

**PROPOSTA DE UM MODELO HÍBRIDO BASEADO NO MODELO
EPIDEMIOLÓGICO
SIR E INTELIGÊNCIA COMPUTACIONAL PARA PREVISÃO DA
COVID-19 NO
MARANHÃO**

Pablo Francisco Melo Bezerra¹; Selmo Eduardo Rodrigues Júnior²

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

Resumo

Esta pesquisa procura desenvolver um modelo híbrido para a previsão do número de casos de infectados pelo COVID-19 no estado do Maranhão utilizando o modelo SIRD (*Susceptible-Infected-Recovered-Death*) e/ou suas variantes em conjunto com um modelo de previsão baseado em séries temporais. Foi observado que o modelo SIRD possui parâmetros que não podem ser encontrados nos bancos de dados analisados, sendo os mesmos importantes para o modelo híbrido indicando características importantes da doença. Foi determinado também que parâmetros possuem uma variação no tempo para que o sistema do conjunto de EDO do SIRD acompanhe o gráfico de infectados do conjunto de dados. Com isso o objetivo do modelo de previsão recebe como entrada os dados obtidos e como saída proverá o valor dos parâmetros para um certo período de tempo adiante, aplicando os mesmos no modelo SIRD gerando assim o número de infectados durante os próximos dias. Como resultado inicial, foram obtidos os parâmetros para horizontes de previsão de tempo de 7, 14 e 30 dias com uma margem de erro relativamente baixa, e logo após foi realizada a avaliação entre os diferentes modelos usados para indicar o mais efetivo.

Palavras-chaves: Covid-19; Modelo epidemiológico SIRD; Modelo de Previsão Híbrido; Métodos de otimização matemática; ARIMA; LSTM.

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

Introdução

Desde o início da pandemia do COVID-19, vários estudos têm aplicado o modelo SIRD (Suscetíveis, Infectados, Recuperados, Falecidos), desenvolvido por Kermack, McKendrick e Walker (1927), para prever a propagação da doença e avaliar a eficácia de intervenções de saúde pública. A ampla flexibilidade e aplicações do modelo se estendem em diversos campos como a avaliação de políticas e de intervenções, nos estudos de Gupta et al. (2020) e Morato et al. (2020), até em análises envolvendo os casos não reportados Bastos et al. (2021).

Com isso foi utilizado técnicas para a previsão do modelo SIRD para estimar a propagação futura do Covid-19, que envolvem o uso de redes neurais (ANN's) como o LSTM, usado por Lai e Pai (2023) e Bousquet et al. (2021), e o modelo estatístico ARIMA, apresentado nos trabalhos Brockwell e Davis (2016) e Fatimah et al. (2022), com o objetivo de identificar o método de previsão mais efetivo e seu impacto utilizando a configuração híbrida que mescla as características dos dois modelos impactando nas decisões tomadas por hospitais e municípios que estabelecem medidas para o tratamento e prevenção da população.

Metodologia

Utilizou-se um ambiente de notebooks Jupyter Notebook que não requer configuração, pois já possui por padrão a linguagem Python e é executado na nuvem disponibilizada pelo Google, permitindo o acesso de várias redes de computadores. A base de dados sobre o COVID-19 no Maranhão foi extraída dos boletins diários do Portal da Secretaria de estado da saúde (SES, 2023).

O modelo de equações diferenciais ordinárias (EDO) utilizado neste projeto foi a variação *Susceptible Infected Recovered Deceased* (SIRD). Esse modelo tem por objetivo utilizar os parâmetros da EDO e os dados iniciais para modelar o andamento da doença. O modelo pode ser descrito pelo conjunto de equações:

Figura 1 - Equação SIRD

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

$$\begin{cases} \frac{dS(t)}{dt} = -\beta \frac{I(t)S(t)}{N} \\ \frac{dI(t)}{dt} = \beta \frac{I(t)S(t)}{N} - \gamma I(t) - \mu I(t) \\ \frac{dR(t)}{dt} = \gamma I(t) \\ \frac{dD(t)}{dt} = \mu I(t) \end{cases}$$

Onde:

- S(t): Suscetíveis atualmente;
- I(t): Infectados atualmente;
- R(t): Recuperados atualmente;
- D(t): Mortos atualmente;
- N : Quantidade da população;
- β : Taxa de infecção;
- γ : Taxa de recuperação;
- μ : Taxa de mortalidade

Logo realizou-se a extração dos parâmetros internos do modelo SIRD utilizando o algoritmo de Levenberg-Marquardt com auxílio do método de Runge-Kutta de quarta ordem para as EDO's.

Figura 2 - Aplicação do Runge-Kutta de quarta ordem.

$$\begin{cases} x = [S, I, R, D]^T \\ F = \left[-\frac{\beta SI}{N}, \frac{\beta SI}{N} - I(\gamma + \mu), \gamma I, \mu I \right]^T \\ \frac{dx}{dt} = F(t, x) \\ k_1 = F(t_n, x_n) \\ k_2 = F\left(t_n + \frac{h}{2}, x_n + \frac{hk_1}{2}\right) \\ k_3 = F\left(t_n + \frac{h}{2}, x_n + \frac{hk_2}{2}\right) \\ k_4 = F(t_n + h, x_n + hk_3) \\ x_{n+1} = x_n + \frac{h}{6} (k_1 + 2k_2 + 2k_3 + hk_3) \end{cases}$$

Onde:

- x_n : Solução aproximada da EDO em t_n ;
- h: Diferença entre t_n e t_{n+1} ;
- k_1, k_2, k_3, k_4 : Calculados a partir das derivadas.

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

Figura 3 - Algoritmo de Levenberg-Marquardt

$$(\beta, \gamma, \mu) = \underset{\beta, \gamma, \mu}{\operatorname{argmin}} \sum_{i=1}^m \left[x_{n+1,i} - \left(x_{n,i} + \frac{h}{6} (k_1 + 2k_2 + 2k_3 + hk_3) \right) \right]^2$$

Com os dados dos parâmetros obtidos podemos utilizar os mesmos como material de aprendizado/estimação para alguns modelos de previsão para diversificar a eficácia geral dos modelos o conjunto de dados foi dividido em 3 partes arbitrárias na qual estão apresentadas na Tabela 1. Com isso é feita uma análise em cada uma dessas janelas de tempo para identificar se o modelo responde bem nesses diferentes instantes com características diferentes. Os horizontes de previsão também se alteraram durante as análises nas janelas de tempo, sendo escolhidos os horizontes de 7, 14 e 30 dias totalizando diferentes instantes de tempo com características diferentes para verificar a consistência do modelo de previsão, onde no qual os dados de teste do modelo estão apresentados na Tabela 2.

Tabela 1 - Divisão dos períodos de tempo.

Períodos	Data inicial	Data final
Dataset 1	01/10/2020	24/07/2021
Dataset 2	01/10/2020	17/05/2022
Dataset 3	01/10/2020	11/03/2023

Tabela 2 - Dados de treino para cada período e horizonte de previsão

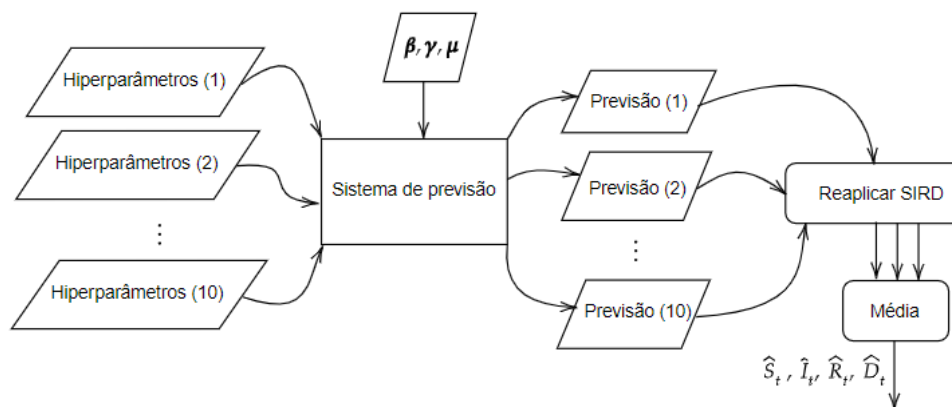
Dataset	Período	Data Inicial	Data Final
Dataset 1	7	17/07/2021	24/07/2021
	14	10/07/2021	
	30	24/06/2021	
Dataset 2	7	10/05/2022	17/05/2022
	14	03/05/2022	
	30	17/04/2022	
Dataset 3	7	04/03/2023	11/03/2023
	14	25/02/2023	
	30	09/02/2023	

Visando minimizar as variações de erros na previsão foram realizadas previsões com os modelos utilizando 10 hiperparâmetros, escolhidos de forma arbitrária, internos dos modelos para gerar previsões nas quais foram feitas a média dos resultados para gerar uma previsão mais consistente.

Figura 5 - Arquitetura dos modelos de previsão

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.



Durante a modelagem de previsão usou-se o modelo ARIMA, segundo TSAY (2005), é um modelo de previsão constituído por três componentes principais AR, I e MA onde a parte auto regressiva (AR) captura a relação linear entre os valores passados da série e o seu valor atual. A parte de média móvel (MA), por outro lado, modela o efeito dos erros passados na série. Já a parte integrada (I) é responsável por remover tendências e outros padrões não estacionários da série, tornando-a estacionária.

Figura 4 - Equação ARIMA

$$y_t^d = c + \phi_1 y_{t-1}^d + \phi_2 y_{t-2}^d + \dots + \phi_p y_{t-p}^d + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Onde:

- y_t : Valor da série temporal;
- c : É uma constante de média;
- ϕ : Parâmetro auto regressivo (AR);
- θ : Parâmetro de média móvel (MA);
- p, d, q : Níveis para auto regressão, integração e média móvel;
- ϵ_t : Erro aleatório no momento t .

Foram feitos 3 modelos diferentes para prever β , γ e μ utilizando uma biblioteca que consegue estimar os melhores hiperparâmetros do modelo ARIMA automaticamente tendo como métrica interna o valor obtido a partir do critério de avaliação de Akaike (AKAIKE, 1974). A arquitetura geral da rede neural artificial Long-Short Term Memory, mais conhecido como LSTM (HOCHREITER; SCHMIDHUBER, 1997) que apresenta a capacidade de trabalhar com dados de forma sequencial, foi dividida em duas abordagens de previsão, chamadas de Single Step e Multi Step, onde os modelos possui 3 entradas (β , γ , μ), onde as

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

mesmas receberem dados de 7 dias anteriores buscando aprimorar as previsões do modelo com dados mais relevantes. A abordagem Single Step possui a configuração de realimentação dos dados previstos no modelo para realizar uma previsão de passo único, ou seja um dia de cada vez utilizando realimentação, já na abordagem utilizando o Multi Step o modelo realiza uma previsão em múltiplos instantes de tempo de forma direta.

Foram utilizadas métricas de avaliação entre os dados para estimar a qualidade da previsão, os métodos escolhidos foram o Erro Médio Absoluto (MAE), a Raiz do Erro Médio Quadrático (RMSE) e o Erro Médio Percentual Absoluto (MAPE).

Figura 5 - Equações de erros MAE, RMSE, MAPE

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- N : Número de pontos de dados;
- y_i : Dados reais no ponto i;
- \hat{y}_i : Estimação de dados no ponto i.

No qual as métricas avaliam com critérios diferentes em que há penalidades que são de formas absolutas, de acordo com o tamanho do erro e em escala percentual para MAE, RMSE e MAPE respectivamente. Logo após foi feita a comparação dos modelos utilizando uma previsão direta, sem modelagem matemática, em relação ao procedimento do modelo SIRD avaliando assim o desempenho do modelo SIRD para a previsão. Para identificar o modelo mais consistente foi feita a média entre os diferentes modelos em cada horizonte de previsão utilizando a Figura 6 validando a eficácia do modelo como um todo, focando nos suscetíveis, infectados, recuperados e mortos de forma igualitária.

Figura 6 - Equação de critério de erro

$$ERRO_{SIRD} = \frac{ERRO_S + ERRO_I + ERRO_R + ERRO_D}{4}$$

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

RESULTADOS

Sobre o comportamento do modelo SIRD padrão, em relação aos dados reais, foi necessária uma modificação na forma em que os parâmetros do modelo epidemiológico (β , γ , μ) fossem variáveis no tempo, levando em conta a flexibilidade e comportamentos físicos da doença ao longo do tempo para que as equações pudessem obter exatamente os mesmos valores que o conjunto de dados, utilizando assim o algoritmo de Levenberg-Marquardt para determinar esses parâmetros implícitos.

Após feita a previsão dos parâmetros dos parâmetros utilizando os os sistemas de previsão do Auto-ARIMA, LSTM Single-Step e Multi-Step e aplicando no modelo SIRD obtivemos as previsões e com elas foram realizadas as avaliações de erros pertinentes utilizando os critérios de avaliação discutidos.

Durante os erros apresentados foram realizadas as comparações do uso do modelo SIRD utilizando a previsão dos parâmetros com uma previsão direta realizada sem nenhum princípio matemático, e foi concluído que para o modelo ARIMA os resultados são bem próximos, e utilizando o LSTM os erros foram exorbitantes, uma das possíveis causas para esse erro nestes LSTM de forma direta pode ser *Overfitting* onde o modelo durante sua fase de treinamento possui erros baixos porém durante o teste os erros são altos, devido a forma em que o sistema em que os modelos LSTM foram construídos há três possibilidades, *Overfitting*, o modelo pode não aprender corretamente durante o treino ou a quantidade de entradas pode ser muito baixa para modelar os dados desejados, chamado de *Underfitting*.

Tabela 3 - Erros para cada modelo

Sistemas de previsão	Modelo aplicado	MAE	RMSE	MAPE (%)
LSTM-SINGLE	SIRD	18,63	26,10	2,31
ARIMA		13,63	17,83	4,2
LSTM-SINGLE	Forma direta	83.516,5	83.584,34	198
ARIMA		34,06	44,05	5,3

Após as medidas de erros em cada período, foi feita a média entre os 3 instantes de tempo para identificar o modelo com uma maior consistência na maior parte do conjunto

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

de dados, onde foi observado que para os modelos LSTM e ARIMA o horizonte de evento com menor erro foi o de 7 dias assim sendo escolhidos os dois melhores de cada técnica de previsão que foram os modelos ARIMA e LSTM-SINGLE. É observado que apesar do modelo ARIMA (SIRD) estar desempenhando melhor na Tabela 3, o seu erro em percentual é maior em relação ao LSTM-SINGLE (SIRD) o que indica que o modelo teve maiores erros em dados que possuem uma escala alta de valores, como os suscetíveis na casa dos milhões, mesmo que suas outras métricas de erro tenham se apresentado levemente mais eficazes em relação ao LSTM Single, o modelo ARIMA é quase duas vezes inferior quanto ao erro percentual para todas as condições do sistema de equações SIRD.

CONCLUSÃO

Por fim durante a comparação entre o modelo SIRD e utilizando a previsão direta na Tabela 3, há uma limitação de utilizar os modelos LSTM para uma previsão direta, no qual os dados além de mostrarem uma previsão totalmente fora dos padrões, a previsão também pode ser afetada por não haver as restrições que o modelo SIRD propõe utilizando seus parâmetros físicos (β , γ , μ) evidenciando que o modelo SIRD modela de forma mais efetiva os comportamentos da doença, com isso o uso do modelo híbrido e que sua contribuição tem um impacto efetivo nos modelos de previsão abordados podendo ser útil para antecipar o comportamento futuro de doenças que possuem uma volatilidade considerável. Como propostas para o futuro, as metodologias podem ser testadas para outros conjuntos de dados, pois cada região é impactada de uma forma diferente pela doença, e validando a metodologia empregada neste projeto, também pode-se modificar os métodos utilizados realizando a substituição do método de otimização para encontrar os parâmetros, utilizando algoritmos genéticos ou *particle swarm optimization*, também é válida a mudança no modelo para previsão de séries temporais utilizando algoritmos baseados em *Neuro-Fuzzy* ou *Reinforcement Learning*.

REFERÊNCIAS

AKAIKE, H. **A new look at the statistical model identification.** *IEEE Transactions on Automatic Control*, Institute of Electrical and Electronics Engineers (IEEE), v. 19, n. 6, p. 716–723, dez. 1974. Disponível em: <<https://doi.org/10.1109/tac.1974.1100705>>.

BASTOS, S. B. et al. **The covid-19 (sars-cov-2) uncertainty tripod in brazil: Assessments**

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.

on model-based predictions with large under-reporting. Alexandria Engineering Journal, v. 60, p. 4363–4380, 2021.

BOUSQUET, A. et al. **Deep learning forecasting using time-varying parameters of the sird model for covid-19.** Scientific Reports, n. 12, p. 3030, 2021.

BROCKWELL, P. J.; DAVIS, R. A. **Introduction to Time Series and Forecasting.** 3. ed. [S.l.]: Springer, 2016. ISBN 978-3-319-29852-8.

FATIMAH, B. et al. **A comparative study for predictive monitoring of covid-19 pandemic.** Applied Soft Computing, v. 122, 2022.

GUPTA, R. K. et al. **Modelling the covid-19 epidemic and implementation of population-wide interventions in italy.** Nature Medicine, v. 26, p. 855–860, 2020.

HOCHREITER, S.; SCHMIDHUBER, J. **Long short-term memory.** Neural Computation, MIT Press - Journals, v. 9, n. 8, p. 1735–1780, nov. 1997. Disponível em: <<https://doi.org/10.1162/neco.1997.9.8.1735>>.

IBGE. **INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA.** 2020. Disponível em: <<https://cidades.ibge.gov.br/brasil/ma/panorama>>. Acesso em: 29/03/2023.

KERMACK, W. O.; MCKENDRICK, A. G.; WALKER, G. T. **A contribution to the mathematical theory of epidemics.** The Royal Society Publishing, v. 115, p. 700 – 721, 1927.

LAI, J.-P.; PAI, P.-F. **A dual long short-term memory model in forecasting the number of COVID-19 infections.** Electronics, MDPI AG, v. 12, n. 3, p. 759, fev. 2023. Disponível em: <<https://doi.org/10.3390/electronics12030759>>.

MORATO, M. M. et al. **A parametrized nonlinear predictive control strategy for relaxing covid-19 social distancing measures in brazil.** PubMed, v. 124, p. 197–214, 2020.

SES. **Portal da Secretaria de Estado da Saúde.** 2023. Disponível em: <<https://www.saude.ma.gov.br/>>. Acesso em: 31/05/2023.

TSAY, R. S. **Analysis of Financial Time Series.** 2. ed. [S.l.]: Wiley Interscience, 2005. ISBN 13 978-0-471-69074-0.

¹ Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: pablomelo@acad.ifma.edu.br.

² Professor do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; Titulação: Dr.; E-mail: selmo.junior@ifma.edu.br.