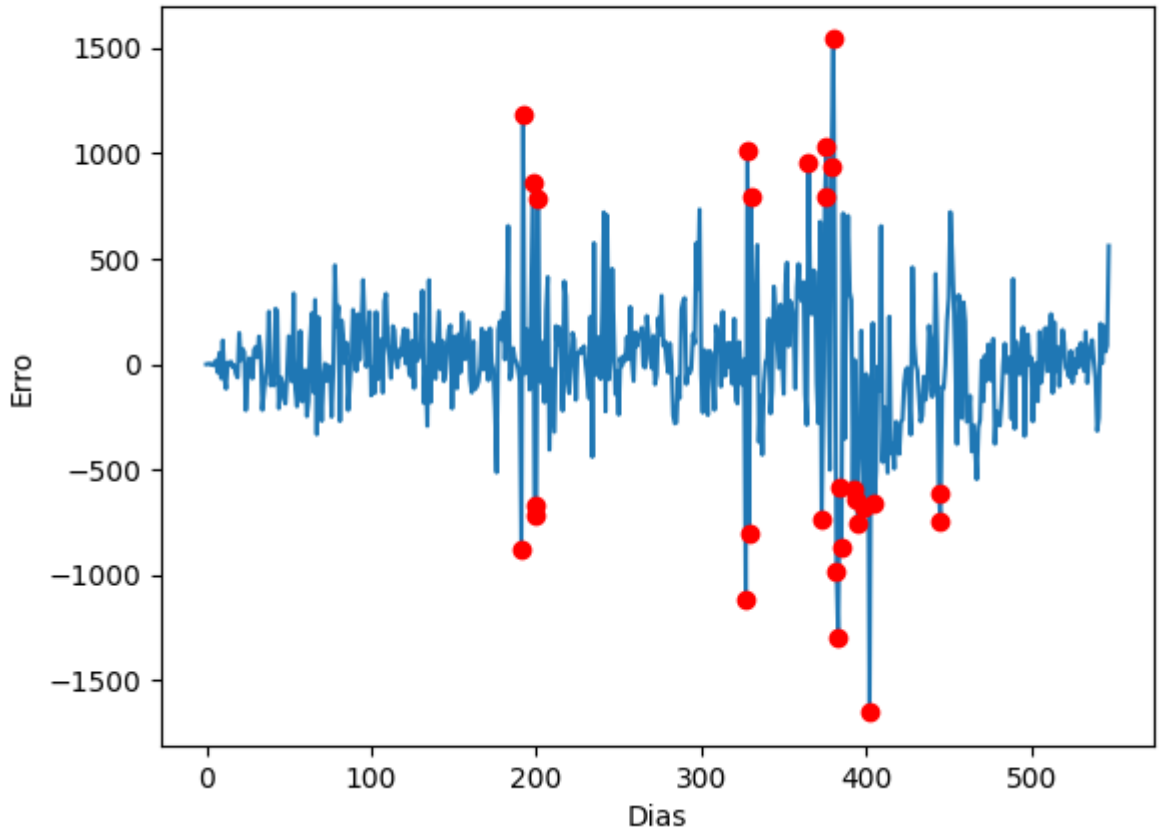
Já os pontos vermelhos do gráfico se referem à estimativa dos outliers através do método da floresta de isolamento, sendo que estes representam os dias em que houve uma variação anormal em relação aos pontos anteriores.

Analisando a posição dos outliers, é possível notar que os primeiros outliers estão próximos do 200º dia, onde aparecem 6 outliers. Ao verificar o número diário de mortes no Brasil para este mesmo período, percebe-se que se trata do final da primeira onda de mortes no país, logo, os outliers estão registrando a queda da média anteriormente considerada normal de mortes.

Os próximos pontos estão registrados após o dia 300 e antes do que seria o dia 350, sendo registrados 4 outliers neste período. Comparando com os números de mortes diárias registrados para este período, é possível associar o aparecimento destas anomalias com o aumento de casos que houve logo após as festas de final de ano, as quais poderiam justificar o aumento dos casos.

Por fim, o restante dos outliers se encontram próximos ao dia 400 (fevereiro de 2021), que coincide com o início da segunda onda de casos e mortes no país, a qual, segundo Hojo-Souza (2021), poderia estar relacionada com a disseminação de variantes mais transmissíveis da doença, bem como a diminuição da imunização das pessoas que já tinham contraído a doença e se recuperado.

Figura 2 – Resíduos (erros) e outliers da identificação da série temporal



Fonte: Autoria própria (2021.

4 CONCLUSÃO

Após a análise dos resultados foi possível concluir que o método dos mínimos quadrados recursivos combinado com a floresta de isolamento é bastante útil para identificar dados anormais dentro de uma série temporal, neste caso em específico, das mortes diárias por COVID-19 no Brasil. Portanto, estes resultados ajudam a entender como a série temporal e o uso de outliers podem ser usados para avaliar eventos reais próximos às datas encontradas, como por exemplo, no caso do aumento de mortes devido à disseminação de novas variantes da SARS-CoV-2, sendo estes resultados diferente de outros trabalhos pesquisados, os quais buscam prever futuros cenários para a pandemia.

Uma sugestão para o futuro em relação ao aprofundamento do artigo seria como o aumento da quantidade de outliers impactaria na visualização dos dados, ou ainda como datas comemorativas ou medidas governamentais impactaram no aumento ou decréscimo do número de casos.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e o primeiro autor agradece a UTFPR por uma bolsa de iniciação científica (CNPq, processo: 103509/2021-6/PQ).

REFERÊNCIAS

WHO. **Origin of SARS-CoV-2**, World Health Organization, WHO reference number: WHO/2019-nCoV/FAQ/Virus_origin/2020.1, 2020.

FONGARO, Gislaine; STOCO, Patrícia Hermes; SOUZA, Doris Sobral Marques; GRISARD, Edmundo Carlos; MAGRI, Maria Elisa; ROGOVSKI, Paula; SCHÖRNER, Marcos André; BARAZZETTI, Fernando Hartmann; CHRISTOFF, Ana Paula; DE OLIVEIRA, Luiz Felipe Valter; BAZZO, Maria Luiza; WAGNER, Glauber; HERNÁNDEZ, Marta; RODRÍGUEZ-LÁZARO, David. **The presence of SARS-CoV-2 RNA in human sewage in Santa Catarina, Brazil, November 2019.** Science of the Total Environment. 2021.

BARRIA-SANDOVAL, Claudia; FERREIRA, Guillermo; BENZ-PARRA, Katherine, LÓPEZ-FLORES, Pablo. **Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques**: A comparative study, PLoS ONE 16(4): e0245414. https://doi.org/10.1371/journal.pone.0245414. 2021.

POMBO, Rodrigo. **Covid19**. Disponível em: <https://github.com/pomber/covid19>. Acesso em: 09/09/2021.

CSSEGISandData. **COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Disponível em:** <https://github.com/CSSEGISandData/COVID-19>. Acesso em 09 de setembro de 2021.

MAURYA, Sujeet; SINGH, Shikha. **Time Series Analysis of the Covid-19 Datasets**, 2020 IEEE International Conference for Innovation in Technology (INOCON), Bengaluru, India, 6-8 Nov. 2020.

COELHO, Antônio Augusto Rodrigues; COELHO, Leandro dos Santos. **Identificação de Sistemas Dinâmicos Lineares**. Florianópolis: editora da UFSC, 2004.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. **Isolation Forest**. 2008 Eighth IEEE International Conference on Data Mining. 2018.

HOJO-SOUZA, Fernanda Sumika; HOJO-SOUZA Natália Satchiko; SILVA, Cristiano Maciel; GUIDONI, Daniel Ludovico. **Second wave of COVID-19 in Brazil**: younger at higher risk, European Journal of Epidemiology, volume 36, pag. 441–443. 2021.