Casos e mortes por Covid-19 no Brasil: uma abordagem por redes neurais de grafos

Covid-19 cases and deaths in Brazil: a graph neural network approach

Lucas Caldeira de Oliveira *, Dalcimar Casanova†

31 de agosto de 2021

RESUMO

Dado o impacto que o vírus SARS-CoV-2 tem causado no mundo, o entendimento de como as curvas de contágios e óbitos vão evoluir possuem grande relevância, permitindo que decisões no âmbito de conter o avanço da pandemia sejam tomadas assertivamente e em tempo hábil. Este trabalho aborda o problema de previsão de séries temporais no contexto das curvas municipais de casos e mortes por Covid-19, utilizando um modelo de Redes Neurais de Grafos. A modelagem do problema deu-se na forma de um grafo, contemplando os 5570 municípios brasileiros e suas dependências espaciais e temporais, sendo as relações de vizinhança expressas pelas malhas viárias (rodovias, ferrovias, hidrovias e rotas aéreas), enquanto que as relações no tempo são compostas pelas curvas a serem previstas. O modelo proposto foi implementado como uma GNN espaço-temporal, e os resultados qualitativos obtidos se mostraram promissores, embora necessitem de análises quantitativas para se determinar a real eficácia do modelo na previsão dos números da pandemia.

Palavras-chave: Covid-19. Redes Neurais de Grafos. Grafo do Brasil. Previsão de casos e mortes.

ABSTRACT

Given the impact that the SARS-CoV-2 virus has had on the world, the understanding of how the contagion and death curves will evolve has great relevance, allowing decisions to contain the spread of the pandemic to be taken assertively and in a timely manner. This work proposes to address the problem of forecasting time series in the context of municipal curves of Covid-19 cases and deaths, applying a Graph Neural Network model. The modeling of the problem was carried out in the form of a graph, covering the 5570 brazilian municipalities and their spacial and temporal dependences, where the neighborhood relations are expressed by the road networks (highways, railways, waterways and air routes), while the time relations are composed by the curves to be predicted. The proposed model was implemented as a spatiotemporal GNN, and the qualitative results obtained proved to be promising, although they need quantitative analysis to determine the actual effectiveness of the model in predicting pandemic numbers.

Keywords: Covid-19. Graph Neural Networks. Brazil graph. Cases and deaths forecasting.

1 INTRODUÇÃO

A pandemia do vírus SARS-CoV-2, causador da Covid-19, fez com que muitas pesquisa fossem desenvolvidas com o intuito central de diminuir os impactos da doença na sociedade (LALMUANAWMA; HUSSAIN; CHHAKCHHUAK, 2020). Em especial, a área da epidemiologia e os estudos de propagação do vírus servem como embasamento para a correta tomada de decisões por parte dos gestores.

* Engenharia de Computação; lucasoliveira.2017@alunos.utfpr.edu.br.
† Departamento de Informática; dalcimar@utfpr.edu.br; https://orcid.org/0000-0002-1905-4602 .

Na literatura, existem diversas abordagens que se propõem a estimar a progressão das curvas temporais da Covid-19, seja quanto ao número de casos de contágio ou de mortes decorrentes do vírus. Dentre elas, os métodos estatísticos são tradicionalmente utilizados devido ao amplo embasamento teórico e à grande interpretabilidade dos resultados (e.g., Espinosa et al. (2020) e Siqueira et al. (2020)). Outra vertente de pesquisa se pauta em métodos tradicionais de aprendizado de máquina (ML), que tendem a ser mais flexíveis quanto à estruturação das estimativas (e.g., da Silva et al. (2020) e Pereira et al. (2020)).

Entretanto, esses trabalhos, assim como outros tantos divulgados desde 2020, concentram-se em antecipar as curvas de casos e/ou mortes sob uma perspectiva que generaliza macrorregiões, em geral estados ou países. Além disso, conforme pesquisas recentes de Xu et al. (2019) e Cai et al. (2019) sobre a pandemia da gripe A (H1N1), os meios de transporte exercem papel fundamental na propagação da doença, de tal forma que o aspecto espacial se torna relevante no estudo e entendimento das curvas temporais da pandemia.

E se a pandemia do novo coronavírus puder ser melhor entendida, e sua dinâmica melhor modelada, ao se considerar tanto as dependências temporais quanto as dependências espaciais, em especial os modais viários?

Este trabalho tem como objetivo abordar o problema da previsão das curvas temporais de casos e mortes por Covid-19 no Brasil utilizando Redes Neurais de Grafos (do inglês, *Graph Neural Network*, GNN). Para isso, foi gerado um *dataset* do Brasil composto pelos 5570 municípios e os modais viários que os interligam, modelado na forma de um grafo simétrico ponderado.

2 METODOLOGIA

Esta pesquisa foi conduzida conforme os seguintes passos: modelagem do *dataset*, construção da rede neural, treino e teste do modelo de GNN proposto. Os detalhes de cada uma das etapas são apresentados a seguir.

2.1 Modelagem do *Dataset*

Para representar a relação existente entre os municípios brasileiros, considerando suas características individuais e relacionais, optou-se pela modelagem do problema como um grafo, com os municípios e suas características compondo os vértice e as arestas representando as conexões intermunicipais existentes. Dada a inexistência de um *dataset* que contivesse tanto as curvas temporais quanto as características estruturais no escopo do problema, utilizou-se alguns bancos de dados em acesso público para construção do *dataset* desejado.

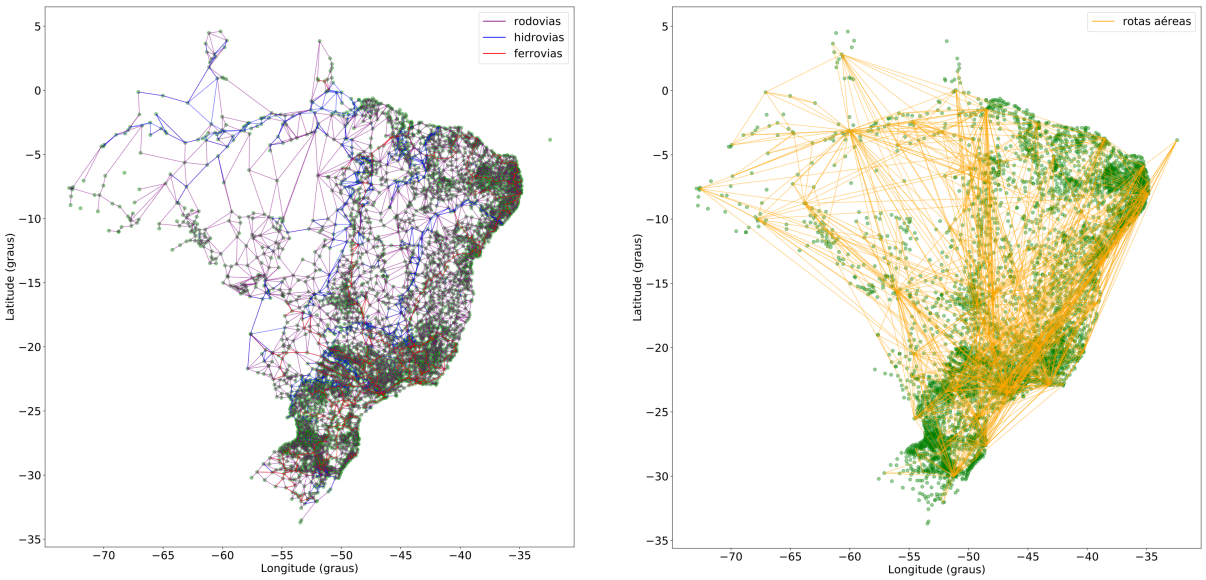
Assim, foram utilizadas duas bases de dados do Instituto Brasileiro de Geografia e Estatística (IBGE) em conjunto, referentes à mapas cartográficos (IBGE, 2019) e de do setor de logística de transportes (IBGE, 2014), que, em com a utilização do *software* QGIS, foi possível construir 4 malhas de conectividade: rodoviária; ferroviária; hidroviária e rotas aéreas, com as quais atingiu-se uma conectividade de 99,6% – apenas os municípios de Itamarati (AM) e Jordão (AC) não são alcançados.

Como pode ser visto na Fig. 1, juntas elas interligam quase todos os municípios do Brasil e permitem uma análise integrada em escala nacional.

Quanto aos vértices e seus atributos, primariamente é necessário obter os números de contágios e óbitos, os quais foram coletados no *site* Brasil.io (*https://brasil.io/dataset/covid19/files/*). Na data de coleta (25/01/2021), o Brasil encontrava-se no 336º dia da pandemia, totalizando 8.659.640 casos de contágio e 213.896 mortes.

Além disso, é sabido que uma alta concentração de pessoas acelera a disseminação do vírus (CHU et al., 2020), então foram utilizadas a densidade demográfica (em habitantes por km^2) e população estimada de cada

Figura 1 – Conectividade dos municípios brasileiros pelos modais rodoviário, ferroviário, hidroviário e transporte aéreo.



Fonte: autoria própria (2021).

município para o ano de 2020 (IBGE, 2019).

Outro fator importante é a capacidade de cuidar dos enfermos, que foi representado pelo número total de leitos hospitalares instalados até dezembro de 2019 (IBGE, G., 2020).

Também foram coletadas as coordenadas de latitude e longitude, as quais serviram apenas para o cálculo de distâncias no ponderamento das arestas e, portanto, não fazem parte do vetor de atributos municipais.

2.2 Solução Proposta

Com o intuito de abordar o problema da previsão das curvas temporais da pandemia sob a ótica das GNNs, se adotou a arquitetura de GNN espaço-temporal (WU et al., 2021), isto é, uma rede neural convolucional de grafo acoplada com uma rede recorrente (do inglês, *Recurrent Neural Network*, RNN), permitindo que o modelo lide com as dependências espaciais e temporais de forma dinâmica.

Conforme Hamilton (2020), uma GNN convolucional (que opera por passagem de mensagens) possui dois operadores fundamentais: o de agregação e o de atualização – as funções $\mathcal{F}$ e $\mathcal{G}$, respectivamente. Enquanto $\mathcal{F}$ realiza uma ponderação $\mathbf{M}_\tau$ do estado atual $\mathbf{X}_\tau$ dos vértices da vizinhança, $\mathcal{G}$ é responsável por considerar o valor ponderado por $\mathcal{F}$ em torno de cada município para atualizar o estado latente $\mathbf{Z}_\tau$ desses municípios.

$$\mathbf{M}_\tau = \mathcal{F}\left(\mathbf{X}_\tau, \mathbf{A}\right) \tag{1}$$

$$\mathbf{Z}_\tau = \mathcal{G}\left(\mathbf{X}_\tau, \mathbf{M}_\tau\right) \tag{2}$$

A matriz de estado $\mathbf{Z}_\tau$ é, então, a entrada para a RNN, a qual a partir da entrada e do estado oculto anterior $\mathbf{H}_{\tau-1}$, gera um novo estado oculto. Esse estado oculto passa por uma rede do tipo *Multi Layer Perceptron* (MLP) para ajuste de escala e a saída $\mathbf{Y}_\tau$ é a previsão para o próximo dia.

$$\mathbf{H}_\tau = \mathrm{RNN}\left(\mathbf{Z}_\tau, \mathbf{H}_{t-1}\right) \tag{3}$$

$$\mathbf{Y}_\tau = \mathrm{MLP}\left(\mathbf{H}_t\right) \tag{4}$$

2.3 Implementação Computacional

Tendo o modelo do problema, foram feitas algumas adaptações, permitindo que o algoritmo da rede neural executasse mais rápido, além de tornar o código mais legível e diminuir a complexidade de espaço referente ao carregamento dos tensores do *dataset* na memória principal.

Assim, o grafo do Brasil é representado por: uma matriz de adjacência $\mathbf{A}$, contendo as arestas ponderadas dos 4 modais viários; vetores de atributos estáticos $\mathbf{x}_s \in \mathbb{R}^3$ por município, com os valores de densidade demográfica, população absoluta e número de leitos; e a matriz de atributos dinâmicos $\mathbf{X}_d = (\mathbf{x}_d(t)) \in \mathbb{R}^{336\times 2}$ por município, contendo os dados de casos e óbitos dia-a-dia.

Dessa forma, o algoritmo de treino executa uma composição a cada dia τ, unindo $\mathbf{x}_s$ e $\mathbf{x}_d(\tau)$ na forma de um vetor expandido de atributos, conforme a Eq. (5), para cada um dos 5570 municípios.

$$\mathbf{x}_\tau = [\mathbf{x}_s \mid \mathbf{x}_d(\tau)] \in \mathbb{R}^5 \tag{5}$$

Adicionando a notação sobrescrita $[k]$ para se referir ao k-ésimo município, é possível definir matriz de entrada do algoritmo (Eq. (6)), para um instante τ, que corresponde ao estado latente de cada vértice antes da etapa de agregação.

$$\mathbf{X}_\tau = \left(\mathbf{x}_\tau^{[k]}\right) \in \mathbb{R}^{5570\times 5} \tag{6}$$

Para o modelo de GNN, utilizou-se módulos *Gated Recurrent Unit* (GRU) tanto para a função de atualização (operador $\mathcal{G}$) quando para compor a rede de recorrência, e a convolução é feita com 1-*hop*.

A implementação da rede neural utilizou as bibliotecas PyTorch e PyTorch Geometric versão 1.8.1, e a validação cruzada foi feita separando o conjunto de dados no eixo do tempo, em grupos sequenciais de treino, validação e teste.

O treino foi realizado para previsão de 1 dia a frente, isto é, são informados t dias para a rede neural, no formato do tensor $(\mathbf{X}_1 \mid \ldots \mid \mathbf{X}_t) \in \mathbb{R}^{5570\times 5\times t}$, e se avalia o resultado previsto para o dia $t+1$.

3 RESULTADOS

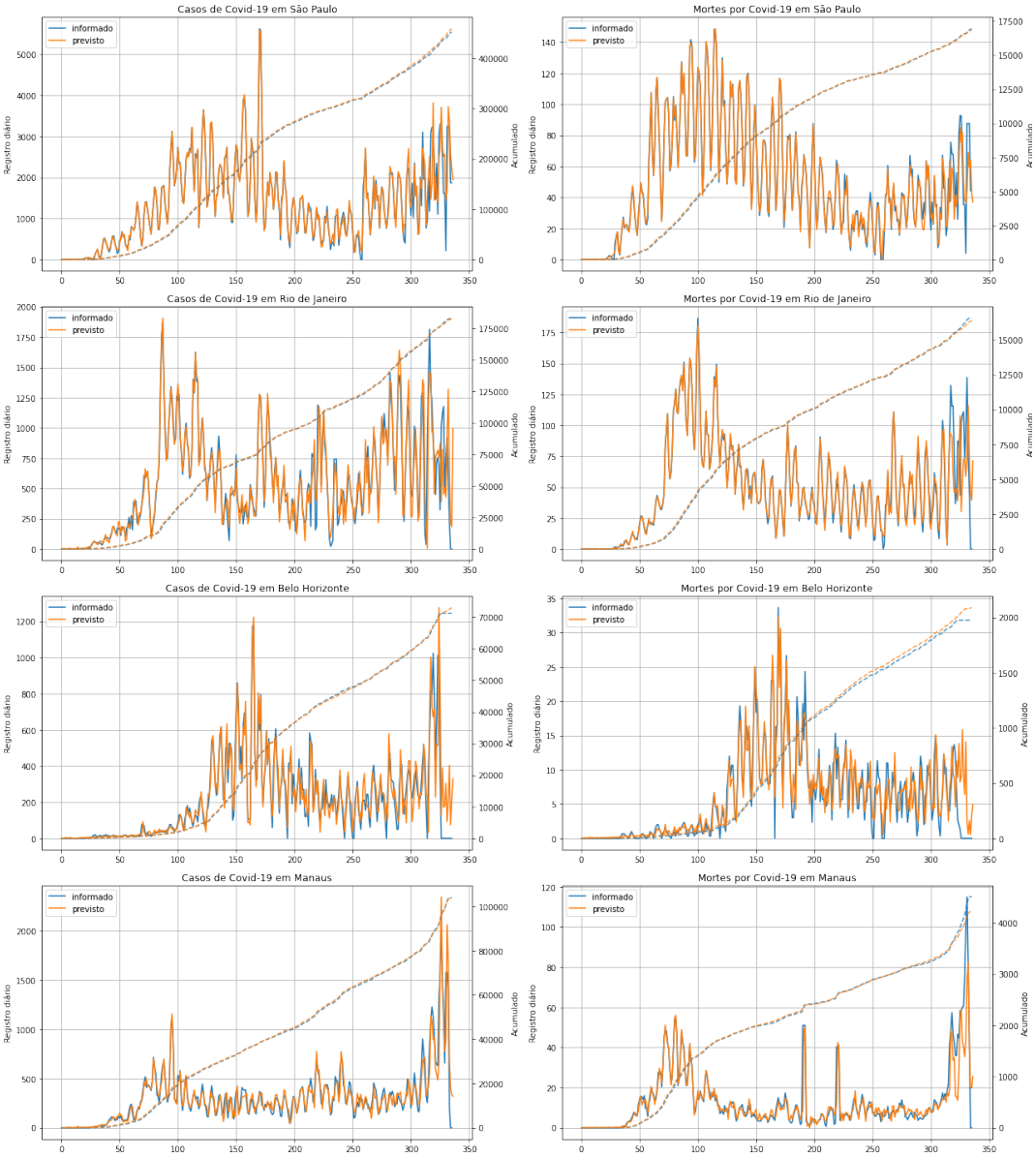
Ao examinar as curvas previstas e as curvas reais esperadas na Fig. 2, é possível notar uma boa qualidade na aproximação dos números diários, demonstrando que o modelo foi capaz de incorporar a dinâmica espacial e temporal da pandemia.

4 CONCLUSÕES

Experimentar uma abordagem por redes neurais de grafos, em contrapartida aos métodos estatísticos e de ML tradicionais costumeiramente encontrados na literatura, mostrou-se extremamente promissor.

Uma vez que a GNN implementada é um modelo único, ela permite a troca de informações entre as realidades individuais vivenciadas em cada município, ao passo que permite perguntas e análises não só quanto à evolução

Figura 2 – previsão diária de casos e mortes para capitais estaduais.



Fonte: autoria própria (2021).

temporal da pandemia, mas também quanto à sua propagação espacial.

Os resultados obtidos neste trabalho evidenciam que tanto o modelo quando a premissa de utilizar os modais viários como indicador de conectividade intermunicipal permitem uma boa modelagem da dinâmica da pandemia e servem de motivação para que mais pesquisas sejam feitas nesta área.

Entretanto, ainda que as curvas previstas aparentem ter boa acurácia, é preciso avaliar a qualidade do modelo com métricas que sejam adequadas ao problema e ao *dataset*, além de comparar o modelo com *baselines* para obter um melhor julgamento na interpretação dos resultados.

Como trabalhos futuros, espera-se avaliar a qualidade do modelo proposto de forma mais rigorosa, com uso de métricas e *baselines*, e também avaliar o modelo quanto a sua capacidade de prever os números da pandemia em um horizonte de previsão maior que 1 dia.

AGRADECIMENTOS

Agradeço à Fundação Araucária pelo subsídio, que permitiu uma maior dedicação à pesquisa e contribuiu para o desenvolvimento da ciência nacional.

REFERÊNCIAS

CAI, Jun et al. Roles of Different Transport Modes in the Spatial Spread of the 2009 Influenza A (H1N1) Pandemic in Mainland China. **International Journal of Environmental Research and Public Health**, v. 16, n. 2, 2019. ISSN 1660-4601. DOI: 10.3390/ijerph16020222. Disponível em: .

CHU, Derek K et al. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. **The Lancet**, v. 395, n. 10242, p. 1973–1987, 2020. ISSN 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(20)31142-9. Disponível em: .

DA SILVA, Ramon Gomes et al. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. **Chaos, Solitons & Fractals**, v. 139, p. 110027, 2020. ISSN 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110027. Disponível em: .

ESPINOSA, Mariano Martinez et al. Predição de casos e óbitos de COVID-19 em Mato Grosso e no Brasil. **Journal of Health & Biological Sciences**, v. 8, n. 1, p. 1–7, 2020.

HAMILTON, William L. Graph representation learning. **Synthesis Lectures on Artifical Intelligence and Machine Learning**, Morgan & Claypool Publishers, v. 14, n. 3, p. 1–159, 2020.

IBGE. **Bases Cartográficas Contínuas – Brasil**. [S.l.: s.n.], 2019. Disponível em: . Acesso em: 16 dez. 2020.

IBGE. **Logística dos Transportes – Brasil**. [S.l.: s.n.], 2014. Disponível em: . Acesso em: 26 nov. 2020.

IBGE, GeoHub. **Leitos totais em 2019**. [S.l.: s.n.], 2020. Disponível em: . Acesso em: 25 jan. 2021.

LALMUANAWMA, Samuel; HUSSAIN, Jamal; CHHAKCHHUAK, Lalrinfela. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. **Chaos, Solitons & Fractals**, v. 139, p. 110059, 2020. ISSN 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110059. Disponível em: .

PEREIRA, Igor Gadelha et al. Forecasting COVID-19 Dynamics in Brazil: A Data Driven Approach. **International Journal of Environmental Research and Public Health**, MDPI AG, v. 17, n. 14, p. 5115, jul. 2020. ISSN 1660-4601. DOI: 10.3390/ijerph17145115. Disponível em: .

SIQUEIRA, Elton et al. Temporal Prediction Model of the Evolution of Confirmed Cases of the New Coronavirus (SARS-CoV-2) in Brazil. **IEEE Latin America Transactions**, v. 100, 1e, 2020.

WU, Zonghan et al. A Comprehensive Survey on Graph Neural Networks. **IEEE Transactions on Neural Networks and Learning Systems**, v. 32, n. 1, p. 4–24, 2021. DOI: 10.1109/TNNLS.2020.2978386.

XU, Bo et al. Impacts of Road Traffic Network and Socioeconomic Factors on the Diffusion of 2009 Pandemic Influenza A (H1N1) in Mainland China. **International Journal of Environmental Research and Public Health**, v. 16, n. 7, 2019. ISSN 1660-4601. DOI: 10.3390/ijerph16071223. Disponível em: .