Modelagem e Identificação de Dados Epidemiológicos Associados à Pandemia de COVID-19 em Santa Catarina

Eduard Hermes Anschau*, Alexandro Garro Brito*, Pablo Andretta Jaskowiak*
*Universidade Federal de Santa Catarina
Rua Dona Francisca, 8300, Distrito Industrial
CEP: 89219–600, Joinville, Santa Catarina, Brasil
Emails: eduardhermes@hotmail.com, alexandro.brito@ufsc.br, pablo.andretta@ufsc.br

***Resumo*—O novo coronavírus (COVID-19) difundiu-se por todo o globo e tornou-se uma das grandes mazelas da contemporaneidade, impactando profundamente o Brasil, o qual configura como uma das nações mais afetadas pela doença. Desse modo, a necessidade por sistemas tecnológicos de combate à crise sanitária tornou-se ainda mais urgente nesse país. À vista disso, o presente artigo apresenta um estudo comparativo entre duas técnicas de modelagem e previsão de dados epidemiológicos associados à pandemia de COVID-19 no Brasil, especificamente no estado de Santa Catarina. Foram considerados modelos do tipo *Non-Linear Autoregressive model with eXogenous input* (NARX) polinomiais como contraponto à modelagem de séries temporais por meio da construção de redes neurais recorrentes da variante *Long short-term memory* (LSTM) para importantes séries de dados associadas à doença. O desempenho preditivo dos modelos, avaliado por meio da aplicação de métricas de desempenho tradicionais, mostrou que, para três das quatro séries temporais consideradas, o modelo NARX obteve resultados mais satisfatórios.**

***Palavras-chave*—COVID-19, aprendizado de máquina, NARX, long short-term memory**

I. Introdução

A modelagem e a simulação de sistemas são ferramentas de decisão importantes, as quais podem ser úteis na análise e controle de doenças humanas [1,2,3]. No entanto, como cada comorbidade apresenta características biológicas particulares, os modelos de previsão precisam ser adaptados a cada caso específico, a fim de se tornarem aptos a enfrentar situações reais [4,5].

Na contemporaneidade, destaca-se o COVID-19 – doença infecciosa, cujos primeiros casos de infecção são datados de dezembro de 2019. Em 30 de janeiro de 2020, a Organização Mundial da Saúde (OMS) declarou que tal fenômeno epidêmico se tratava de uma Emergência de Saúde Pública de Importância Internacional [6]. A epidemia se difundiu rapidamente por todo o mundo e, em 11 de fevereiro de 2020, a mesma instituição a renomeou como SARS-CoV-2 [7]. Em 11 de março de 2020, a doença já havia sido confirmada em mais de 118 000 casos relatados globalmente em 114 países, com mais de 90 por cento dos casos concentrados em apenas quatro deles. Dessa forma, a OMS declarou se tratar de uma pandemia. No Brasil, o primeiro caso confirmado da doença ocorreu em 2 de fevereiro de 2020, considerando que, até o dia 21 de maio de 2021, foram contabilizados aproximadamente 16 milhões de casos e 446 mil mortes, o que faz com que esse país seja uma das nações mais afetadas pela doença. Especificamente, no estado de Santa Catarina, foram confirmados em torno de 944 mil casos e pouco mais de 14 mil óbitos pela enfermidade, até o dia 21/05/2021, bem como uma taxa de ocupação de leitos de UTI no Sistema Único de Saúde (SUS) de 93,93% [8]. À vista desses fatos, corrobora-se a criticidade da situação sanitária a que o Brasil, bem como o estado de Santa Catarina, estão submetidos. Em um cenário em que se deve ponderar objetivos tão conflitantes, como manter a atividade econômica e o devido isolamento da população, é fundamental que as autoridades sanitárias e governamentais estejam dotadas de recursos tecnocientíficos para análise e descrição de dados associados ao avanço da epidemia.

Dentre as ferramentas matemáticas pioneiras em matéria de análise epidemiológica, destaca-se o modelo compartimental SIR, fundamentado na tríade "suscetíveis, infectados e recuperados" [9]. Esse modelo descreve a epidemia como um sistema de equações diferenciais que relaciona as parcelas da população contidas nos compartimentos S, I e R, de forma análoga à modelagem da interação entre partículas segundo o princípio da ação de massas [10]. Entretanto, o SIR não é capaz de explicar a persistência ou erradicação de doenças infecciosas [11,12]. Defende-se que a principal razão para isso é de que esse modelo considera a distribuição de indivíduos espacial e temporalmente homogênea [10]. Dessa forma, uma abordagem para lidar com a questão de populações heterogêneas, estudada em ecologia, são os chamados Modelos Baseados em Indivíduos, MBI (ou IBM, do inglês *Individual Based Model*) [13,14,15].

O estado da arte dos modelos descritivos e preditivos é em grande parte representado por tecnologias mais avançadas, a exemplo da inteligência artificial (IA), do aprendizado de máquina (AM) e da aprendizagem profunda (AP), as quais podem ser empregadas para identificar e prever distintos aspectos de dados epidêmicos. As principais áreas onde essas técnicas podem ser aplicadas são no diagnóstico precoce de

doenças, no rastreamento de contato, no desenvolvimento de medicamentos e vacinas, bem como na previsão de casos de contração de doenças [16,17]. No âmbito da identificação de séries temporais associadas à pandemia de COVID-19, destaca-se a aplicação de Redes Neurais Recorrentes, as quais são capazes de lidar com não linearidades, assim como com interdependências inerentes aos dados [18,19]. Nesse contexto, um estudo baseado em Aprendizado Profundo, mediante redes neurais LSTM é utilizado para prever os casos de transmissão de COVID-19 em países como os EUA, Itália e Canadá. Esse método mostrou bom desempenho preditivo devido à capacidade do LSTM em manipular séries de dados temporalmente dependentes. Hochreiter [20] faz uma análise comparativa de cinco modelos de AP para previsão de dados globais acerca dos casos confirmados de COVID-19, sendo que o modelo *Variational AutoEncoder* (VAE) obteve os melhores resultados, seguido pelas redes BiLSTM (variante bidirecional da rede LSTM) e LSTM. Outro [21] realizou a previsão dos casos confirmados de COVID-19 no Brasil por meio de seis modelos baseados em AP. Nesse trabalho, dois métodos resultaram em bom desempenho preditivo: *Support Vector Regression* e redes neurais recorrentes fundamentadas em LSTM.

Outra classe de técnicas bem sucedida na identificação de séries de dados corresponde ao modelo NARX, o qual mostra bom desempenho na identificação de séries temporais não-lineares, bem como agrega uma importante característica: a facilidade com que certos tipos de conhecimentos podem ser extraídos e incorporados [19]. No entanto, esse modelo é pouco explorado na temática de identificação dos dados epidemiológicos acerca da pandemia do novo coronavírus.

O presente trabalho propõe metodologias de modelagem não-linear para análise e previsão de dados epidemiológicos acerca da pandemia de COVID-19. Para tanto, propõe-se uma análise comparativa entre duas técnicas de previsão aplicadas às séries temporais da pandemia no estado de Santa Catarina, a saber: o modelo NARX e a rede neural recorrente do tipo LSTM. Acredita-se que esta pesquisa trará impactos tanto do ponto de vista científico como social. De posse de tais ferramentas, cientistas e autoridades serão capazes de prever comportamentos e tomar decisões para frear o avanço da epidemia e prover ações de manutenção da atividade econômica em nível local e estadual.

II. Materiais e Métodos

A. Descrição dos Dados

Para a conformação do conjunto de dados atinentes à pandemia de COVID-19 em Santa Catarina, utilizou-se como fonte os boletins de ocorrência fornecidos pelo site oficial da Secretaria de Saúde do Estado de Santa Catarina. Por meio desses boletins foram captadas as seguintes séries de dados: quantidade acumulada de casos confirmados; quantidade acumulada de óbitos; quantidade acumulada de casos de pacientes recuperados da doença; número de leitos SUS ocupados por enfermos de COVID-19.

Ressalte-se que o primeiro ponto de dados informado pela Secretaria de Saúde é datado do dia 13 de março de 2020, quando foram computados três casos confirmados da doença no estado catarinense. O último ponto de dados refere-se ao dia 01 de março de 2021. Todos os dados foram compilados em formato .csv, o que permite sua manipulação e análise por meio da biblioteca *Pandas*, em linguagem de programação Python. Em suma, os dados provenientes da Secretaria de Saúde de Santa Catarina estão dispostos graficamente na Figura 1. Por meio de simples inspeção visual das curvas dispostas nessa figura chamam atenção algumas características: há uma descontinuidade associada ao dia 14 de setembro de 2020 para as séries de dados atinentes à quantidade de casos e óbitos notificados, assim como a porção final de fevereiro de 2021 indica um regime atípico de crescimento de todas as quantidades representadas na Figura 1.

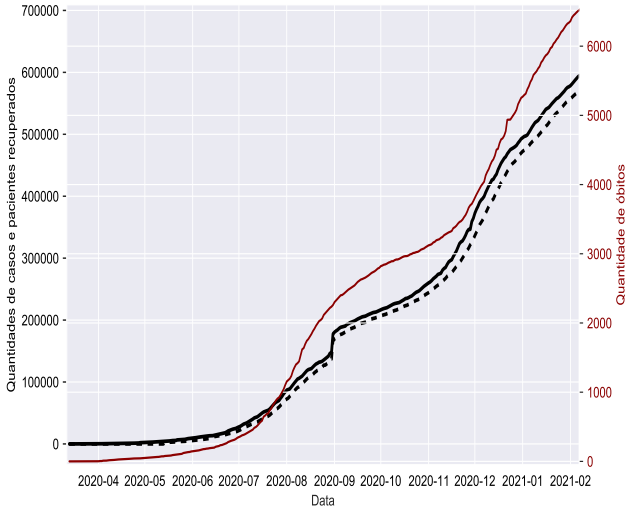

Fig. 1. Gráfico representativo dos dados epidemiológicos correspondentes à pandemia de COVID-19 em Santa Catarina.

De fato, a descontinuidade mencionada deve-se a um problema de atualização do banco de dados, à época, por parte da Secretaria de Saúde. Complementarmente, naquele momento vivenciava-se considerável crescimento das quantidades relacionadas à pandemia no Brasil. A Figura 2 representa graficamente a quantidade diária de leitos SUS de Santa Catarina ocupados por pacientes com COVID-19, informação que passou a ser notificada pelos boletins epidemiológicos emitidos pela Secretaria de Saúde somente a partir do dia 21 de abril de 2020. A curva exibida na Figura 2 indica aspecto semelhante de crescimento dessa quantidade ao visto nas outras variáveis, com taxa de crescimento substancial na porção final do mês de fevereiro de 2021. Por fim, a Tabela I sumariza o perfil estatístico das principais séries temporais epidemiológicas referidas por este trabalho.

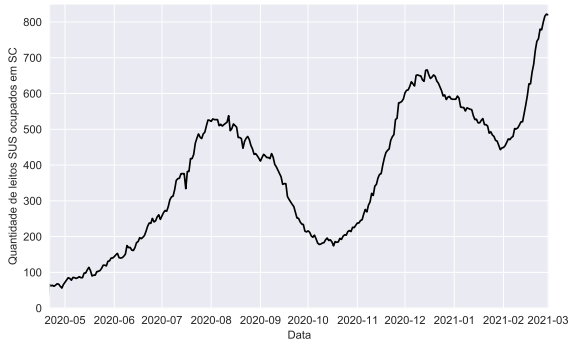

Fig. 2. Gráfico representativo da curva associada à quantidade diária de leitos SUS de Santa Catarina ocupados por pacientes com COVID-19.

TABELA I
CARACTERIZAÇÃO ESTATÍSTICA DAS SÉRIES DE DADOS RELACIONADAS À PANDEMIA DE COVID-19 EM SC COM BASE NAS QUANTIDADE DIÁRIAS.

Parâmetro	Casos	Óbitos	Recuperados	Leitos ocupados
Média	1963,9	21,9	1866,5	375,7
Desvio Padrão	2375,2	21,3	2190,4	191,3
Valor mín.	0	0	0	56
Valor máx.	30193	9,74	30186	822
Q-0,25	429,8	174	317,2	199,5
Q-0,5	1504,5	16	1525,5	418
Q-0,75	2903	32	2750,5	521,5

B. Modelagem das Séries Temporais

1) Modelo NARX polinomial: a identificação de sistemas dinâmicos não-lineares por meio do NARMAX (Non-Linear Autoregressive model with eXogenous input) de mapeamento polinomial é detalhada em [19]. O NARMAX, de forma geral, é um modelo autorregressivo munido de média móvel, cuja variante MISO (*multiple input, single output*) é expressa matematicamente pela Equação (1).

$$\begin{aligned} y(k) = F^l[&y(k-1), y(k-2),...,y(k-n_y),u_1(k-1),\\ &u_1(k-2),...,u_1(k-n_{u_1}), u_2(k-1), u_2(k-2),\\ &...,u_2(k-n_{u_2}),...,u_n(k-1),u_n(k-2),...,\\ &u_n(k-n_{u_n}),e(k-1), e(k-2),...,e(k-n_e)], \end{aligned} \tag{1}$$

em que F^l representa uma função não-linear geral munida de grau de não-linearidade l. As variáveis $y(k)$ e $e(k)$ são, respectivamente, a saída e o ruído aditivo do sistema, cujos atrasos máximos são representados por n_y e n_e. Similarmente, as n entradas exógenas utilizadas para conformação do modelo são denotadas por $u_i(k)$, com atrasos máximos denotados por n_{u_i}, de forma que i=1,...,n. No caso do presente estudo, F^l corresponde a uma expansão polinomial com grau de não-linearidade l. Adicionalmente, considera-se que as séries temporais epidemiológicas não possuem atraso puro de tempo e que nenhum dos parâmetros a ser estimado depende de $e(k)$. Dessa maneira, os modelos utilizados neste artigo correspondem ao modelo denominado NARX - caso especial da modelagem NARMAX que não se utiliza de fatores associados a médias móveis. Uma vez que o NARX de mapeamento polinomial produz uma estrutura linear nos parâmetros, a construção computacional do modelo é facilitada. A estimação de parâmetros é concebida por meio de aplicação de algoritmo baseado em mínimos quadrados (MQ), o qual confere minimização dos erros de previsão. Adicionalmente, o nível de importância dos termos regressivos utilizado para conformar a expressão matemática descrita pela Equação (1) é determinado por meio do parâmetro ERR (*error reduction ratio*), o qual representa quantitativamente a capacidade de cada termo regressivo explicar a variância do sinal de saída identificado. O objetivo de usar o critério ERR é tornar o modelo apto a organizar um conjunto de regressores candidatos em ordem decrescente de relevância [22]. A quantidade de termos utilizada para compor o modelo, ou seja, o ponto de corte em que o ERR deixará de incluir termos no modelo pode ser determinado usando outro critérios complementares [22]. Para esse fim, o presente trabalho utilizou-se do critério de informação de Akaike e Rissanen (AIC).

2) Modelos LSTM: a rede neural recorrente LSTM corresponde a um modelo sofisticado de aprendizado de máquina, fundamentado em unidades de memória e concebido para mitigar o problema de dissipação do gradiente [23]. Essencialmente, esse tipo de rede possui três elementos denominados como portas, a saber: a porta de entrada, a porta de esquecimento e a porta de saída. Esses componentes controlam o fluxo de informação pela rede, sendo que o estado de cada célula de memória é gerenciado pelas portas de entrada e de saída. O conteúdo de saída da camada LSTM é gerado a partir da porta de saída, que corresponde à informação memorizada e efetivamente utilizada pela rede neural. Esse mecanismo permite que a rede guarde informações importantes por considerável período de tempo durante o processo de treinamento. Sendo assim, pode-se afirmar que as características vantajosas dessas redes para a concepção do presente estudo são a grande capacidade para assimilação de dependências temporais de longo prazo e a significativa destreza no que tange à manipulação computacional de dados dispostos cronologicamente. Denotando a série temporal de entrada como X_t e o número de células LSTM por n, as portas da rede obedecem às seguintes relações:

- porta de entrada: $I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$;
- porta de esquecimento: $F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$;
- porta de saída: $O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$;
- estado intermediário da célula: $C_{int} = tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$;
- estado da célula (próxima entrada de memória): $C_t = F_t \cdot C_{t-1} \cdot C_{int}$;
- estado atualizado da célula: $H_t = O_t \odot tanh(C_t)$.

onde

- W_{xi}, W_{xf}, W_{hc}, W_{hf} e W_{ho} referem-se aos parâmetros de peso e b_i, b_f e b_o denotam os parâmetros de viés.
- W_{xc} e W_{hc} denotam parâmetros de peso, b_c é parâmetro

de viés, e $\odot$ refere-se ao operador matemático de produto escalar;
- $tanh$ e σ correspondem, respectivamente, às funções de ativação tangente hiperbólica e sigmoide.

A função sigmoide é usada pela porta de esquecimento para determinar o estado a ser preservado. A porta de entrada se utiliza da próxima função sigmoide e da primeira função $tanh$, que determina as informações a serem retidas no estado da célula ou quais informações devem ser desprezadas por esta. A última função sigmoide associa-se à porta de saída, que extrai a informação útil transmitida à célula posterior.

C. Métricas de desempenho

O desempenho preditivo dos modelos de identificação foi avaliado com base no cálculo de métricas tradicionais de avaliação de desempenho preditivo [24]. Neste estudo, as métricas de avaliação consideradas são: Erro Médio Absoluto (MAE), Erro Percentual Médio Absoluto (MAPE), Raiz do Erro Quadrático Médio (RMSE) e Raiz do Erro Quadrático Relativo (RRSE). As expressões matemáticas para cada uma das métricas são descritas pelas Equações (2) – (5).

$$\text{MAE} = \sum_{t=1}^{n} \frac{|y(t)-\hat{y}(t)|}{n}, \tag{2}$$

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y(t)-\hat{y}(t)}{y(t)} \right|, \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{100}{n} \sum_{t=1}^{n} (y(t) - \hat{y}(t))^2}, \tag{4}$$

$$\text{RRSE} = \sqrt{\sum_{t=1}^{n} \frac{y(t)-\hat{y}(t)}{\overline{y}-\hat{y}(t)}}. \tag{5}$$

Onde

- $y(t)$, $\hat{y}(t)$, $\overline{y}(t)$ referem-se, respectivamente, aos sinais de saída real, previsto pelo modelo de identificação não-linear e previsto por um preditor simples baseado na média aritmética;
- n indica a quantidade de observações utilizadas para o cálculo das métricas.

O MAE mede a magnitude média dos erros em um conjunto de previsões, sem considerar sua direção. O MAPE indica percentualmente quanto, em média, o preditor erra, sem compensar erros negativos com erros positivos. Já a raiz do erro quadrático médio (RMSE) de um estimador, calcula a média dos quadrados dos erros - ou seja, a diferença quadrática média entre os valores estimados e o que é estimado - normalizada pela aplicação da raiz quadrada. Para modelos de identificação perfeitos, as métricas supracitadas resultam em valores nulos. O RRSE é uma métrica estatística relativa à eficiência de predição caso um preditor simples tivesse sido aplicado sobre os dados. Esse preditor simples, mais especificamente, corresponde a média aritmética dos valores reais dos dados. Assim, o erro quadrático relativo, calculado pelo RRSE, computa o erro quadrático total e o normaliza por meio da divisão pelo erro quadrático total do preditor simples. Para modelos de identificação ideais essa métrica possuirá valor nulo, sendo que sua variação numérica está contida no intervalo real positivo.

D. Estratégia de modelagem

Posteriormente à coleta das séries de dados, estas foram submetidas a uma etapa de pré-processamento, a qual consistiu essencialmente na exclusão de dados irrelevantes como, a exemplo, a remoção de dados com valores negativos. Adicionalmente, aos dados efetivamente utilizados, foi aplicado procedimento de normalização matemática para cada cenário de identificação. Após, a massa de dados foi separada em dois conjuntos: dados para treinamento e dados para validação dos modelos de identificação, considerando que o grupo de dados de validação utilizado corresponde, por padrão, aos dados referentes ao mês de fevereiro de 2021. A partir da geração de diversos modelos, aqueles que obtiveram melhores características, no que concerne às métricas de desempenho, foram utilizados para a previsão do conjunto de dados de validação.

III. Modelagem e Identificação das Séries Temporais

Para a validação dos modelos NARX polinomiais, a estratégia utilizada consistiu na produção de diversos modelos a partir de toda a massa de dados disponíveis. Constatou-se que desprezar dados mais antigos das séries temporais acarretou em melhor desempenho preditivo para essa técnica e, por padrão, utilizou-se as séries temporais dos meses de novembro e dezembro do ano de 2020, para predição dos dados atinentes ao mês de fevereiro de 2021. Testes adicionais mostraram que a utilização de dados de momentos anteriores ao mês de dezembro, para produção dos modelos NARX, acarretavam em inadequadas características preditivas.

No âmbito da expressão matemática dos modelos NARX polinomiais obtidos, será utilizado o seguinte padrão de notação: $M(k)$, $R(k)$, $L(k)$ e $D(k)$ referem-se, respectivamente, às quantidades à tempo discreto associadas às quantidades de mortes, pacientes recuperados, leitos SUS ocupados e a variável correspondente aos dias da semana. Já $y(k)$ denota a quantidade em tempo discreto da variável de saída identificado.

No que se refere à concepção dos modelos LSTM, a fim de prevenir a sobreparametrização desses modelos, foram adicionadas camadas de *dropout* de valor 0,3 para todas as redes neurais criadas. Esses elementos são adicionados após cada camada LSTM, para conferir retenção da exatidão dos modelos. De forma complementar, ao final da rede LSTM é utilizada uma camada de densidade para interpretar os valores de saída e a função de ativação usada ($tanh$). Adicionalmente, todos os modelos produzidos por meio desta técnica empregaram apenas uma camada LSTM, já que essa configuração apresenta menor custo computacional, bem como demonstrou exatidão semelhante a redes neurais munidas de mais camadas. Com essa abordagem, 90% de toda a massa

de dados foi utilizada para treinamento e, posteriormente, as redes concebidas foram utilizados para prever os dados do mês de fevereiro de 2021. Dessa forma, comparou-se ambas as técnicas de identificação aplicadas por meio da avaliação de seus desempenhos quanto à capacidade de previsão. Ademais, analisou-se a correlação cruzada entre as variáveis estimadas e seus respectivos resíduos, no âmbito da identificação por modelos NARX, bem como foram obtidas as curvas de perdas de treinamento das redes neurais LSTM concebidas.

A. Visão geral dos modelos NARX e LSTM produzidos

1) Previsão de casos diários: o modelo de identificação obtido, por meio do emprego da técnica NARX, para a série de dados correspondente aos casos de COVID-19, possui as seguintes características:

- grau de não linearidade $l = 2$;
- as séries de dados de óbitos, pacientes recuperados, leitos ocupados e os dados relacionados aos dias da semana foram utilizadas como entradas exógenas;
- 16 termos regressores compõem o modelo final.

A expressão final do modelo NARX para a série temporal de casos previstos é dada pela Equação (6).

$$\begin{aligned} y(k) = {} & 8{,}2 \times 10^{-4} L(k-3)R(k-1) - 1{,}1 \times 10^{-4} R(k-5)R(k-2) \\ & +1{,}3 \times 10^{-3} M(k-6)y(k-7) + 5{,}7 \times 10^{-2} D(k-3)y(k-1) \\ & -9{,}4 \times 10^{-5} R(k-3)R(x-1) + 34{,}0 D(k-4)D(k-2) \\ & +2{,}9 \times 10^{-4} L(k-1)R(5) - 4{,}7 \times 10^{-3} R(k-6)M(k-4) \\ & +4{,}9 \times 10^{-1} y(k-2) - 6{,}1 \times 10^{-1} M(k-3)^2 \\ & +2{,}3 D(k-6)M(k-1) + 4{,}0 \times 10^{-2} D(k-5)y(k-4) \\ & +58{,}9 M(k-3) - 5{,}8 \times 10^{-2} D(k-1)y(k-2) \\ & +2{,}7 D(k-3)M(k-2) \\ & -5{,}4 \times 10^{-3} R(k-2)M(k-3). \end{aligned} \tag{6}$$

De outra parte, o modelo LSTM para previsão da quantidade de casos notificados fundamenta-se nas seguintes características:

- a única variável de entrada do modelo corresponde à série de dados de pacientes recuperados da doença;
- são utilizadas 256 células ocultas;
- taxa de aprendizado de 0,3;
- aplicação de 1000 épocas de treinamento.

A Figura 3 exibe graficamente as curvas de previsão de casos de COVID-19 obtidas por ambas as técnicas empregadas.

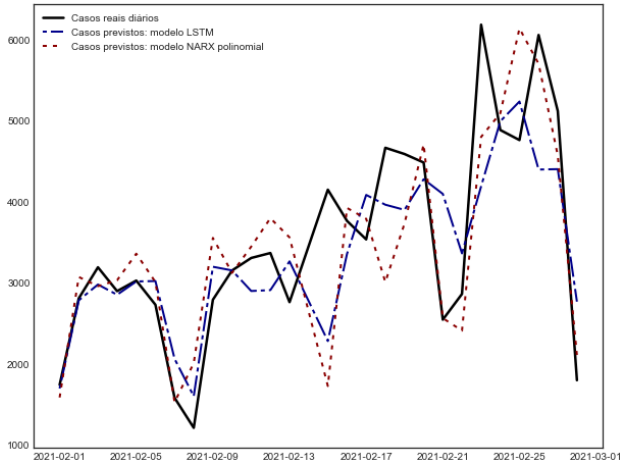

Fig. 3. Gráfico das curvas da quantidade real de casos confirmados de COVID-19 e as quantidades previstas pelos modelos NARX polinomial e LSTM.

2) Previsão de óbitos diários: o modelo de identificação do tipo NARX obtido para os dados de óbitos confirmados por COVID-19 em Santa Catarina possui as seguintes particularidades:

- grau de não linearidade $l = 3$;
- as séries de dados associadas aos leitos ocupados e os dados relacionados aos dias da semana foram utilizadas como entradas exógenas;
- ao final 4 termos regressores compõem o modelo resultante.

A expressão matemática do modelo NARX resultante para essa variável é dada pela Equação (7).

$$\begin{aligned} y(k) = {} & 2{,}38 \times 10^{-7} L(k-6)^2 L(k-3) \\ & -1{,}28 \times 10^{-4} y(k-5)y(k-4)y(k-3) \\ & +3{,}64 \times 10^{-2} D(k-5)D(k-3)y(k-6) \\ & -2{,}03 \times 10^{-3} D(k-3)y(k-6)y(k-1). \end{aligned} \tag{7}$$

Já o modelo LSTM para previsão da quantidade de óbitos diários baseia-se nas seguintes características:

- as variáveis de entrada utilizadas correspondem às séries de dados referentes às quantidades de pacientes recuperados, leitos ocupados e os dados associados aos dias da semana;
- são utilizadas 500 células ocultas;
- taxa de aprendizado da rede fixada em 0,5;
- aplicação de 230 épocas de treinamento.

A Figura 4 exibe as curvas de previsão para a quantidade de óbitos por COVID-19 obtidas por ambas as técnicas empregadas.

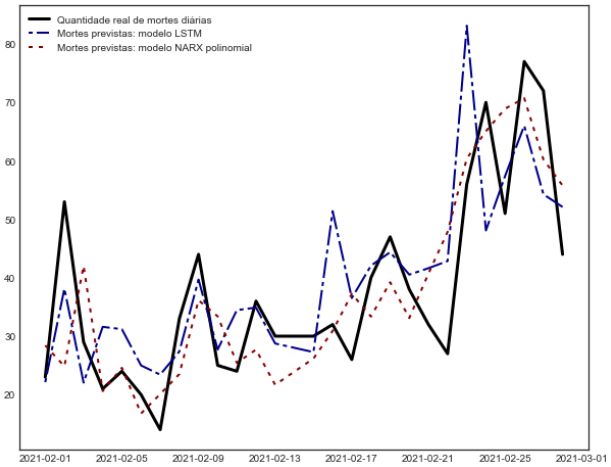

Fig. 4. Gráfico das curvas da quantidade real de óbitos de COVID-19 e as quantidades previstas pelos modelos NARX polinomial e LSTM.

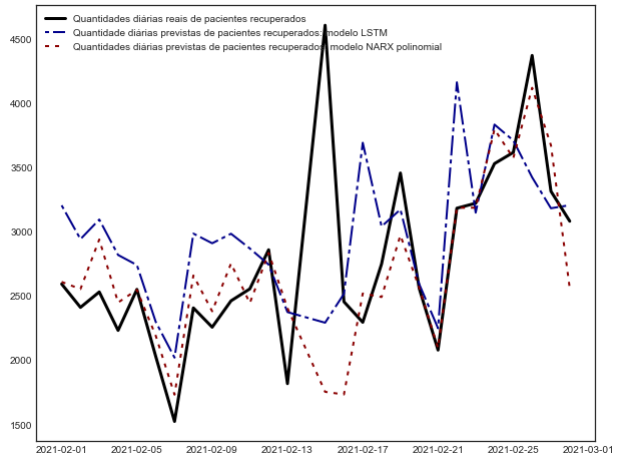

Fig. 5. Gráfico das curvas da quantidade real de pacientes recuperados e as quantidades previstas pelos modelos NARX polinomial e LSTM.

3) Previsão de pacientes recuperados: o modelo polinomial NARX obtido para a identificação dos casos de recuperação possui as seguintes características:

- grau de não linearidade do modelo: $l = 2$;
- as séries de dados associadas aos leitos ocupados e os dados relacionados aos dias da semana foram utilizadas como entradas exógenas;
- a composição final do modelo é munida de 7 termos regressores.

A expressão matemática final para o modelo de identificação das quantidades diárias de pacientes recuperados da doença é dada pela Equação (8).

$$\begin{aligned} y(k) = 0{,}15 &+ 1{,}07 \times 10^{-1} D(k-5)^2 y(k-5) \\ &+ 6{,}50 \times 10^{-2} D(k-5) y(k-4) \\ &8{,}28 \times 10^{-2} D(k-6) y(k-1) \\ &+ 2{,}00 \times 10^{-3} L(k-1)^2 \\ -4{,}3 \times 10^{-5} y(k-5) y(k-2) &+ 3{,}5 D(k-2) D(k-2). \end{aligned} \tag{8}$$

O modelo obtido via emprego da rede neural LSTM para a mesma variável é munida dos seguintes parâmetros:

- as variáveis de entrada utilizadas corresponde às séries de dados correspondentes às quantidade de casos diários, óbitos diários e leitos ocupados;
- são utilizadas 1000 células ocultas;
- taxa de aprendizado de 0,5;
- aplicação de 500 épocas de treinamento.

Na Figura 5 são mostradas as curvas de previsão da referida variável por COVID-19 obtidas mediante aplicação dos modelos de identificação.

4) Previsão de leitos SUS ocupados: quanto à previsão da quantidade de leitos ocupados em fevereiro de 2021, o modelo NARX resultante agrega as seguintes propriedades:

- grau unitário de não-linearidade;
- as entradas do modelo correspondem às séries de dados associadas aos leitos ocupados, aos casos confirmados de COVID-19, bem como aos casos de recuperação da doença;
- a composição final do modelo possui 20 termos regressores.

A expressão matemática final para o modelo de identificação NARX, referente à quantidade de leitos ocupados em fevereiro de 2021, é dada pela Equação (9).

$$\begin{aligned} y(k) = 3{,}24 &+ 9{,}44 \times 10^{-1} y(k-1) + 10{,}20 \times 10^{-2} y(k-2) \\ &+ 9{,}40 \times 10^{-3} y(k-3) - 9{,}75 \times 10^{-3} y(k-4) \\ &- 6{,}44 \times 10^{-2} y(k-5) + 4{,}50 \times 10^{-2} y(k-6) \\ &- 1{,}26 \times 10^{-1} y(k-7) + 5{,}03 \times 10^{-5} C(k-1) \\ &+ 1{,}16 \times 10^{-3} C(k-2) + 7{,}18 \times 10^{-4} C(k-3) \\ &+ 3{,}68 \times 10^{-4} C(k-4) + 1{,}05 \times 10^{-3} C(k-5) \\ &+ 1{,}22 \times 10^{-3} C(k-6) - 2{,}95 \times 10^{-4} R(k-1) \\ &- 7{,}13 \times 10^{-4} R(k-2) - 2{,}36 \times 10^{-4} R(k-3) \\ -1{,}25 \times 10^{-4} R(k-4) &- 9{,}63 \times 10^{-4} R(k-5) - 1{,}27 \times 10^{-3} R(k-6), \end{aligned} \tag{9}$$

O modelo obtido via emprego da rede neural LSTM para essa variável possui a seguinte constituição:

- o único atributo da rede neural criada refere-se à série de dados dos casos diários de recuperação;
- são utilizadas 512 células ocultas;
- taxa de aprendizado de 0,5;
- aplicação de 1000 épocas de treinamento.

A Figura 6 exibe graficamente as curvas de previsão quantidade de óbitos por COVID-19 obtidas por ambas as técnicas empregadas.

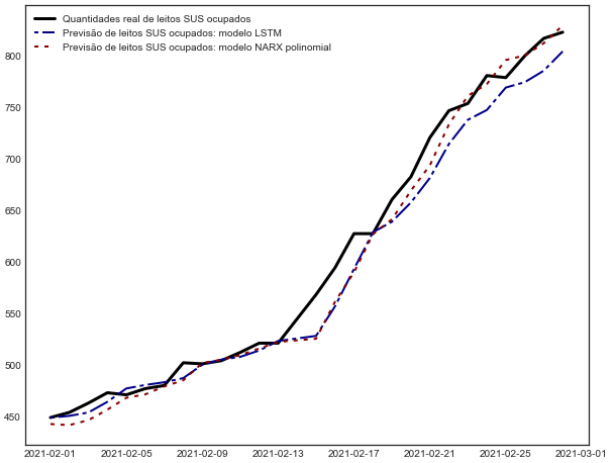

Fig. 6. Gráfico das curvas da quantidade real de leitos SUS ocupados e as quantidades previstas pelos modelos NARX polinomial e LSTM.

TABELA II
MÉTRICAS DE DESEMPENHO PARA OS MODELOS DE PREVISÃO.

Variável	Modelo	MAE	MAPE	RRSE	RMSE
Casos	NARMAX	608,18	0,20	0,72	818,02
	LSTM	430,52	0,12	0,61	696,36
Óbitos	NARMAX	6,55	0,18	0,62	9,18
	LSTM	9,74	0,25	1,07	12,84
Recuperados	NARMAX	252,53	0,08	0,93	552,03
	LSTM	557,88	0,18	1,18	740,87
Leitos Oc.	NARMAX	9,24	0,02	0,15	14,49
	LSTM	17,79	0,03	0,20	23,17

É facilmente visto, por meio de análise dos valores mostrados na Tabela II, que o modelo de identificação NARX polinomial supera a técnica LSTM em aspecto preditivo para três das quatro séries temporais - séries de óbitos, casos de recuperação e ocupação de leitos - no que se refere às métricas de desempenho utilizadas. De fato, o emprego preditivo das redes neurais mostrou-se superior apenas na previsão dos casos confirmados de COVID-19, já que, para esses dados, os valores obtidos para os parâmetros estatísticos são menores do que os resultantes da aplicação do modelo NARX.

B. Correlação presente nos resíduos para os modelos NARX

Uma característica relevante, atinente à qualidade dos modelos NARX, refere-se à análise das características dos resíduos presentes nesse modelo, a qual pode ser realizada por meio do estudo da correlação estatística. A Figura 7 exibe graficamente os resultados dos testes de autocorrelação dos resíduos, bem como a correlação cruzada entre os resíduos e as variáveis de entrada, resultantes da identificação NARX polinomial.

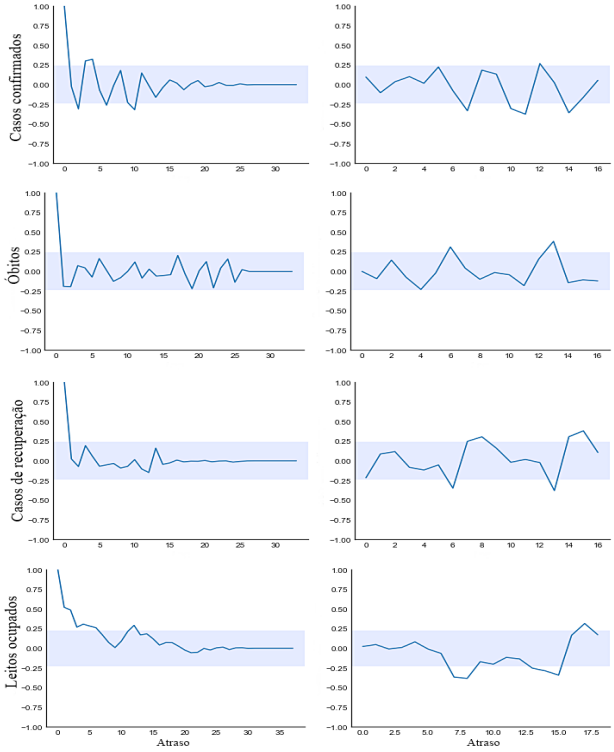

Fig. 7. Curvas concernentes aos testes de correlação residual dos modelos NARX.

Na Figura 7, a coluna à esquerda e à direita representam, respectivamente, as curvas de autocorrelação dos resíduos e as curvas de correlação cruzada entre os resíduos e as entradas dos modelos de identificação, com base em um intervalo associado a um nível de confiança probabilística de 75%. Verifica-se que parcela majoritária dos dados referentes à autocorrelação dos resíduos situa-se no intervalo de confiança, indicando que os resíduos, para cada cenário de identificação, são não correlacionados. Com respeito à correlação cruzada entre os resíduos e as variáveis de entrada de cada modelo, nota-se magnitudes mais elevadas, sem convergência à nulidade para dados mais distante no passado. Nesse caso, os desvios das curvas em relação aos limites de confiança indicam não-linearidades não modeladas pela técnica NARX. Dessa forma, pode-se inferir que os modelos NARX são razoáveis na identificação das variáveis epidêmicas.

C. Perdas associadas às redes neurais LSTM

Para avaliação da capacidade de aprendizado dos modelos LSTM, é adequado que as perdas dessas redes neurais - parâmetro inversamente proporcional ao erro de previsão cometido - convirjam para valores reduzidos. A Figura 8 mostra as curvas de evolução das perdas como função da quantidade de épocas de treinamento para cada rede neural LSTM concebida no âmbito do presente estudo. A Figura 8 nos permite inferir que todas as perdas convergem, mesmo que uma convergência relativamente mais lenta seja vista para o cenário de previsão de casos confirmados de COVID-19. Dessa forma, verifica-se desempenho adequado de aprendizagem

para as redes neurais produzidas. Ressalte-se que, como o conjunto de dados utilizados na modelagem era reduzido, circunstâncias melhores de aprendizado das redes seriam obtidas a partir do uso de conjuntos maiores de dados.

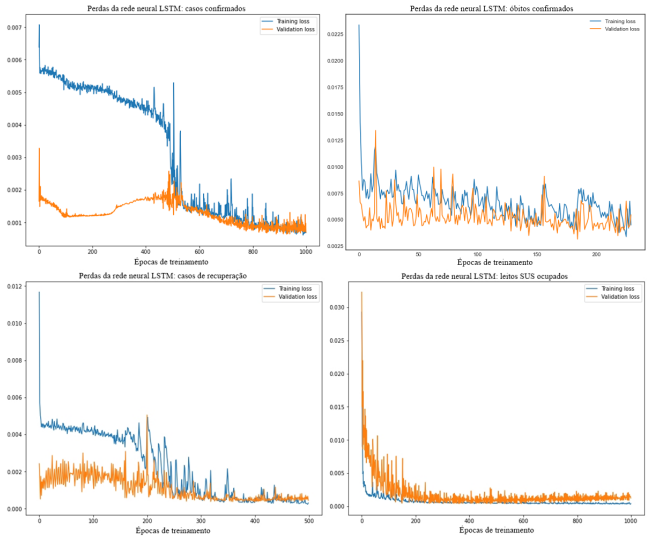

Fig. 8. Evolução das perdas das redes neurais LSTM.

IV. Conclusões

O presente estudo teve como objetivo a obtenção de modelos de identificação de séries temporais de dados associados à pandemia de COVID-19, no estado brasileiro de Santa Catarina. Para tanto, foram utilizados dois métodos de previsão de sistemas não lineares: o NARX polinomial e a modelagem mediante redes neurais recorrentes da variante LSTM. A pesquisa se fundamentou nas quantidades de casos, óbitos, pacientes recuperados e leitos SUS ocupados por pacientes com COVID-19.

Os resultados oriundos da análise de correlação residual presente nos modelos NARX mostram razoabilidade na identificação das séries epidêmicas, com algumas evidências de dinâmicas não apropriadamente descritas por essa técnica. Nesse ensejo, a utilização de outras variáveis de entrada, a inclusão de termos de média móvel nos modelos matemáticos, bem como a aplicação de distintos algoritmos de estimação de parâmetros, são alternativas úteis para aprimoramento dos modelos polinomiais, as quais serão tratadas em pesquisas posteriores. A análise das perdas, obtidas no âmbito da modelagem via redes neurais, revela boa capacidade de aprendizado em todos os cenários de identificação. Adicionalmente, os resultados do trabalho indicam superioridade do modelo NARX polinomial para três das quatro variáveis epidemiológicas, com base nos valores obtidos para quatro métricas de desempenho tradicionais.

Ressalte-se que a realização de distintos ajustes paramétricos, bem como a concepção de novas características e a aplicação de validação cruzada, no âmbito das redes LSTM, otimizariam o desempenho dessa técnica. Adicionalmente, a pouca quantidade e o incerto nível de confiabilidade dos dados, juntamente com escasso conhecimento sobre aspectos do da dinâmica de disseminação do novo coronavírus, configuram, por si só, limitações em esfera de epidemiologia matemática. De fato, os fatores técnicos supracitados são passíveis de estudo como relevantes objetos de pesquisa em trabalhos futuros.

Agradecimentos

Os autores agradecem à FAPESC pelo auxílio financeiro concedido.

Referências Bibliográficas

[1] M, Anderson. Population biology of infectious diseases: part 1. Nature, 280 (1979). pp. 361-367.

[2] B. Ivorra et al. “Mathematical formulation and validation of the Be-FAST model for classical swine fever virus spread between and within farms”, *Annals of Operations Research*, pp. 25-47, 2014.

[3] H.R. Thieme. Mathematics in population biology Mathematical Biology Series, Princeton University Press (2003).

[4] F. Brauer, C. Castillo-Chávez, “Mathematical Models in Population Biology and Epidemiology”, *Texts in Applied Mathematics*, Springer (2001).

[5] D. Yan, H. Cao, “The global dynamics for an age-structured tuberculosis transmission model with the exponential progression rate Applied Mathematical Modelling”, (2019), pp. 769-786.

[6] W.H. Organization. Director-general’s opening remarks at the media briefing on COVID-19 - 11 march 2020. Disponível em: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-COVID-19—11-march-2020. Acesso em 12 de Outubro de 2020.

[7] Gorbalenya, A.E. *et al*. Severe acute respiratory syndrome-related coronavirus: the species and its viruses - a statement of the coronavirus study group bioRxiv (2020), pp. 1-20.

[8] COVID-19: Governo de SC prevê agravamento da pandemia. Disponível em:https://g1.globo.com/sc/santa-catarina/noticia/2021/05/21/COVID-19-com-aumento-no-numero-de-casos-ativos-da-doenca-governo-de-sc-projeta-agravamento-da-pandemia.ghtml. Acesso em 21 de 21 de maio de 2021.

[9] H. W. Hethcote, “The mathematics of infectious diseases”, ”SIAM Review”, pp. 599– 653, 2000.

[10] M. J. Keeling, P. Rohani. (2002). “Estimating spatial coupling in epidemiological systems: a mechanistic approach”, *Ecology Letters*, pp. 20–29, 2002.

[11] A. L. Lloyd. “Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics”, *Theoretical Population Biology*, pp. 59–71, 2001.

[12] M. Keeling, B. Grenfell. “Individual based perspectives on R-0”, *Journal of Theoretical Biology*, pp. 51–61, 2000.

[13] E. G. Nepomuceno et al. “Modelagem de sistemas epidemiológicos por meio de modelos baseados em indivíduos”, *Anais do XVI Congresso Brasileiro de Automática*, pp. 2399–2404, 2006.

[14] R. Vaishya et al. “Artificial Intelligence (AI) applications for COVID-19 pandemic”, *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, pp. 337-339, 2020.

[15] W. Naudé. “Artificial Intelligence against COVID-19: An Early Review”, *IZA Institute of Labor Economics*, 2020.

[16] X. Zhu et al. “Attention-based recurrent neural network for influenza epidemic prediction”, *BMC Bioinformatics*, 2019.

[17] H. Hewamalage , C. Bergmeir, K. Bandara. “Recurrent Neural Networks for Time Series Forecasting: Current status and future directions”. *Int Journal of Forecasting*, pp. 388-427, 2021.

[18] M. Ribeiro et al. “Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil”, *Chaos Solitons Fractals*, 2020.

[19] L. A. Aguirre. (2007). Introdução à Identificação de Sistemas: técnicas lineares e não lineares aplicadas a sistemas reais, Editora da UFMG. 3a edição.

[20] S. Hochreiter, J. Shmidhuber. “Long short-term memory”, *Neural Computation*, pp. 1735–1780, 1997.

[21] S. Kaushik et al. “AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures”. *Frontiers in Big Data*, 2020.