Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 19, Iss. 1, pp. 33-44, 2021

SOBRE O PROBLEMA DE PREVISÃO DA COVID-19 UTILIZANDO MODELOS MORFOLÓGICO-LINEAR PROFUNDOS

Ricardo de A. Araújo
[1]Laboratório de Inteligência Computacional do Araripe
Instituto Federal do Sertão Pernambucano, Ouricuri, PE, Brasil.
ricardo.araujo@ifsertao-pe.edu.br

Resumo – A doença do coronavírus 2019 (COVID-19) foi declarada pela Organizaçao Mundial de Saúde (*World Health Organization*, WHO) como uma pandemia sem precedentes, tendo sobrecarregado os sistemas de saúde devido à alta demanda de internações em unidades de terapia intensiva. Neste contexto, estimar a dinâmica da pandemia da COVID19 é essencial para lidar com as limitações do sistema de saúde. Portanto, neste trabalho desenvolvemos um estudo empírico sobre séries temporais relacionadas a pandemia da COVID-19 e, baseado neste estudo, apresentamos um modelo morfológico-linear profundo, treinado por um processo de aprendizagem baseado em gradiente descendente, capaz de prever este tipo particular de série temporal. Para avaliar o desempenho preditivo do modelo proposto, foram utilizadas séries temporais da COVID-19, com frequência diária, no Brasil e Estados Unidos da América. Os resultados alcançados mostram que modelo proposto supera os modelos de aprendizagem de máquina clássicos e recentemente apresentados na literatura para estimar a dinâmica da pandemia da COVID-19.

Palavras-chave – COVID-19, Séries Temporais, Previsão, Modelo Morfológico-Linear, Aprendizagem Profunda.

Abstract – The coronavirus disease 2019 (COVID-19) has been declared by the World Health Organization (WHO) as an unprecedented pandemic in the present days, straining healthcare systems due to the high demand for admissions to intensive care units. In this context, estimating the dynamics of the COVID19 pandemic is essential to deal with health system drawbacks. Therefore, in this work we developed an empirical study on time series related to the COVID-19 pandemic and, based on this study, we present a deep morphological-linear model, trained by a gradient-based learning process, able to predict this particular kind of time series. Trying to assess the predictive performance of the proposed model, we use daily COVID-19 time series in Brazil and United States of America. The achieved results show that the proposed model outperforms classical and recent machine learning models to estimate the dynamics of the COVID-19 pandemic.

Keywords – COVID-19, Time Series, Prediction, Morphological-Linear model, Deep Learning.

1. INTRODUÇÃO

Nos dias atuais, o controle de doenças virais ainda é considerado um grande desafio para a saúde pública mundial [1]. O coronavírus é um vírus de ácido ribonucléico envelopado que tem sido responsável pelo surgimento de diversas doenças no aparelho respiratório humano [2]. Devido a prevalência, a constante recombinação de seus genes e a alta diversidade genética, surgiu em Wuhan - Hubei (China), em meados de 2019, um novo coronavírus, conhecido como *severe acute respiratory syndrome coronavirus 2* (SARS-Cov-2) [3, 4], originado de morcegos [5, 6] e sendo o agente causador da doença do coronavírus 2019 (COVID-19) [3, 7–9].

Devido a sua disseminação acelerada, a COVID-19 foi considerada como uma pandemia em Março de 2019 pela Organização Mundial de Saúde [10], atingindo todo o globo terrestre e tendo desencadeado uma crise catastrófica sem precedentes, bem como representando uma ameaça real para a saúde e economia globais [11–13]. Na maioria dos casos confirmados os sintomas são leves e sem complicações [10, 14].

No entanto, devido ao surgimento de variantes da COVID-19 associadas a complicações moderadas em cerca de 15-20% dos pacientes (tendo a necessidade de internação com suporte de oxigênio) e graves em torno de 5-12% dos pacientes (tendo urgência para internação em unidade de terapia intensiva), houve uma sobrecarga além do esperado nos sistemas de saúde mundiais [15–17]. Para evitar o colapso de seus sistemas de saúde, chefes de estado empregaram medidas de quarentena, fechamento de escolas, distanciamento social, uso de máscara e restrição de viagens visando retardar a disseminação acelerada da COVID-19 [18–21].

Tais medidas provocaram a paralisação de diversas atividades econômicas, tendo impacto negativo no produto interno bruto (PIB) e na manutenção de empregos em todo o mundo [12]. Neste sentido, estimar a dinâmica da pandemia da COVID-19 é indispensável para nortear ações estratégicas para lidar com o aumento crescente da demanda por serviços de saúde e os efeitos diretos e indiretos na economia [22–24]. Logo, esforços ainda precisam ser aplicados no desenvolvimento de estudos sobre o fenômeno gerador de séries temporais da COVID-19 e, consequentemente, modelos para predizê-los.

Desta forma, na tentativa de encontrar respostas a questões sobre a dinâmica da COVID-19, apresentamos um estudo sobre séries temporais relacionadas ao seu fenômeno gerador. Baseado neste estudo, é investigado um modelo de rede neural crescente-

decrescente-linear profunda, projetado a partir de um processo de aprendizagem baseado em gradiente descendente, para produzir um mapeamento capaz de estimar a dinâmica deste tipo particular de série temporal.

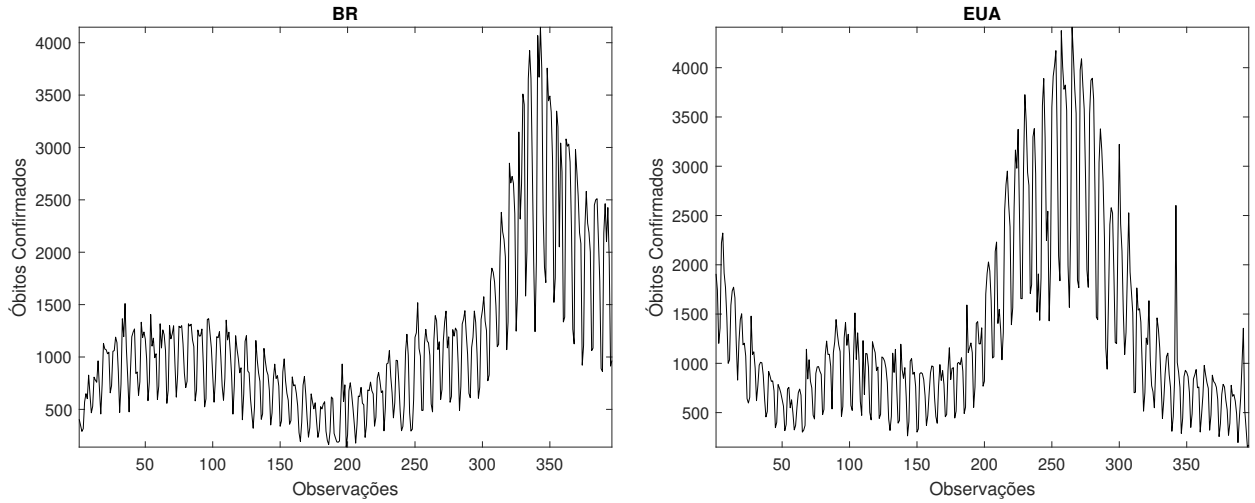
Séries temporais com frequência diária relacionadas a óbitos confirmados pela COVID-19 em dois países (Brasil e Estados Unidos da América) são utilizadas em nossa análise experimental. O erro médio absoluto (*mean absolute error*, MAE), o erro médio absoluto percentual (*mean absolute percentage error*, MAPE) e o erro médio quadrático (*mean squared error*, MSE) são utilizados para avaliar o desempenho preditivo, que é validado estatisticamente utilizando os testes de Friedman e Tukey. Os resultados alcançados demonstram desempenho superior do modelo proposto quando comparado a modelos clássicos e recentemente apresentados na literatura.

Este trabalho é organizado da seguinte forma. A Seção 2 apresenta uma análise das séries temporais investigadas. Na Seção 3, é descrito o modelo proposto. Posteriormente, na Seção 4, as simulações e os resultados experimentais são apresentados. Por fim, na Seção 5, as conclusões e as direções futuras deste trabalhos são descritas.

2 ANÁLISE DAS SÉRIES TEMPORAIS

As séries temporais de óbitos confirmados no Brasil (BR) e Estados Unidos da América (EUA) são investigadas neste trabalho, uma vez que estes são os países onde houveram a maior quantidade de óbitos reportados por COVID-19. Cada série temporal analisada consiste de observações diárias no período de 01/05/2020 a 31/05/2021, como apresentado na Figura 1.

Figura 1: Gráfico das séries temporais investigadas.



Fonte: Elaborada pelo autor.

A seguir, apresentamos uma análise empírica para avaliar as características do fenômeno gerador das séries temporais investigadas. Inicialmente, investigamos a função de autocorrelação (*autocorrelation function*, ACF) [25] e a função de autocorrelação parcial (*partial autocorrelation function*, PACF) [25], ilustradas nas Figuras 2 e 3, para analisar a ocorrência de dependência linear.

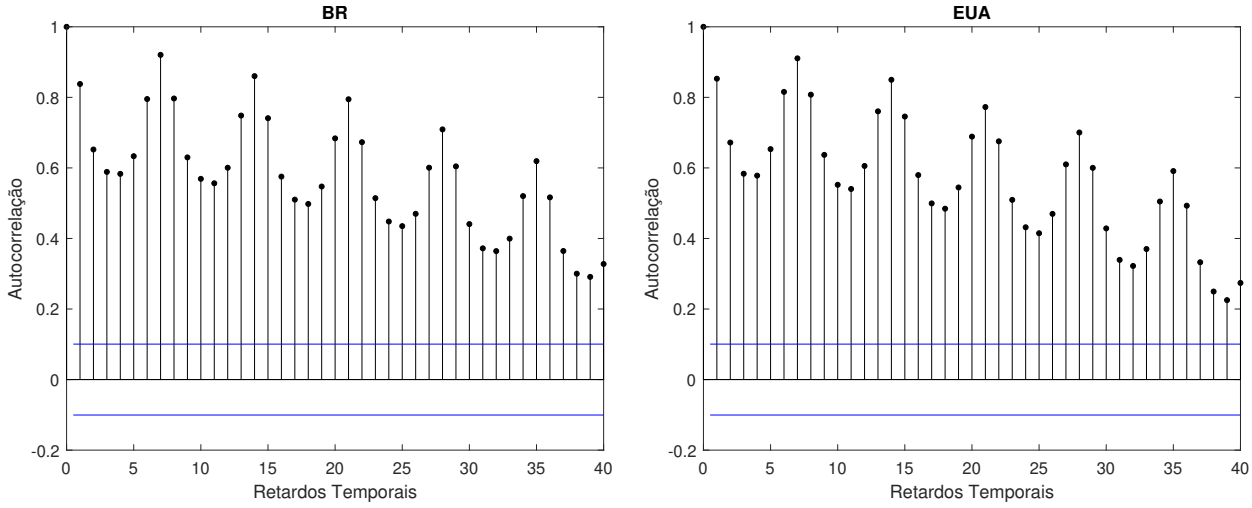
Na Figura 2 é possível verificar na ACF um decaimento lento com ciclos bem definidos e regulares (sazonalidade), sugerindo a presença de dependência linear e não linear. Na Figura 3 é possível verificar os retardos temporais 1, 2, 3, 5, 6, 7, 8, 11, 14, 15, 16, 17, 18, 21, 22, 26, 28, 32 e 34 (para a série BR) e os retardos temporais 1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 15, 16, 23, 28, 30 e 33 (para a série EUA) como significantes para caracterização do fenômeno gerador das séries investigadas. No entanto, a análise da ACF e a PACF não possibilita uma correta avaliação da dependência não linear. Assim, para superar esta limitação, é apresentada a informação mútua média (*mean mutual information*, MMI) [26, 27], ilustrada na Figura 4.

De acordo com a Figura 4, é também possível observar um comportamento de decaimento lento com ciclos regulares, sugerindo a presença de dependência não linear. No entanto, a MMI não é capaz de avaliar a natureza de tal dependência e, portanto, utilizamos o parâmetro de Hurst (*Hurst parameter*, HP) [28–30], ilustrado na Figura 5, para superar tal limitação. A Figura 5 revela que as séries temporais investigadas são geradas por processos auto-similares com dependência não linear de longo prazo (os valores do HP são próximos de 0), sugerindo que, teoricamente, estas podem ser previstas.

3 O MODELO PROPOSTO

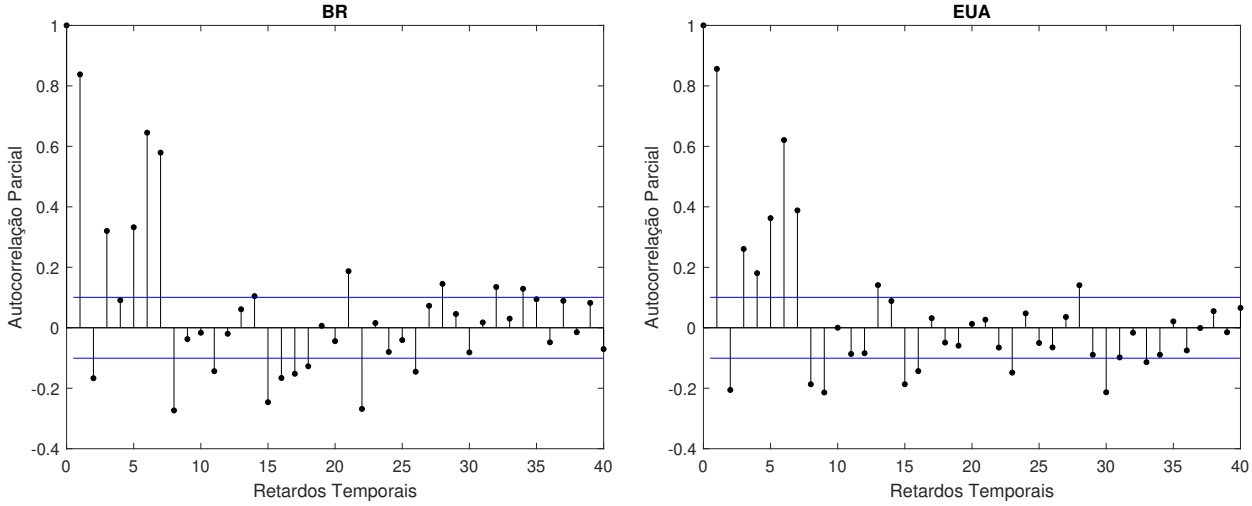
A motivação deste trabalho está relacionada ao fato que modelos epidemiológicos não obtiveram sucesso na previsão da pandemia da COVID-19 [31] devido a dependência humana para a definição de seus parâmetros e, consequentemente, comprometendo o projeto automatizado de um modelo com desempenho minimamente aceitável [32]. Ao contrário dos modelos

Figura 2: Gráfico da função de autocorrelação das séries temporais investigadas.



Fonte: Elaborada pelo autor.

Figura 3: Gráfico da função de autocorrelação parcial das séries temporais investigadas.



Fonte: Elaborada pelo autor.

epidemiológicos clássicos para prever a disseminação de vírus [31, 33–39], este trabalho emprega uma abordagem baseada em séries temporais com uma tecnologia de aprendizagem profunda para estimar cenários futuros da pandemia da COVID-19.

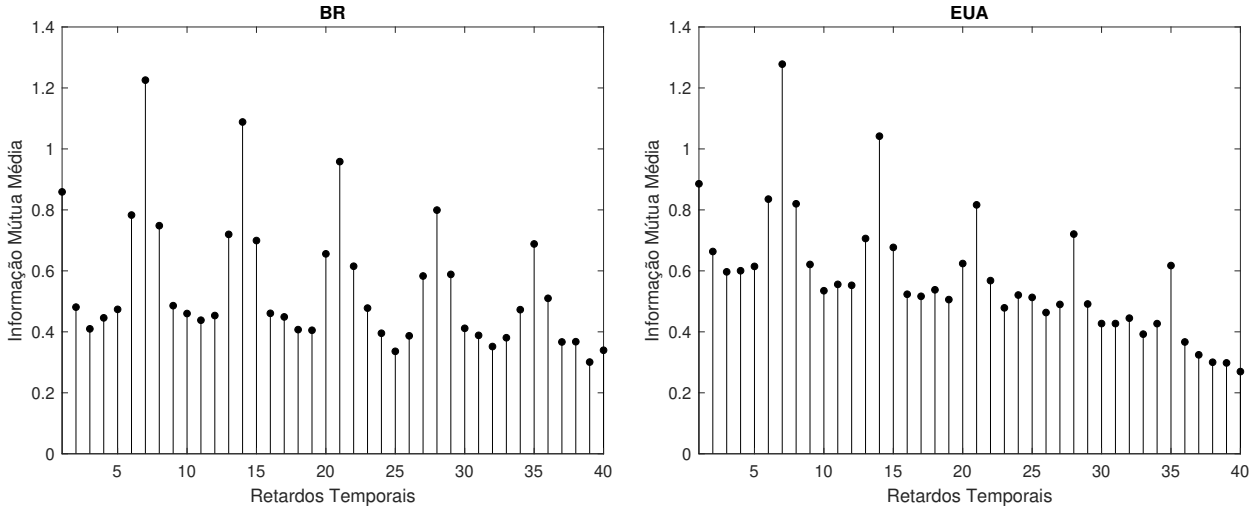
De acordo com Araújo [40], é possível utilizar operadores crescente-decrescente-linear (increasing-decreasing-linear, IDL) para prever séries temporais geradas por processos auto-similares com dependência não linear de longo prazo (comportamento anti-persistente). Assim, este trabalho apresenta uma arquitetura de rede neural profunda baseada em operadores IDL, referida como *deep increasing-decreasing-linear neural network* (DIDLNN), para prever séries temporais da COVID-19. A Figura 6 apresenta a arquitetura do modelo proposto.

A camada de entrada do modelo é composta por d retardos temporais da série (assumindo que x_t é a observação da série temporal (x) no tempo t, então os retardos temporais serão $x_{t-1}, x_{t-2}, ..., x_{t-d}$). As H camadas a seguir (conhecidas como escondidas) do modelo são compostas por: i) técnica de regularização por *dropout* [41] (as unidades de processamento neurais são inativadas aleatoriamente, utilizando uma taxa de 10%, durante o processo de treinamento, reduzindo o impacto do problema de *overfitting*), ii) k unidades IDL, e iii) unidades sigmoides (uma função de ativação aplicada a saída das unidades IDL). Por fim, a camada de saída é composta por: i) uma unidade IDL, e ii) uma unidade sigmoide. Note que é utilizada apenas uma unidade de processamento nesta camada porque este trabalho foca no problema de previsão do tipo um-passo-adiante.

A n-ésima saída da unidade de processamento IDL da l-ésima camada da DIDLNN é dada por

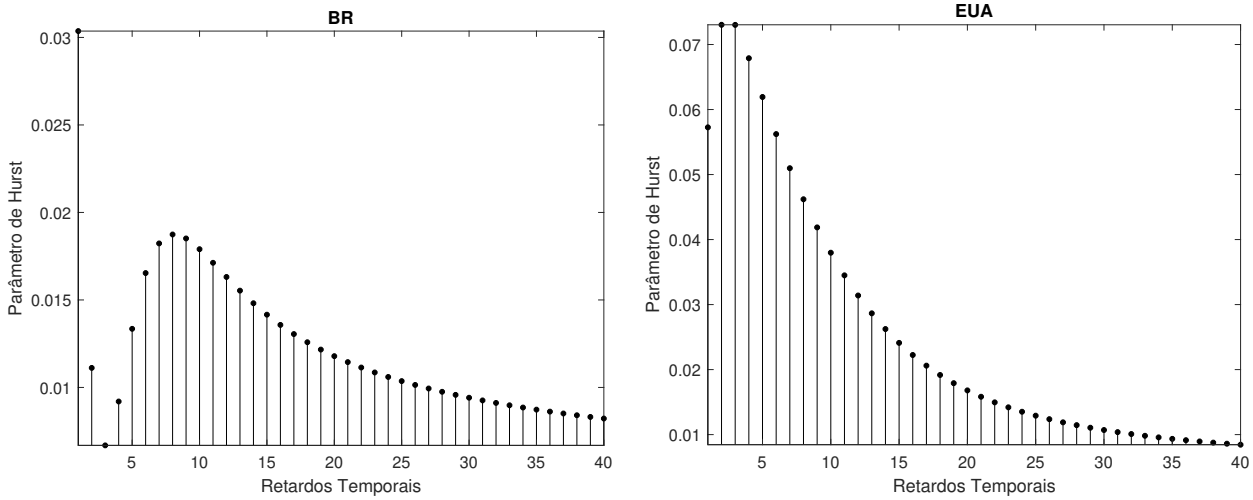
$$y_n^{(l)} = f\left(u_n^{(l)}\right), \qquad n = 1, \ldots, N_l, \tag{1}$$

Figura 4: Gráfico da informação mútua média das séries temporais investigadas.



Fonte: Elaborada pelo autor.

Figura 5: Gráfico do parâmetro de Hurst das séries temporais investigadas.



Fonte: Elaborada pelo autor.

em que

$$u_n^{(l)} = \lambda_n^{(l)} \alpha_n^{(l)} + \left(1 - \lambda_n^{(l)}\right) \beta_n^{(l)}, \qquad \lambda_n^{(l)} \in [0,1], \tag{2}$$

com

$$\beta_n^{(l)} = \sum_{i=1}^{N_{l-1}} y_i^{(l-1)} p_{n,i}^{(l)} + \rho_n^{(l)}, \tag{3}$$

e

$$\alpha_n^{(l)} = \theta_n^{(l)} \tau_n^{(l)} + \left(1 - \theta_n^{(l)}\right) \kappa_n^{(l)}, \qquad \theta_n^{(l)} \in [0,1], \tag{4}$$

onde

$$\tau_n^{(l)} = \varphi_n^{(l)} \delta_n^{(l)} + \left(1 - \varphi_n^{(l)}\right) \varepsilon_n^{(l)}, \qquad \varphi_n^{(l)} \in [0,1], \tag{5}$$

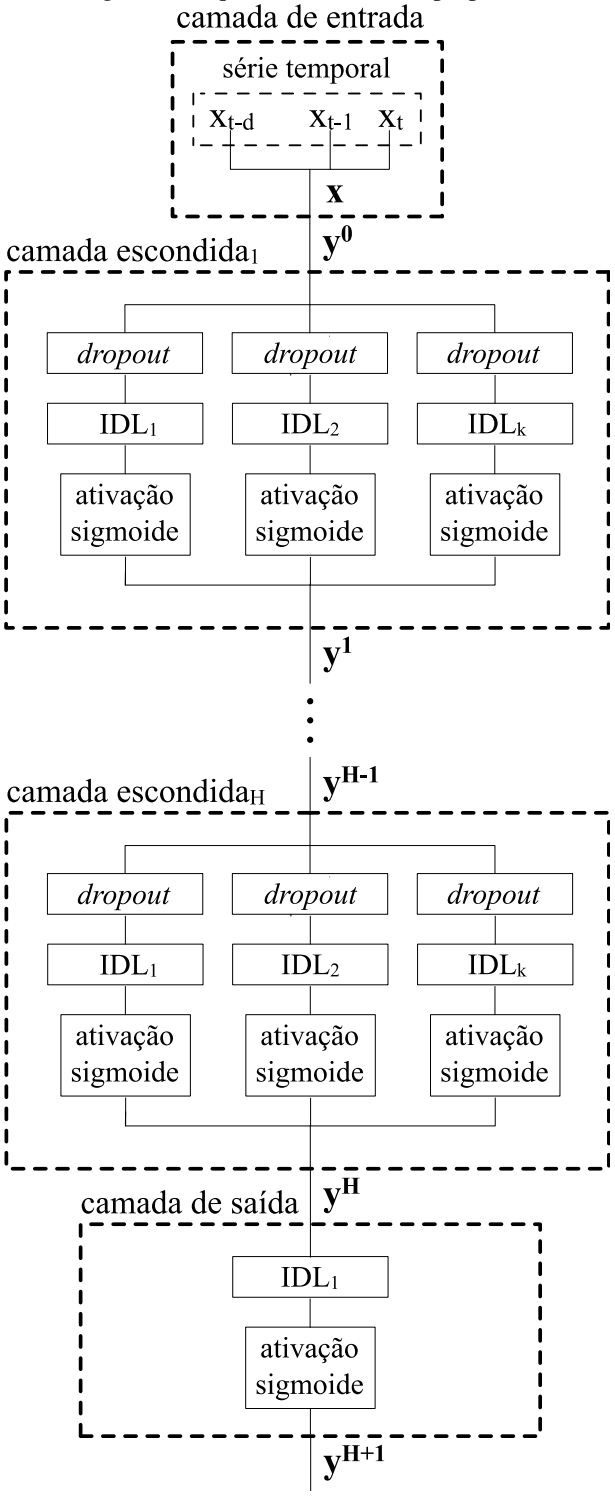
e

$$\kappa_n^{(l)} = \omega_n^{(l)} \overline{\delta}_n^{(l)} + \left(1 - \omega_n^{(l)}\right) \overline{\varepsilon}_n^{(l)}, \qquad \omega_n^{(l)} \in [0,1], \tag{6}$$

com

$$\delta_n^{(l)} = \delta_{\mathbf{a}_n^{(l)}} \left(\mathbf{y}^{(l-1)}\right) = \bigvee_{i=1}^{N_{l-1}} \left(y_i^{(l-1)} + a_{n,i}^{(l)}\right), \tag{7}$$

Figura 6: Arquitetura do modelo proposto.



Fonte: Elaborada pelo autor.

$$\varepsilon_n^{(l)} = \varepsilon_{\mathbf{b}_n^{(l)}}\left(\mathbf{y}^{(l-1)}\right) = \bigwedge_{i=1}^{N_{l-1}} \left(y_i^{(l-1)} +' b_{n,i}^{(l)}\right), \tag{8}$$

$$\overline{\delta}_n^{(l)} = \overline{\delta}_{\mathbf{c}_n^{(l)}}\left(\mathbf{y}^{(l-1)}\right) = \bigwedge_{i=1}^{N_{l-1}} \left[\left(y_i^{(l-1)}\right)^* +' c_{n,i}^{(l)}\right], \tag{9}$$

$$\overline{\varepsilon}_n^{(l)} = \overline{\varepsilon}_{\mathbf{d}_n^{(l)}}\left(\mathbf{y}^{(l-1)}\right) = \bigvee_{i=1}^{N_{l-1}} \left[\left(y_i^{(l-1)}\right)^* + d_{n,i}^{(l)}\right]. \tag{10}$$

O termo N_l denota a quantidade de unidades de processamento na l-ésima camada do modelo. Os parâmetros $\lambda_n^{(l)}$, $\theta_n^{(l)}$, $\varphi_n^{(l)}$, $\omega_n^{(l)}$, $\rho_n^{(l)} \in [0,1]$, e os parâmetros $\mathbf{a}_n^{(l)}, \mathbf{b}_n^{(l)}, \mathbf{c}_n^{(l)}, \mathbf{d}_n^{(l)}, \mathbf{p}_n^{(l)} \in \mathbb{R}^{N_{l-1}}$. Note que a saída de cada unidade de processamento IDL emprega uma função de ativação sigmoide, dada por

$$f\left(u_n^{(l)}\right) = \frac{1}{1+\exp\left(-u_n^{(l)}\right)}. \tag{11}$$

A ativação interna ($u_n^{(l)}$) da unidade de processamento IDL é composta por uma combinação (o termo de mistura é dado por $\lambda_n^{(l)}$) entre $\beta_n^{(l)}$ e $\alpha_n^{(l)}$, que representam os módulos linear e não linear, respectivamente, da unidade de processamento. O módulo linear é composto pelo *perceptron* clássico. O módulo não linear é composto pelo operador morfológico crescente-decrescente (*increasing-decreasing*, ID), que é definido em termos de uma combinação (o termo de mistura é dado por $\theta_n^{(l)}$) entre o módulo crescente ($\tau_n^{(l)}$) e o módulo decrescente ($\kappa_n^{(l)}$). O módulo crescente é composto por uma combinação (o termo de mistura é dado por $\varphi_n^{(l)}$) entre um operador de dilatação ($\delta_n^{(l)}$) e um operador de erosão ($\varepsilon_n^{(l)}$). O módulo decrescente é composto por outra combinação (o termo de mistura é dado por $\omega_n^{(l)}$) entre um operador de anti-dilatação ($\overline{\delta}_n^{(l)}$) e um operador de anti-erosão ($\overline{\varepsilon}_n^{(l)}$).

No processo de treinamento do modelo proposto os pesos são ajustados de acordo com um critério de erro (função objetivo) até a convergência, isto é, até o critério de parada ser alcançado. Vale mencionar que a função objetivo gera uma superfície de erro que reside no espaço $\mathbb{R}^W$ (W representa a quantidade de parâmetros do modelo). O principal problema na minimização da função objetivo é encontrar um ponto de ótimo neste espaço que minimiza o erro entre a saída gerada pelo modelo e a saída desejada. Como o processo de aprendizagem do modelo proposto é do tipo supervisionado, este será equivalente a um problema de otimização não linear irrestrito, onde a superfície de erro será minimizada a partir do ajuste iterativo do vetor de pesos. Para o processo de aprendizagem do modelo proposto utilizamos o algoritmo de retropropagação do erro [40].

4 SIMULAÇÕES E RESULTADOS EXPERIMENTAIS

Nesta Seção focamos em descrever a configuração e os resultados das simulações realizadas para avaliar o modelo proposto quando aplicado ao problema de previsão de séries temporais relacionadas a COVID-19. Neste contexto, para todas as simulações realizadas, normalizamos todas as séries temporais no intervalo $[0.2, 0.8]$ e dividimos os dados em dois conjuntos: as últimas 31 observações (de 01/05/2021 a 31/05/2021) para o conjunto de teste e as observações restantes para o conjunto de treinamento.

Comparamos os resultados do modelo proposto com os obtidos por modelos clássicos e recentemente apresentados na literatura: i) modelos estatísticos (autoregressivo integrado de médias móveis - ARIMA [42], arvores de regressão - REGTREE [43] e regressão de vetor de suporte - SVR [44]), ii) modelos de redes neurais (*perceptron* multicamadas - MLP [45], função de base radial - RBF [46], memória de curto e longo prazo - LSTM [47], convolucional - CNN [48]), e iii) modelos dinâmicos (rede neural autoregressiva não linear com entradas exógenas - NARX [49]).

Para os experimentos com os modelos ARIMA e REGTREE, empregamos a metodologia sugerida em [42] e [43], respectivamente, utilizando a *econometrics toolbox* do *MATLAB*. Para os experimentos com o modelo SVR, definimos o *kernel* como função de base radial e empregamos a metodologia sugerida por [44], para determinar Γ, $cost$ e $tolerance$, utilizando a biblioteca *scikit-learn* do *Python*.

Para os experimentos com os modelos MLP, RBF e NARX, empregamos os procedimentos apresentados em [45], [46] e [49], respectivamente, para determinar seus parâmetros, utilizando a *neural networks toolbox* do *MATLAB*. Para os experimentos com os modelos LSTM e CNN, empregamos o procedimento descrito em [47] e [48], respectivamente, para determinar seus parâmetros, utilizando a biblioteca *keras* do *Python*.

Para os experimentos com o modelo proposto, definimos uma arquitetura dada por DIDLNN (I;$\mathbf{H} = h_1, \ldots, h_k$;$O$;$\mu$;$\sigma$). O termo I representa a dimensionalidade de entrada, o termo h_k define a quantidade de unidades de processamento na k-ésima camada escondida, o termo O representa a dimensionalidade de saída (como este trabalho foca em problemas de previsão de um-passo-adiante, definimos O=1), o termo μ representa a taxa de aprendizagem e o termo σ representa o fator de suavização.

Para a quantidade de camadas escondidas (k), empregamos uma metodologia empírica, utilizando a validação cruzada, onde os valores 1, 3 e 5 são investigados. A Tabela 1 resume os valores investigados para cada h_k.

Tabela 1: Valores investigados para cada h_k.

k	h_1	h_2	h_3	h_4	h_5
1	10	-	-	-	-
3	100	50	10	-	-
8	100	100	50	50	10

Vale mencionar que o melhor valor encontrado para k é 3. Para a taxa de aprendizagem (μ), empregamos uma metodologia empírica, utilizando validação cruzada, onde os valores 0.001, 0.01 e 0.1 são investigados. Neste contexto, o melhor valor encontrado para μ é 0.01. Para o fator de suavização (σ), também foi empregado uma metodologia empírica, utilizando validação cruzada, onde os valores 0.005, 0.05 e 0.5 são investigados. Note que o melhor valor encontrado para σ é 0.05. A escolha da dimensionalidade de entrada (I) foi baseada na análise das séries temporais apresentada na Seção 2, onde foram utilizados os

retardos temporais 1, 2, 3, 5, 6, 7, 8, 11, 14, 15, 16, 17, 18, 21, 22, 26, 28, 32 e 34 (para a série BR) e os retardos temporais 1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 15, 16, 23, 28, 30 e 33 (para a série EUA). Por fim, o modelo proposto foi ajustado por 10000 épocas de treinamento.

Três medidas relevantes são utilizadas para avaliação do desempenho preditivo no contexto de previsão de séries temporais da COVID-19: erro médio absoluto (*mean absolute error*, MAE), erro médio absoluto percentual (*mean absolute percentage error*, MAPE) e erro médio quadrático (*mean squared error*, MSE), respectivamente definidos por

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_t - \hat{x}_t|, \tag{12}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_t - \hat{x}_t|}{x_t}, \tag{13}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_t - \hat{x}_t)^2, \tag{14}$$

em que N representa a quantidade de pontos da série temporal, e x_t e $\hat{x}_t$ são os valores reais e previstos, respectivamente, para o t-ézimo ponto da série temporal.

Note que, para cada configuração, realizamos trinta experimentos, onde são calculados a média (MEAN) e o desvio padrão (STD). Também, aplicamos o teste de Friedman [50] com nível de significância $\alpha = 0.05$, uma vez que este estabelece um *rank* de desempenho para os modelos investigados. Além disso, utilizamos um teste *post hoc*, conhecido como teste de Tukey [51] com $\alpha = 0.05$, para avaliar o desempenho par a par dos modelos investigados. Vale mencionar que ambos os testes são consideradas todas as séries temporais em conjunto, de acordo com a metodologia apresentada por Demsar [52].

4.1 Análise das Medidas de Desempenho

A Tabela 2 resume os resultados obtidos para as séries temporais BR e EUA, apresentando as estatísticas MEAN e STD, para as medidas MAE, MAPE e MSE. Vale a pena mencionar que, independente da medida de desempenho analisada, o modelo proposto neste trabalho alcançou o melhor desempenho dentre os modelos investigados. Note que o MAE alcancado pelo modelo proposto no intervalo $[1.9e-02, 3.2e-02]$ indica que o valor absoluto da diferença entre o valor previsto e o valor real está relativamente baixo. Além disso, podemos abservar que os valores obtidos com o modelo proposto para a medida MAPE no intervalo $[7.5e-02, 8.2e-02]$ sugerem que a previsão tem desvios percentuais significativamente baixos. Por fim, verificamos que os valores alcançados pelo modelo proposto para a medida MSE no intervalo $[6.8e-04, 2.3e-03]$ indicam que as previsões estão bastante próximas dos valores reais das séries investigadas.

Tabela 2: Desempenho preditivo do conjunto de teste para as medidas MAE, MAPE e MSE.

	MAE		MAPE		MSE	
Modelo	BR	EUA	BR	EUA	BR	EUA
ARIMA [42]	0.1097 ±0.0000	4.8655e-02 ±0.0000	0.2437 ±0.0000	0.2007 ±0.0000	1.9320e-02 ±0.0000	3.6788e-03 ±0.0000
CNN [48]	7.3158e-02 ±3.2320e-02	2.7210e-02 ±3.0575e-03	0.1617 ±7.4057e-02	0.1015 ±1.2510e-02	8.5947e-03 ±6.0796e-03	1.4044e-03 ±3.6180e-04
DIDLNN	**3.7886e-02** ±2.1017e-03	**1.9198e-02** ±2.3210e-03	**8.2017e-02** ±3.3199e-03	**7.5059e-02** ±1.0798e-02	**2.3929e-03** ±2.0227e-04	**6.8687e-04** ±1.0187e-04
LSTM [47]	6.7615e-02 ±2.1691e-02	2.1849e-02 ±4.4319e-03	0.1509 ±4.4201e-02	8.1534e-02 ±1.6333e-02	7.1806e-03 ±3.9812e-03	8.2749e-04 ±2.3035e-04
MLP [45]	3.9796e-02 ±2.6945e-06	2.7386e-02 ±1.9234e-06	8.8177e-02 ±6.3641e-06	0.1090 ±8.5830e-06	2.4089e-03 ±3.9560e-07	1.3164e-03 ±1.6909e-07
NARX [49]	5.9036e-02 ±7.0472e-03	3.0348e-02 ±1.1499e-02	0.1283 ±1.8193e-02	0.1181 ±4.6300e-02	5.4962e-03 ±1.4629e-03	1.4288e-03 ±8.1062e-04
RBF [46]	0.1563 ±0.0000	2.6112e-02 ±0.0000	0.3245 ±0.0000	9.9056e-02 ±0.0000	3.7363e-02 ±0.0000	1.1868e-03 ±0.0000
REGTREE [43]	9.8182e-02 ±0.0000	2.8865e-02 ±0.0000	0.2042 ±0.0000	0.1126 ±0.0000	1.4717e-02 ±0.0000	1.3120e-03 ±0.0000
SVR [44]	6.0757e-02 ±0.0000	2.2974e-02 ±0.0000	0.1410 ±0.0000	8.9807e-02 ±0.0000	4.3832e-03 ±0.0000	8.8686e-04 ±0.0000

A Tabela 3 apresenta o resultado do teste de Friedman para as medidas MAE, MAPE e MSE. Aqui é possível confirmar estatisticamente os resultados apresentados na Tabela 2. Note que o modelo proposto obteve o menor valor de *rank*, sugerindo que este é o modelo de previsão mais preciso, de acordo com todas as medidas investigadas, para a modelagem de séries temporais da COVID-19 investigadas neste trabalho.

Tabela 3: Resultados do teste de Friedman para as medidas MAE, MAPE e MSE.

	MAE		MAPE		MSE	
Posição	Modelo	Valor de *Rank*	Modelo	Valor de *Rank*	Modelo	Valor de *Rank*
1	DIDLNN	1.00	DIDLNN	1.00	DIDLNN	1.00
2	LSTM	3.50	LSTM	3.50	SVR	3.00
3	SVR	3.50	SVR	3.50	LSTM	3.50
4	MLP	4.00	MLP	4.00	MLP	4.00
5	CNN	5.50	CNN	5.50	NARX	6.00
6	NARX	5.50	NARX	5.50	REGTREE	6.00
7	RBF	6.50	RBF	6.50	CNN	6.50
8	REGTREE	7.00	REGTREE	7.00	RBF	6.50
9	ARIMA	8.50	ARIMA	8.50	ARIMA	8.50
	χ^2=10.80 and p-value=2.13e-01		χ^2=10.80 and p-value=2.13e-01		χ^2=11.20 and p-value=1.91e-01	

Na Tabela 4 são apresentados os resultados do teste de Tukey para as medidas MAE, MAPE e MSE. A análise por pares revela que o modelo proposto apresenta melhor desempenho em relação a todos os pares. Os menores valores da estatística do teste de Tukey sugerem que o modelo proposto tem, estatisticamente, melhor desempenho em relação aos modelos investigados.

Tabela 4: Resultados do teste de Tukey para as medidas MAE, MAPE e MSE.

	MAE		MAPE		MSE	
Par	Estatística	p-valor	Estatística	p-valor	Estatística	p-valor
DIDLNN - ARIMA	7.50	6.1699e-03	7.50	6.1699e-03	7.50	6.1699e-03
DIDLNN - CNN	4.50	1.0035e-01	4.50	1.0035e-01	5.50	4.4610e-02
DIDLNN - LSTM	2.50	3.6131e-01	2.50	3.6131e-01	2.50	3.6131e-01
DIDLNN - MLP	3.00	2.7332e-01	3.00	2.7332e-01	3.00	2.7332e-01
DIDLNN - NARX	4.50	1.0035e-01	4.50	1.0035e-01	5.00	6.7889e-02
DIDLNN - RBF	5.50	4.4610e-02	5.50	4.4610e-02	5.50	4.4610e-02
DIDLNN - REGTREE	6.00	2.8460e-02	6.00	2.8460e-02	5.00	6.7889e-02
DIDLNN - SVR	2.50	3.6131e-01	2.50	3.6131e-01	2.00	4.6521e-01

4.2 Análise do Comportamento da Previsão

Na Figura 7, apresentamos uma análise comparativa entre o valor real e previsto, do conjunto de teste, gerados pelo modelo proposto e pelos dois melhores preditores em nossa análise. Nestas figuras, é possível verificar, de maneira subjetiva, que as previsões são similares, mesmo com todas as medidas indicando melhor desempenho do modelo proposto.

4.3 Discussão

De acordo com os resultados apresentados nas seções anteriores, é possível verificar algumas evidências sugerindo que o modelo proposto tem, estatisticamente, melhor desempenho quando comparado aqueles obtidos com modelos clássicos e recentemente apresentados na literatura, considerando as medidas MAE, MAPE e MSE. Neste contexto os testes de Friedman e de Tukey confirmaram estatisticamente os resultados alcançados.

Também, é possível verificar que o mapeamento gerado pelo modelo proposto, composto por uma combinação entre operadores lineares e operadores morfológicos crescente-decrescente, mostra uma alta capacidade para reconstruir a dinâmica das séries temporais investigadas. Desta forma podemos confirmar a hipótese que as séries temporais da COVID-19 podem ser modeladas em termos de uma combinação de componentes lineares e não lineares. No entanto, outras evidências empíricas e teóricas desse comportamento devem ser investigadas em trabalhos futuros.

Alem disso, observamos um desempenho de previsão similar quando consideramos o modelo proposto e os modelos LSTM e SVR, a partir de uma análise subjetiva (gráfica) da previsão. Uma justificativa para este comportamento pode estar relacionado a capacidade destes modelos estimarem um mapeamento híbrido composto de componentes lineares e não lineares, combinados de alguma forma. No entanto, vale mencionar que mesmo com desempenho similar, nenhum destes modelos é capaz de superar o modelo proposto neste trabalho em termos das medidas investigadas.

Note que outras evidências podem dar suporte ao melhor desempenho do modelo proposto: i) a hipótese do fenômeno gerador de séries temporais da COVID-19 poder ser modelado em termos de uma combinação balanceada de componentes lineares e não lineares com comportamento crescente e decrescente. Neste sentido, o modelo proposto pode ser visto como um mapeamento direto dessas séries temporais, e ii) a capacidade do modelo proposto de estimar o percentual de uso das componentes linear e não linear em cada unidade de processamento da estrutura em camadas profunda. Também, é possível verificar que o processo de aprendizagem utilizado tem um comportamento estável (valores baixos para o desvio padrão), convergindo para pontos de ótimo (sub-ótimo) da superfície da função objetivo.

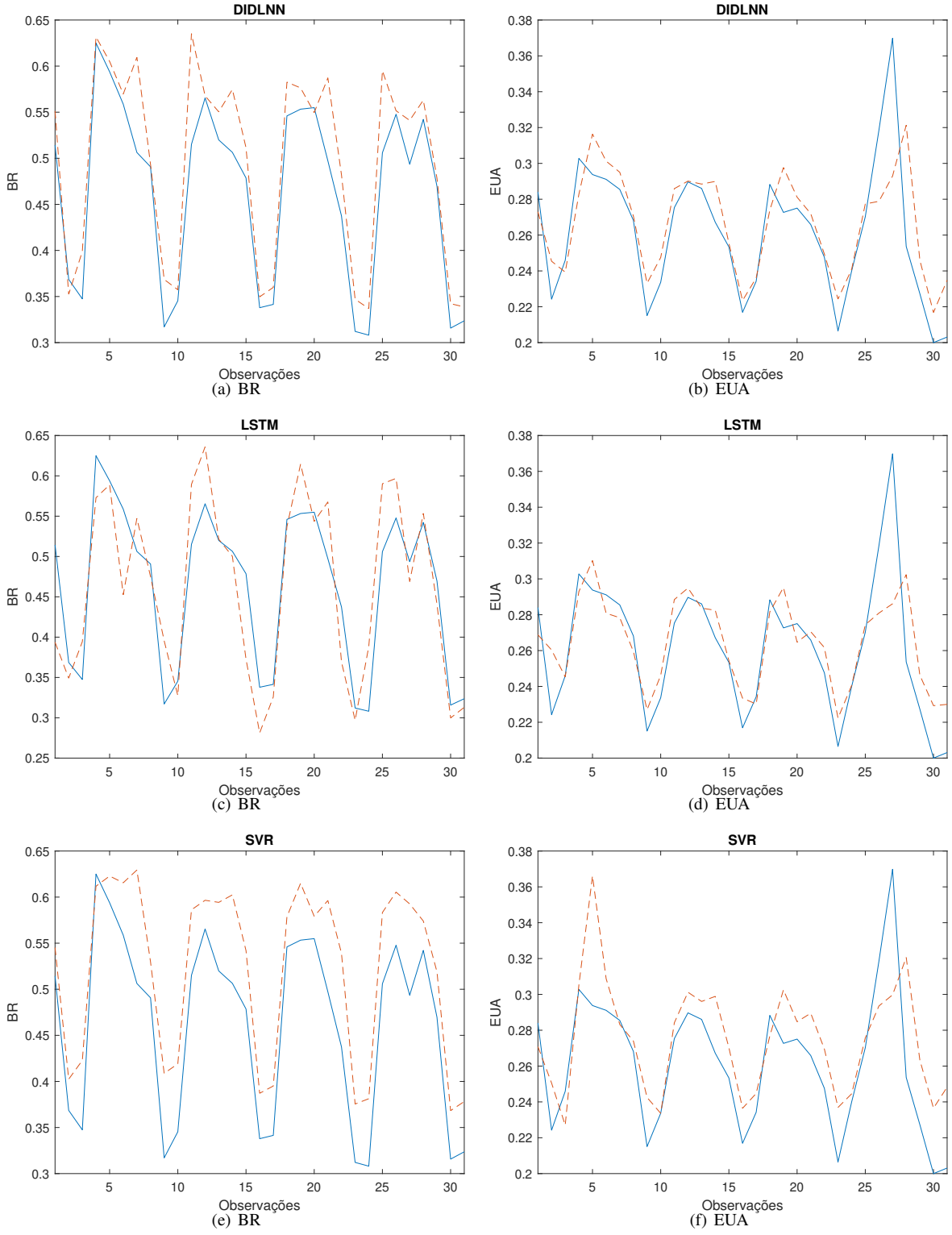


Figura 7: Resultados de previsão do modelo proposto ((a) e (c)) e do melhor preditor investigado ((b) e (d)) para as séries temporais consideradas (conjunto de teste).

5 CONCLUSÃO

Entender a dinâmica da pandemia da COVID-19 tem sido indispensável para dar suporte a ações estratégicas relacionadas a saúde e economia globais. Neste sentido, este trabalho apresenta evidências empíricas de que séries temporais da COVID-19, amostradas em frequência diária, podem ser modeladas em termos de uma combinação entre componentes lineares e não lineares, sugerindo que estas são de alguma forma previsíveis, baseado nos critérios a seguir: i) as funções de autocorrelação e autocorrelação parcial (para análise da dependência linear), ii) a informação mútua média (para análise da dependência não linear), e iii) o parâmetro de Hurst (para avaliar a não linearidade de processos autosimilares em termos de uma análise persistente

e anti-persistente).

Desta forma, desenvolvemos um modelo morfológico-linear profundo, chamado de *deep increasing-decreasing-linear neural network* (DIDLNN), treinado por um processo de aprendizagem baseado em gradiente descendente e capaz de prever este tipo particular de série temporal. Cada camada do modelo proposto é composta por neurônios artificiais híbridos dados por uma combinação de componentes lineares e não lineares para construir uma modelagem direta das séries temporais da COVID-19.

Vale mencionar que, independente da série temporal analizada, o modelo proposto obteve desempenho superior, para todas as medidas consideradas (MAE, MAPE e MSE), em relação a todos os modelos clássicos e recentemente apresentados na literatura. Portanto, podemos concluir que o modelo proposto tem uma eficácia geral muito melhor. Além disso, podemos verificar que o mapeamento gerado pelo modelo proposto é bastante complexo, pois é composto por camadas profundas com várias unidades de processamento dadas por combinações balanceadas entre operadores morfológicos crescente-decrescente e operadores lineares.

Logo, não é necessário saber, *a priori*, a função de combinação de ambas as componentes, uma vez que o modelo proposto ajusta automaticamente o percentual de uso de cada componente de cada unidade de processamento em cada camada. Portanto, este ajuste possibilita a reconstrução do fenômeno gerador das séries temporais da COVID-19 com apenas um estágio de previsão.

Como trabalho futuro, um estudo sobre a complexidade computacional do modelo proposto e seu processo de aprendizagem deve ser realizado. Como a quantidade de unidades de processamento e a quantidade de camadas do modelo foi definido empiricamente, uma análise de sensibilidade deve ser realizada para a escolha ótima (sub-ótima) destes parâmetros. Um estudo adicional sobre a modelagem de resíduos com o modelo proposto deve ser realizada, uma vez que alguma informação pode não ter sido capturada pelo modelo proposto e que pode ser modelada em termos dos resíduos.

6 AGRADECIMENTOS

Este trabalho é parcialmente apoiado pelo INES (instituto Nacional de Ciência e Tecnologia para Engenharia de Software), CNPq processo 465614/2014-0, CAPES processo 88887.136410/2017-00, e FACEPE processos APQ-0399-1.03/17 e PRONEX APQ/0388-1.03/14. Sergio Soares é parcialmente apoiado pelo CNPq, processo 309697/2019-0.

REFERÊNCIAS

[1] G. F. Gao. "From "A" to "Z": attacks from emerging and re-emerging pathogens". *Cell*, vol. 172, pp. 1157–1159, 2018.

[2] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao and W. Tan. "A Novel Coronavirus from Patients with Pneumonia in China, 2019". *The New England Journal of Medicine*, vol. 382, no. 2, pp. 727–733, 2020.

[3] C. Lai, T. Shih, W. Ko, H. Tang and P. Hsueh. "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges". *International Journal of Antimicrobial Agents*, vol. In Press, 2020.

[4] C. Wu, P. G. Postema, E. Arbelo, E. R. Behr, C. R. Bezzina, C. Napolitano, T. Robyns, V. Probst, E. Schulze-Bahr, C. A. Remme and A. A. M. Wilde. "SARS-CoV-2, COVID-19 and inherited arrhythmia syndromes". *Heart Rhythm*, vol. In Press, 2020.

[5] J. She, J. Jiang, L. Ye, L. Hu, C. Bai and Y. Song. "2019 novel coronavirus of pneumonia in Wuhan, China: emerging attack and management strategies". *Clinical and Translational Medicine*, vol. 9, no. 1, pp. 1–19, 2020.

[6] P. Zhou, X. Yang, X. Wang, B. Hu, L. Zhang, W. Zhang, H. Si, Y. Zhu, B. Li, C. Huang, H. Chen, J. Chen, Y. Luo, H. Guo, R. Jiang, M. Liu, Y. Chen, X. Shen, X. Wang, X. Zheng, K. Zhao, Q. Chen, F. Deng, L. Liu, B. Yan, F. Zhan, Y. Wang, G. Xiao and Z. Shi. "Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin". *bioRxiv*, vol. In Press, 2020.

[7] F. He, Y. Deng and W. Li. "Coronavirus Disease 2019 (COVID-19): What we know?" *Journal of Medical Virology*, vol. 92, no. 7, pp. 719–725, 2020.

[8] V. J. Munster, M. P. G. Koopmans, N. V. Doremalen, D. V. Riel and E. Wit. "A Novel Coronavirus Emerging in China - Key Questions for Impact Assessment". *The New England Journal of Medicine*, vol. 382, pp. 692–694, 2020.

[9] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis and R. Agha. "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)". *International Journal of Surgery*, vol. 76, pp. 71–76, 2020.

[10] WHO. "Coronavirus disease 2019 (COVID-19): Situation Report". Technical report, World Health Organization, 2021.

[11] Nature. "Coronavirus latest: global infections pass two million". *15 April*, 2020.

[12] F. K. Ayittey, M. K. Ayittey, N. B. Chiwero, J. S. Kamasah and C. Dzuvor. "Economic impacts of Wuhan 2019-nCoV on China and the world". *Journal of Medical Virology*, vol. 92, no. 5, pp. 473–475, 2020.

[13] W. Cheng-I, P. G. Postema, E. Arbelo, E. R. Behr, C. R. Bezzina, C. Napolitano, T. Robyns, V. Probst, E. Schulze-Bahr, C. A. Remme and A. A. M. Wilde. "SARS-CoV-2, COVID-19 and inherited arrhythmia syndromes". *Heart Rhythm*, vol. 17, no. 9, pp. 1456–1462, 2020.

[14] J. Cao, X. Hu, W. Cheng, L. Yu, W. Tu and Q. Liu. "Clinical features and short-term outcomes of 18 patients with corona virus disease 2019 in intensive care unit". *Intensive Care Medicine*, vol. 46, no. 5, pp. 851–853, 2020.

[15] Z. Wu and J. M. McGoogan. "Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of Cases From the Chinese Center for Disease Control and Prevention". *JAMA*, vol. 323, no. 13, pp. 1239–1242, 2020.

[16] M. Lazzerini and G. Putoto. "COVID-19 in Italy: momentous decisions and many uncertainties". *Lancet Global Health*, vol. 8, no. 5, pp. e641–e642, 2020.

[17] S. J. R. da Silva and L. Pena. "Collapse of the public health system and the emergence of new variants during the second wave of the COVID-19 pandemic in Brazil". *One Health*, vol. 13, pp. 100287, 2021.

[18] W. E. Parmet and M. S. Sinha. "Covid-19 - The Law and Limits of Quarantine". *The New England Journal of Medicine*, vol. 382, pp. 1–28, 2020.

[19] S. G. V. Rosa and W. C. Santos. "Clinical trials on drug repositioning for COVID-19 treatment". *Pan American Journal of Public Health*, vol. 44, no. e40, pp. 1–7, 2020.

[20] C. Musselwhite, E. Avineri and Y. Susilo. "Restrictions on mobility due to the coronavirus Covid19: Threats and opportunities for transport and health". *Journal of Transport & Health*, vol. 20, pp. 101042, 2021.

[21] K. Leung, J. T. Wu and G. M. Leung. "Effects of adjusting public health, travel, and social measures during the roll-out of COVID-19 vaccination: a modelling study". *The Lancet Public Health*, vol. 6, no. 9, pp. e674–e682, 2021.

[22] L. Li, Z. Yang, Z. Dang, C. Meng, J. Huang, H. Meng, D. Wang, G. Chen, J. Zhang, H. Peng and Y. Shao. "Propagation analysis and prediction of the COVID-19". *Infectious Disease Modelling*, vol. 5, pp. 282 – 292, 2020.

[23] R. Vaishya, M. Javaid, I. H. Khan and A. Haleem. "Artificial Intelligence (AI) applications for COVID-19 pandemic". *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 337 – 339, 2020.

[24] F. Petropoulos and S. Makridakis. "Forecasting the novel coronavirus COVID-19". *PLOS ONE*, vol. 15, no. 3, pp. 1–8, 03 2020.

[25] G. E. P. Box, G. M. Jenkins and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, New Jersey, third edition, 1994.

[26] A. Fraser and H. Swinney. "Independent Coordinates for Strange Atractors from Mutual Information". *Physical Review A*, vol. 33, no. 2, pp. 1134–1140, 1986.

[27] M. B. Stojanovic, M. M. Bozic, M. M. Stankovic and Z. P. Stajic. "A methodology for training set instance selection using mutual information in time series prediction". *Neurocomputing*, vol. 141, pp. 236–245, 2014.

[28] E. Hurst. "Long Term Storage Capacity of Reservoirs". *Transactions of the American Society of Civil Engineers*, vol. 116, pp. 770–799, 1951.

[29] J. M. P. Menezes and G. A. Barreto. "Long-term time series prediction with the NARX network: An empirical evaluation". *Neurocomputing*, vol. 71, no. 16-18, pp. 3335–3343, 2008.

[30] N. Diniz, F. G. Lima and A. C. da Silva Filho. "The Impact of the Hurst Window in the financial time series forecast: an analysis through the Exchange rate". *Review of Business Research*, vol. 12, pp. 27–33, 2012.

[31] W. C. Roda, M. B. Varughese, D. Han and M. Y. Li. "Why is it difficult to accurately predict the COVID-19 epidemic?" *Infectious Disease Modelling*, vol. 5, pp. 271 – 281, 2020.

[32] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti and M. Ciccozzi. "Application of the ARIMA model on the COVID-2019 epidemic dataset". *Data in Brief*, vol. 29, pp. 105340, 2020.

[33] N. Imai, I. Dorigatti, A. Cori, S. Riley and N. M. Ferguson. "Report 1: Estimating the potential total number of novel coronavirus (2019-nCoV) cases in Wuhan City". *https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-1-case-estimates-of-covid-19/*, 2020.

[34] M. Shen, Z. Peng, Y. Xiao and L. Zhang. "Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China". *Innovation*, vol. 1, no. 3, pp. 1–2, 2020.

[35] B. Tang, N. L. Bragazzi, Q. Li, S. Tang, Y. Xiao and J. Wu. "An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov)". *Infectious Disease Modelling*, vol. 5, pp. 248–255, 2020.

[36] B. Tang, X. Wang, Q. Li, N. L. Bragazzi, S. Tang, Y. Xiao and J. Wu. "Estimation of the Transmission Risk of the 2019-nCoV and Its Implication for Public Health Interventions". *Journal of Clinical Medicine*, vol. In Press, 2020.

[37] J. T. Wu, K. Leung and G. M. Leung. "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study". *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[38] C. You, Y. Deng, W. Hu, J. Sun, Q. Lin, F. Zhou, C. H. Pang, Y. Zhang, Z. Chen and X. H. Zhou. "Estimation of the Time-Varying Reproduction Number of COVID-19 Outbreak in China". *medRxiv*, 2020.

[39] S. Zhao, S. S. Musa, Q. Lin, J. Ran, G. Yang, W. Wang, Y. Lou, L. Yang, D. Gao, D. He and M. H. Wang. "Estimating the Unreported Number of Novel Coronavirus (2019-nCoV) Cases in China in the First Half of January 2020: A Data-Driven Modelling Analysis of the Early Outbreak". *Journal of Clinical Medicine*, vol. 9, no. 2, pp. 1–6, 2020.

[40] R. de A. Araujo, N. Nedjah, A. L. I. Oliveira and S. R. de L. Meira. "A deep increasing-decreasing-linear neural network for financial time series prediction". *Neurocomputing*, vol. 347, pp. 59–81, 2019.

[41] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[42] H. Alabdulrazzaq, M. N. Alenezi, Y. Rawajfih, B. A. Alghannam, A. A. Al-Hassan and F. S. Al-Anzi. "On the accuracy of ARIMA based prediction of COVID-19 spread". *Results in Physics*, vol. 27, pp. 104509, 2021.

[43] C. M. Yesilkanat. "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm". *Chaos, Solitons & Fractals*, vol. 140, pp. 110210, 2020.

[44] D. Parbat and M. Chakraborty. "A python based support vector regression model for prediction of COVID19 cases in India". *Chaos, Solitons & Fractals*, vol. 138, pp. 109942, 2020.

[45] N. M. Norwawi. "29 - Sliding window time series forecasting with multilayer perceptron and multiregression of COVID-19 outbreak in Malaysia". In *Data Science for COVID-19*, edited by U. Kose, D. Gupta, V. H. C. de Albuquerque and A. Khanna, pp. 547–564. Academic Press, 2021.

[46] S. Dhamodharavadhani and R. Rathipriya. "COVID-19 mortality rate prediction for India using statistical neural networks and gaussian process regression model". *African Health Sciences*, vol. 21, pp. 194–206, 2021.

[47] K. ArunKumar, D. V. Kalaga, C. M. S. Kumar, M. Kawaji and T. M. Brenza. "Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells". *Chaos, Solitons & Fractals*, vol. 146, pp. 110861, 2021.

[48] K. N. Nabi, M. T. Tahmid, A. R. M. E. Kader and M. A. Haider. "Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks". *Results in Physics*, vol. 24, pp. 104137, 2021.

[49] E. Ayyildiz, M. Erdogan and A. Taskin. "Forecasting COVID-19 recovered cases with Artificial Neural Networks to enable designing an effective blood supply chain". *Computers in Biology and Medicine*, vol. 139, pp. 105029, 2021.

[50] M. Friedman. "A Comparison of Alternative Tests of Significance for the Problem of m Rankings". *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[51] J. W. Tukey. "Comparing individual means in the analysis of variance". *Biometrics*, vol. 5, pp. 99–114, 1949.

[52] J. Demsar. "Statistical Comparisons of Classifiers over Multiple Data Sets". *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.