Otimização de Modelo Híbrido para Previsão de Séries Temporais Não-lineares

Luana Goncalves

***Resumo*— A previsão de séries temporais é vital em diversas áreas do conhecimento. Métodos tradicionais como ARIMA, apesar de populares, se limitam a séries temporais lineares estacionárias. O uso de métodos de Aprendizado de Máquina tem se mostrado eficaz para modelar não-linearidades. Este estudo propõe uma metodologia híbrida que combina ARIMA e Redes Neurais, associados à otimização dos intervalos de atraso no tempo para a estimação. O objetivo é melhorar a precisão da previsão em séries temporais complexas e dinâmicas.**

***Palavras-Chave*— Séries temporais, redes neurais, ARIMA.**

***Abstract*— The prediction of time series is crucial in various areas of knowledge. Traditional methods like ARIMA, although popular, are limited to linear and stationary time series. The use of Machine Learning methods has proven effective in modeling non-linearities. This study proposes a hybrid methodology that combines ARIMA and Neural Networks, associated with the optimization of time lag intervals for estimation. The aim is to improve the accuracy of prediction in complex and dynamic time series.**

***Keywords*— Time Series Artificial, Neural Networks, ARIMA.**

I. Introdução

A possibilidade de antecipar comportamentos a priori é essencial na tomada de decisão a respeito a eventos futuros em quase todas as áreas do conhecimento. Contudo, prever valores resultantes no tempo para séries que por natureza são estocásticas e não estacionárias não é uma tarefa simples, como explicado por [1]. Uma vez que, apesar de ser possível a decomposição de séries em tendências e sazonalidades, tais sequências por vezes apresentam flutuações de curto prazo que não são sistemáticas, previsíveis ou lineares [2]. Conforme definido por [3], uma série temporal é qualquer conjunto de observações ordenadas no tempo. Do ponto de vista estatístico, isso se resume à uma sequência de variáveis aleatórias, a qual pode ser referida como um processo estocástico discreto ao longo do tempo [4]. Uma suposição geral admitida para previsão de tais dados, é que o comportamento passado da série temporal contém informações necessárias para prever sua atuação futura, com isso, é possível construir modelos probabilísticos ou estocásticos, tanto no domínio do tempo, como no domínio da frequência. Dessa forma, a literatura sobre previsão de séries temporais é um tópico bastante discutido nos últimos tempos, o que hoje nos permite: investigar o mecanismo gerador de séries temporais; fazer previsões a curto ou longo prazo de valores futuros, além de procurar periodicidades relevantes nos dados.

Devido à sua popularidade nas últimas décadas, a maioria das tentativas de previsão de séries temporais emprega abordagens como o *Auto-Regressive Integrated Mean Average* (ARIMA), com enfoque em séries temporais univariadas [1], [5], [6]. No entanto, a maior limitação da abordagem ARIMA é devido suas duas premissas: (i) que a série temporal prevista seja linear e (ii) estacionária, ou seja, suas propriedades estatísticas básicas, como média, variância e covariância, devem permanecer constantes ao longo do tempo[1], [7].

Outra abordagem de previsão é o emprego de métodos de Aprendizado de Máquina. Em oposição aos modelos estatísticos, buscam descrever as propriedades dos dados sem o conhecimento prévio da distribuição dos mesmos, o que representa uma considerável vantagem na previsão de séries complexas e altamente não-lineares. Em [8], três algoritmos de inteligência artificial para a previsão de valores financeiros foram implementados e comparados, os quais são: Regressão Linear, *k-nearest neighbors* (KNN) e *support vector machine* (SVM). Recentemente, para o cenário de séries multivariadas, Redes Neurais Artificiais (RNA) também tornaram-se uma ferramenta acessível, isso porque são demasiadamente efetivas em mapear a dinâmica de séries temporais não estacionárias, não-paramétricas e não-lineares, devido sua alta capacidade adaptativa.

Em [9], uma predição financeira foi treinada utilizando RNA a partir do método *Backpropagation*. Entretanto, os modelos Multi-Layer Perceptron (MLP) frequentemente encaram problemas como *overfitting*, decaimento de erro retropropagado, e não determinam automaticamente os intervalos de tempo ideais durante a modelagem [10]. Em [11]–[13], modelos híbridos de previsão com ambos métodos, RNA e ARIMA, combinam esforços para mapear os domínios lineares e não-lineares das séries. Nesse contexto, o objetivo desse trabalho é apresentar uma nova metodologia de previsão de séries temporais, ao combinar modelos lineares e não-lineares associados à otimização dos intervalos de atraso no tempo para a estimação. O modelo proposto é dividido em três etapas: na primeira, um modelo ARIMA é estimado a fim de identificar a existência de componentes lineares nos dados. Por conseguinte, uma MLP não-linear, é treinada para prever o resíduo da estimação ARIMA. Os resultados de ambas etapas se combinam como entrada para o treinamento de uma segunda MLP para a estimação final.

O resto do texto está organizado como segue: a Seção II explica os materiais necessários, a Seção III apresenta a metodologia proposta; a Seção III contempla os resultados obtidos, assim como, sua discussão.

Luana Goncalves, Belém-PA, e-mail: luanagoncalves4@live.com.

II. Previsão de Séries Temporais

A. Auto-Regressive Integrated Mean Average

O modelo ARIMA (p,d,q), também denominado como metodologia de Box-Jenkins, resulta da combinação de três componentes: sendo p o Auto-regressivo (AR), o qual preconiza que a variável de interesse pode ser obtida a partir de seus próprios atrasos; d representa o filtro de Integração (I), que indica as diferenças entre os valores atuais e os valores passados e q é o componente de Médias Móveis (MA), esse utiliza os últimos valores históricos para predizer o próximo valor [14]. Assim, a aplicação do modelo é composta de quatro etapas: identificação, estimação, verificação e previsão [15]. A etapa de identificação consiste em determinar os filtros (p, d, q) e a ordem que melhor representa a série temporal.

Em [16], propõem-se que caso a série seja estacionária, o filtro d é zero e a série não precisa de diferenciação, neste caso tem-se um modelo ARIMA (p,0,q). Se a série não for estacionária, o número de diferenciações para torná-la estacionária é representado pelo filtro d. Para os demais parâmetros, faz-se o uso da função autocorrelação (FAC) e a função de autocorrelação parcial (PACF).

Uma vez que identificados os parâmetros p, q e d; a fase de estimação consiste na aproximação por mínimos quadrados do modelo de regressão dos p parâmetros ϕ e dos q parâmetros θ e da variância ε do modelo de regressão, dado por:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_p \varepsilon_{t-p} \quad (1)$$

Logo, a verificação deve investigar se o modelo está super especificado, ou seja, contém parâmetros em excesso, ou está subespecificado por falta de parâmetros. Finalmente, após a identificação, estimação e verificação, a previsão do modelo deve ser testada por meio de simulações. Os modelos ARIMA têm a capacidade de previsão comprometida no longo prazo, então, sempre que possível, deve-se redimensioná-los [14].

B. Rede Neural Artificial

Criadas a partir do estudo do funcionamento do cérebro humano por McCulloch e Pitts, as redes neurais artificiais são modelos computacionais, originalmente concebidas para executar tarefas de natureza estática [17]. Apesar de implicar, portanto, que não foram idealizadas para tratar problemas temporais, o método de janela de tempo foi a primeira adaptação da rede MLP treinada com o algoritmo *backpropagation* para processamento dinâmico, com a premissa que os valores passados da série contém informações necessárias para prever os dados futuros [18]. Tal estrutura de RNA é representada na Figura 1, na qual as variáveis de entrada da rede são seus próprios atrasos, aproximados por uma função cujo objetivo é estimar valores a posteriori.

As saídas das camadas de entrada são conectadas por uma ou mais estruturas, chamadas de camadas ocultas, onde são mapeadas por uma função linear ou não linear. A última camada oculta é conectada aos neurônios da camada de saída, a qual também representa uma função de ativação, um exemplo pode ser dado a partir de uma sigmóide pela seguinte expressão:

$$f(x) = \frac{1}{(1 + e^{-x})} \quad (2)$$

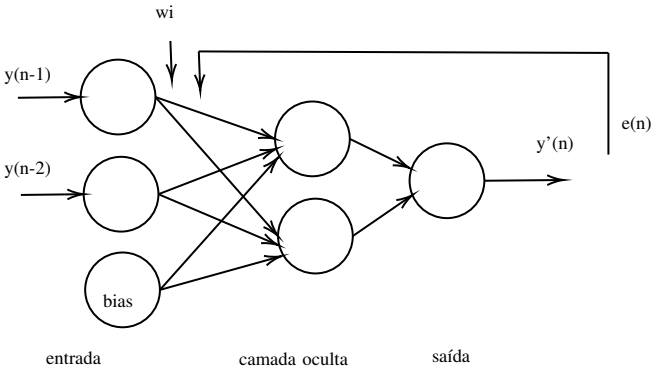

Fig. 1. Exemplo de estrutura RNA – MLP com algoritmo *backpropagation* com duas entradas, uma camada oculta e uma saída.

Vale ressaltar que os pesos inicialmente assumem valores aleatórios, os quais serão atualizados a partir do erro de estimação entre a saída da rede e a série a ser aproximada. Maiores detalhes sobre a utilização de RNA para previsão de séries temporais podem ser obtidos em [19].

C. Método de Newton

Em [20], demonstra-se que não existe forma algébrica para se calcular exatamente as raízes de polinômios com grau maior ou igual a 5. Dessa forma, na busca de processos através dos quais se obtém o melhor valor de uma grandeza e na impossibilidade de expressar exatamente os zeros de funções, uma opção são métodos de busca de, ao menos, uma aproximação deles. Nesse contexto, métodos de otimização numérica são alternativas proveitosas. Um dos mais famosos é o método de Newton, também conhecido como método de aproximações sucessivas, uma vez que consiste em aplicar sucessivamente derivações a partir de um valor inicial. Mais precisamente, consideremos uma função $f : I \to R$ possuindo derivada contínua e não nula em todo ponto do intervalo I. Tomando $x_0 \in I$, definimos:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \quad (3)$$

Se x_1 também pertence a I, é certo que:

$$x_2 = x_2 - \frac{f(x_2)}{f'(x_2)}. \quad (4)$$

Assim, se uma sequência de pontos em I for convergente, usando indutivamente o processo acima, o limite desta será uma raiz da função $f(x)$. Tendendo a infinito nos dois lados da equação, temos:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad (5)$$

e denotando por ao limite da sequência (x_n), teremos que

$$a = a - \frac{f(a)}{f'(a)}. \quad (6)$$

Logo, f(a) = 0. Com isso, permite-se estabelecer a aproximação do valor ótimo de variados processos.

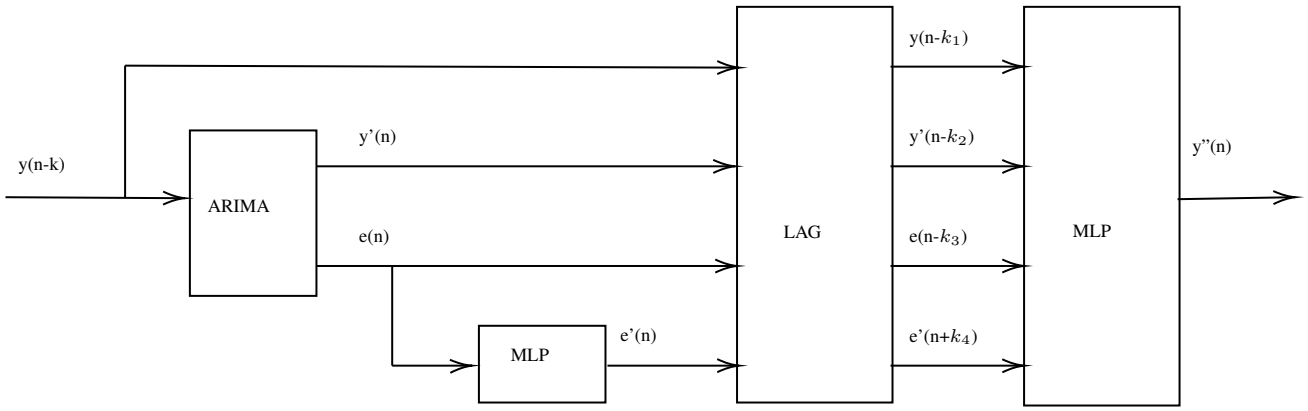

Fig. 2. Fluxograma das etapas de estimação.

III. METODOLOGIA

De acordo com [21], uma série temporal é composta por um estrutura de autocorrelação linear e uma componente não linear, a qual pode ser descrita por:

$$y(n) = y_l(n) + y_n(n), \tag{7}$$

onde, $y(n)$ representa a série a ser estimada, $y_l(n)$ caracteriza a componente linear e $y_n(n)$ a componente não linear. Logo, a proposta desse artigo é um modelo de estimação, representado pelo fluxograma da Figura 2, cuja as entradas consistem nos atrasos das decomposições lineares, não lineares e nos avanços da estimação da componente não linear. Uma vez que, frequentemente é difícil, na prática, determinar se a série em estudo é gerada a partir de processos lineares ou não, além disso, as séries temporais do mundo real raramente são absolutamente lineares ou não lineares. Portanto, há melhores chances de modelar diferentes padrões nos dados combinando variados modelos. Primeiramente, a metodologia proposta consiste em decompor a série a ser estimada, cuja a parte linear é obtida através da estimação ARIMA, dada por $y'(n)$, e a parte não linear é representada a partir de seu resíduo, dado por:

$$e(n) = y(n) - y'(n) \tag{8}$$

onde $e(n)$ representa o erro de estimação entre $y(n)$ e $y'(n)$. Uma vez que a série é decomposta, o passo seguinte é o treinamento de uma MLP para modelagem da componente não linear, dado que sua previsão é mais desafiadora. Então, o modelo RNA treinado para o resíduo é representado por:

$$e(n) = f(e(n-1), \cdots, e(n-k)) + \epsilon(n) \tag{9}$$

$$e(n) = w_0 + \sum_{j=1}^{Q} w_j g\left(w_{0j} + \sum_{i=0}^{k} w_{i,j} e(n-i)\right) + \epsilon(n) \tag{10}$$

onde f representa as funções não lineares nas camadas ocultas da MLP , ϵ_t o erro aleatório e k o número de atrasos na entrada da rede. O modelo apropriado é encontrado quando $\epsilon(n)$ é aleatório.

A despeito da Equação 7 e do método denotado em [21], o modelo proposto nesse artigo considera que uma série temporal deve ser representada por uma função de suas componentes lineares e não lineares. Logo, a composição das componentes da série é dada por:

$$y(n) = f\left(y_l(n), y_n(n)\right) \tag{11}$$

Com isso, uma MLP adicional tem como objetivo modelar a função $f\left(y_l(n), y_n(n)\right)$, representada por:

$$\begin{aligned} y(n) = f(&y(n-1, \cdots, n-k_1), \\ &y'(n-1, \cdots, n-k_2), \\ &e(n-1, \cdots, n-k_3), \\ &e'(n+1, \cdots, n+k_4)), \end{aligned} \tag{12}$$

onde y representa a série a ser estimada; y' a sua estimação ARIMA; e o resíduo entre y e y'; e e' a previsão resultado do modelo de e. Finalmente, para estimar os valores ótimos de k_1, k_2, k_3 e k_4, que são os intervalos de tempo ideais durante a modelagem, foi implementado o método de Newton, no sentido de minimizar o erro obtido entre a série real e a série estimada. A função de *fitness* da etapa de otimização é o *Mean squared error* (MSE) entre a série real, denotada por y, e a série prevista y'', dado por:

$$\textit{fitness} = \frac{1}{t} \sum_{n=0}^{t} (y(n) - y''(n))^2. \tag{13}$$

IV. RESULTADOS

Os testes foram realizados com o uso do *dataset* de validação *Canadian lynx forecasting*, uma vez que representa uma referência em modelagem não linear. Os dados representam o conjunto anual do número de linces caçadas no Distrito Mackenzie River no Noroeste do Canada no período de 1821 a 1934, os quais estão representados na Figura 3. As simulações foram realizadas com um processador "Intel(R) Xeon(R) CPU @ 2.20GHz, RAM 1.0 GB" no ambiente do *Google Colab* e com a linguagem de programação *Python*. Com o auxílio do pacote *Pmdarima*, um modelo autoregressivo de ordem 12, AR (12), foi estimado. Assim como, as estimações baseadas em RNA-MLP foram implementadas com o auxílio do pacote *Scikit-Learn*. Finalmente, o método de otimização foi desenvolvido com o pacote *Optuna*.

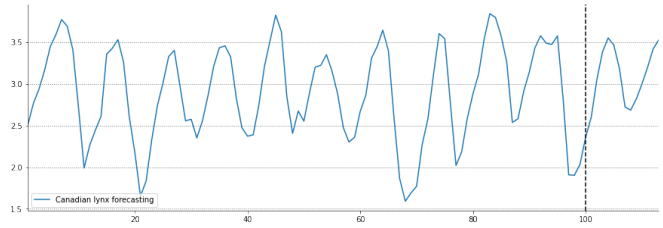

Fig. 3. Representação do conjunto de dados: *Canadian lynx forecasting*, para comparação com outros modelos foi aplicado o logaritmo na base 10 para a análise.

Na Tabela I é apresentada a comparação da performance do método proposto com outros métodos encontrados na literatura e a representação da estimação dos valores da metodologia proposta está apresentada na Figura 4.

TABELA I

Modelo	MSE
Autoregressive integrated moving average (ARIMA)	0.020
Modelo Híbrido de Zhang [21]	0.017
Modelo Híbrido de Aladag [11]	0.009
Modelo Híbidro de Khashei [13]	0.006
Modelo Proposto	**0.004**

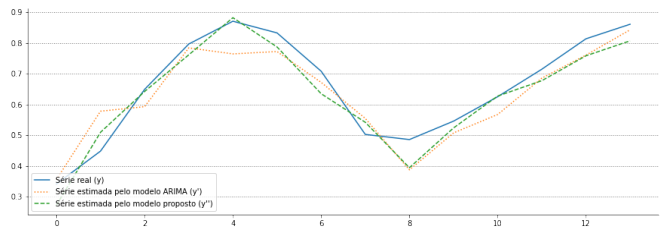

Fig. 4. Representações do conjunto de dados: *Canadian lynx forecasting*, sua estimação ARIMA e a estimação do modelo proposto.

V. Conclusões

Os resultados numéricos da Tabela 1 mostram os modelos híbridos em geral superam o *Autoregressive integrated moving average* (ARIMA) para o conjunto de dados analisado, o que sugere que há melhores chances de modelar diferentes padrões nos dados combinando variados modelos. Além disso, de acordo com os resultados obtidos, percebe-se que a metodologia híbrida proposta melhora a precisão em relação aos modelo de associação de [21] e [11], e até para modelos mais complexos envolvendo o mapeando de funções para combinação da componente linear e não linear, como o modelo de [13]. Tal metodologia indica uma melhoria no MSE de 76,47%; 55,55% e de 33,34% em relação à metodologia híbrida de [21] , [11] e [13], respectivamente.

Referências

[1] M. Kaushik e A. K. Giri, *Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning amp; Deep Learning Techniques*, 2020. DOI: 10.48550/ARXIV.2002.10247. endereço: https://arxiv.org/abs/2002.10247.

[2] P. A. Morettin e W. d. O. Bussab, *Estatística básica*. Saraiva, 2004.

[3] B. G. E. P., G. M. Jenkins, G. C. Reinsel e G. M. Ljung, "Time Series Analysis: Forecasting and Control," em Wiley, 2016.

[4] M. R. Silvestre e M. I. S. Bezerra, "Modelos decomposiÃ§Ãtemporal e de regressÃharmÃ´nica: uma comparaÃ§Ãpara a sÃmensal da temperatura mÃnima e mÃ¡xima mÃde presidente prudente (sp)," pt, *Revista Brasileira de Meteorologia*, v. 30, pp. 457–466, dez. de 2015, ISSN: 0102-7786. endereço: http://old.scielo.br/scielo.php?script=sci_arttext&pid=S0102-77862015000400457&nrm=iso.

[5] J. Fattah, L. Ezzine, Z. Aman, H. Moussami e A. Lachhab, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management*, v. 10, p. 184 797 901 880 867, out. de 2018. DOI: 10.1177/1847979018808673.

[6] J. A. Putra, F. Basbeth e S. Bukhori, "Sugar Production Forecasting System in PTPN XI Semboro Jember Using Autoregressive Integrated Moving Average (ARIMA) Method," em *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2019, pp. 448–453. DOI: 10.23919/EECSI48112.2019.8977010.

[7] L. Cao e F. Tay, "Financial Forecasting Using Support Vector Machines," *Neural Computing and Applications*, v. 10, pp. 184–192, mai. de 2001. DOI: 10.1007/s005210170010.

[8] A. Kumar e M. Chaudhry, "Review and Analysis of Stock Market Data Prediction Using Data mining Techniques," em *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, 2021, pp. 1–10. DOI: 10.1109/ISCON52037.2021.9702498.

[9] R. G. D. Ángel, "Financial time series forecasting using Artificial Neural Networks," *Revista Mexicana de Economía y Finanzas Nueva Época REMEF*, v. 15, n. 1, pp. 105–122, 2020, ISSN: 2448-6795. DOI: 10.21919/remef.v15i1.376. endereço: https://www.remef.org.mx/index.php/remef/article/view/376.

[10] Y. Tian e L. Pan, "Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network," em *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015, pp. 153–158. DOI: 10.1109/SmartCity.2015.63.

[11] C. H. Aladag, E. Egrioglu e C. Kadilar, "Forecasting nonlinear time series with a hybrid methodology," *Applied Mathematics Letters*, v. 22, n. 9, pp. 1467–1470, 2009, ISSN: 0893-9659. DOI: https://doi.org/10.1016/j.aml.2009.02.006.

[12] H. do Nascimento Camelo, P. S. Lucio, J. B. V. L. Junior e P. C. M. de Carvalho, "Métodos de Previsão de Séries Temporais e Modelagem Híbrida ambos Aplicados em Médias Mensais de Velocidade do Vento para

Regiões do Nordeste do Brasil," *Revista Brasileira de Meteorologia*, v. 32, n. 4, pp. 565–574, 2017, ISSN: 1982-4351. DOI: 10.1590/0102-7786324005. endereço: https://www.scielo.br/j/rbmet/a/XCmF6ssBPJvfDVJgWZVDF6t/abstract/?lang=pt#.

[13] M. Khashei e M. Bijari, "A New Hybrid Methodology for Nonlinear Time Series Forecasting," *Modelling and Simulation in Engineering*, v. 2011, jan. de 2011. DOI: 10.1155/2011/379121.

[14] V. L. Fava, "Metodologia de Box-Jenkins para modelos univariados," em Atlas, 2000.

[15] R. Pindyck e D. Rubinfeld, *Econometria: modelos & previsões*. Elsevier, 2004, ISBN: 9788535213430.

[16] P. Morettin e C. Toloi, *Análise de séries temporais: modelos lineares univariados*. BLUCHER., 2018, ISBN: 9788521213529. endereço: https://books.google.com.br/books?id=UwC5DwAAQBAJ.

[17] F. B. Fitch, "Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics, vol. 5 (1943), pp. 115–133.," *Journal of Symbolic Logic*, v. 9, n. 2, pp. 49–50, 1944. DOI: 10.2307/2268029.

[18] E. Cadenas e W. Rivera, "Short term wind speed forecasting in La Venta, Oaxaca, México, using artificial neural networks," *Renewable Energy*, v. 34, n. 1, pp. 274–278, 2009, ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2008.03.014. endereço: https://www.sciencedirect.com/science/article/pii/S0960148108001171.

[19] G. Zhang, B. Eddy Patuwo e M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting*, v. 14, n. 1, pp. 35–62, 1998, ISSN: 0169-2070. DOI: https://doi.org/10.1016/S0169-2070(97)00044-7. endereço: https://www.sciencedirect.com/science/article/pii/S0169207097000447.

[20] N. H. Abel, "Démonstration de l'impossibilité de la résolution algébrique des équations générales qui passent le quatrième degré," em *Oeuvres complètes de Niels Henrik Abel: Nouvelle édition*, L. Sylow e S. Lie, ed., sér. Cambridge Library Collection - Mathematics. Cambridge University Press, 2012, vol. 1, pp. 66–94. DOI: 10.1017/CBO9781139245807.008.

[21] G. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, v. 50, pp. 159–175, 2003, ISSN: 0925-2312. DOI: https://doi.org/10.1016/S0925-2312(01)00702-0. endereço: https://www.sciencedirect.com/science/article/pii/S0925231201007020.