Modelo de regressão *spline*, com efeitos mistos e erros autorregressivos de médias móveis, aplicado aos dados da Covid-19 nos estados do Sul e Sudeste do Brasil

Marcos Antonio Alves Pereira[1]*, Cibele Maria Russo Novelli[2], Mileno Tavares Cavalcante[3]

[1] marcos.stats@gmail.com. Instituto de Formação de Educadores, Universidade Federal do Cariri (UFCA), Brasil
[2] cibele@icmc.usp.br. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP), Campus São Carlos, Brasil
[3] milenoc@yahoo.com. Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro (UFRJ), Brasil
* autor correspondente

Resumo
Este trabalho tem como objetivo apresentar um método para ajustar dados longitudinais de casos confirmados e acumulados de Covid-19, proporcionalmente ao número de habitantes dos estados das regiões Sul e Sudeste do Brasil, considerando o tempo como variável explicativa. Com o modelo proposto, é possível fazer previsões de novos casos da doença como forma de oferecer suporte para gestores públicos e privados na elaboração e planejamento de estratégias para prevenir ou atenuar impactos sociais e econômicos de doenças de propagação viral coma a Covid-19 e outras doenças com propagação similar. Considerou-se modelos de regressão *spline* com efeitos mistos, úteis para ajuste de dados correlacionados que não possuam relação linear, e modelos autorregressivos e de médias móveis (ARMA) para os resíduos, uma vez que estes apresentaram essa característica, além de serem estacionários. Utilizou-se o modelo de regressão *spline* cúbica com efeitos mistos para calcular as taxas de crescimento e para predição de observações futuras do número acumulado de infectados pela Covid-19 nos sete estados estudados. Os resultados obtidos demonstraram boa concordância entre os dados ajustados com o modelo e os dados observados para todos os estados analisados. As previsões para os estados de São Paulo, Espírito Santo e Minas Gerais apresentaram os menores valores absolutos dos desvios relativos entre valores preditos e observados.
Palavras-chave: autorregressivos de médias móveis; covid-19 no Brasil; dados longitudinais; efeitos mistos; regressão *spline*.

Abstract
The main objective of this work is to present a method to fit longitudinal data on confirmed and total number cases of Covid-19, proportionally to the number of inhabitants of the states of South and Southeast regions of Brazil, considering time as an explanatory variable. With the proposed model, it is possible to make predictions of new cases of the disease as a way of offering support to help public and private managers in the elaboration and planning of strategies to prevent or mitigate social and economic impacts of viral propagation diseases such as Covid-19 and other diseases with similar propagation. We worked with spline regression models in the context of mixed-effects models, which are useful for fitting nonlinear correlated data, and autoregressive-moving average modelling (ARMA) for stationary and serially correlated residuals, as it was the case in our analysis. We made use of the cubic spline regression model with mixed-effects to compute the growth rates and to predict future observations of the accumulated number of infected with Covid-19 in the seven states studied. The results obtained showed good agreement between the data fitted with the model and the observed data for all states analyzed. The forecasts for the states of São Paulo, Espírito Santo and Minas Gerais showed the lowest absolute values of relative deviations between predicted and observed values.
Keywords: *autoregressive-moving average; Covid-19 in Brazil; longitudinal data; mixed-effects; spline regression.*

1 Introdução

A eclosão da epidemia de uma doença respiratória na cidade de Wuhan, em dezembro de 2019 na China, e o crescente número de casos, óbitos e de países afetados, levou a comunidade internacional a retomar possíveis alertas sobre o risco de uma pandemia, fato declarado pela Organização Mundial da Saúde (OMS) em março de 2020, devido ao surgimento de mais de cem mil casos ao redor do mundo (GARCIA; DUARTE, 2020). A doença, denominada de Covid-19 (Doença por Coronavírus 2019) pela OMS, é caracterizada por um tipo de pneumonia viral grave, cujo vírus causador foi inicialmente denominado 2019-nCoV e, posteriormente, de SARS-CoV-2 – Síndrome Respiratória Aguda Grave de Coronavírus 2 (ASHOUR *et al*., 2020).

Em fevereiro de 2020, um paciente com problemas no trato respiratório foi atendido no Hospital Albert Einstein em São Paulo e diagnosticado com SARS-CoV-2. Esse paciente, brasileiro e residente na cidade de São Paulo, foi o primeiro caso confirmado da doença registrado no Brasil e, recentemente, havia chegado de uma viagem à Itália, um dos epicentros da Covid-19 na Europa (CRODA; GARCIA, 2020; RODRIGUEZ-MORALES *et al*., 2020). Segundo dados do Ministério da Saúde, após um declínio ou "achatamento da curva" do número acumulado de infectados, ou seja, de casos confirmados da doença nos meses de setembro e outubro de 2020, ocorreu uma escalada no número de casos a partir de novembro de 2020. O primeiro óbito foi notificado em meados do mês de março de 2020 (FRANÇA *et al*., 2020) e até o dia 07 de julho de 2023 já haviam sido computados mais de 37 milhões de casos confirmados e mais de 700 mil óbitos, e o estado de São Paulo apresentava o maior número absoluto de casos e óbitos.

Com o crescimento do número de casos confirmados e de óbitos por Covid-19 espalhados pelo mundo, surgiu a necessidade do tratamento desses dados, o que se tornou um desafio para muitos pesquisadores, mesmo com a limitação na qualidade dos dados, principalmente por conta da subnotificação devido à baixa testagem. Algumas propostas foram apresentadas como Salgotra, Gandomi e Gandomi (2020), que fizeram ajustes por meio de modelos de séries temporais baseados na programação genética evolutiva para analisar dados de casos confirmados de Covid-19 na Índia, enquanto Gomes, Monteiro e Rocha (2020) utilizaram um modelo dinâmico compartimental do tipo SIR (Suscetíveis, Infectados e Removidos), cuja estrutura simples necessita da estimação de poucos parâmetros, numa tentativa de compreender melhor a dinâmica de espalhamento da Covid-19. Por meio do modelo SIRD (Suscetíveis, Infectados, Removidos e Mortos), Parro *et al.* (2021) propuseram uma versão modificada para descrever a dinâmica de uso do sistema de saúde com base nos casos notificados de Covid-19, com os indivíduos considerados susceptíveis como uma proporção da população total. O modelo SIRD modificado exibiu forte aderência aos dados para a maioria dos estados, sendo que para o Brasil como um todo ele apresentou um comportamento mais realista sobre a duração da epidemia que o modelo SIR. Tsallis e Tirnakli (2020) propuseram um modelo não linear, utilizado no mercado de ações, para fazer previsão de picos de Covid-19 pelo mundo. Utilizando dados para o estado de Mato Grosso do Sul, Saraiva e Sauer (2020) apresentaram modelos com curvas de crescimento para estudar o número de casos confirmados da Covid-19, por meio das funções exponencial, logística e Gompertz. O desenvolvimento de modelos estatísticos confiáveis para ajuste e previsão do número de casos e de óbitos da Covid-19 pode ser útil para orientar esforços para prevenir, combater, identificar regiões mais vulneráveis e planejar políticas públicas.

Alguns modelos de regressão usam funções suaves e flexíveis, conhecidas como *splines*, que são definidas como curvas formadas por partes de polinômios, úteis para o ajuste de dados que não possuem relação linear, como é o caso do número de notificações acumuladas da Covid-19. Essas partes de polinômios são unidas por meio de pontos distintos de observações (nós), sendo um polinômio para cada intervalo, cujo objetivo é modelar curvas mais complexas com polinômios mais simples. As principais vantagens de se usar *splines*, são a flexibilidade para o ajuste dos modelos quando comparado ao modelo de regressão linear ou polinomial, a capacidade de modelar comportamentos atípicos dos dados, e, uma vez obtidos os nós, a facilidade em fazer o ajuste, pois um modelo de regressão *spline* é linear nos parâmetros. Além disso, a comparação com modelos de crescimento que preveem apenas um pico da doença, torna a abordagem de *splines* ainda mais competitiva, já que o modelo em que as médias são ajustadas via *splines* não segue um padrão restritivo, ao contrário, se adapta a essas mudanças do comportamento médio dos dados.

Modelos com efeitos mistos, assim como apresentado em Pinheiro e Bates (2000), são úteis para ajustar dados longitudinais, medidas repetidas e multiníveis, levando-se em consideração a correlação entre as observações em cada grupo, e possuem aplicações em áreas do conhecimento como Economia, Engenharia e Farmacocinética (PEREIRA; RUSSO, 2019). Pode-se considerar um modelo de regressão *spline* misto como o modelo linear misto proposto por Laird e Ware (1982). Na medicina, o uso de modelos mistos na forma linear e não linear é frequente, principalmente quando no estudo há interesse de incorporar ao modelo a variabilidade causada pelos indivíduos, como verifica-se em Nordhausen, Oja e Parssinen (2015), que modelaram dados de progressão de miopia, e Grajeda *et al.* (2016), que explicaram a não linearidade de curvas de crescimento em crianças, com ambos utilizando modelos de regressão *spline* com efeitos mistos. Demertzis *et al*. (2020) apresentaram um método para modelagem e previsão do número acumulado de casos de Covid-19 na Grécia baseado na análise de rede complexa e modelo de regressão *spline*.

Pereira *et al.* (2020) propuseram o uso de métodos de inteligência artificial (*data driven approach*) para prever a dinâmica da pandemia de Covid-19 no Brasil. Para prever a propagação do vírus, os autores utilizaram uma rede neural *Modified Auto-Encoder* (MAE) treinado a partir de *clusters* de dados sobre a pandemia no Brasil e em outros países. Os *clusters* foram definidos por similaridade das respostas iniciais ao Covid-19, considerando informações para cada país como um todo e suas províncias/regiões. Tomando como referência métodos mais tradicionais de previsão para o número de casos diários confirmados de Covid-19, como o modelo SIR, o estudo conclui que a abordagem proposta apresentou previsões mais próximas aos valores reais em todos os estados brasileiros no intervalo de fevereiro a meados de maio de 2020. A partir de modelos de séries temporais e de modelos de regressão, Ribeiro *et al.* (2020) realizaram projeções para o número acumulado de casos diários confirmados de Covid-19 para dez estados brasileiros (AM, BA, CE, MG, PR, RJ, RN, RS, SC e SP) e três horizontes de projeção (1, 3 e 6 dias). Os resultados obtidos variaram de acordo com o modelo utilizado, o estado e o horizonte de previsão, com melhor performance geral para o modelo de regressão vetorial. Silva *et al.* (2021) propuseram o uso de métodos de aprendizagem de máquina para análise espaço-temporal do número de casos acumulados de Covid-19 para os municípios dos 27 estados do Brasil e do estado de Pernambuco, com os dados subdivididos por unidade da federação e município. As métricas utilizadas na avaliação foram o coeficiente de correlação e o erro quadrático médio relativo e indicaram que os modelos com melhor performance foram o modelo de regressão linear e *multilayer perceptron* para os municípios brasileiros, e o modelo de regressão linear para o estado de Pernambuco. Oliveira *et al.* (2022) utilizaram um modelo baseado em rede neurais com grafos (*Graph Neural Network*) para a dinâmica espaço-temporal do número diário de casos confirmados e de mortes por Covid-19 nos 5570 municípios brasileiros para o período de 25 de fevereiro de 2020 a 25 de janeiro de 2021. As principais vias de acesso aos municípios do Brasil por diferentes modais de transporte foram mapeadas pelos autores para captar a influência da situação da pandemia nos municípios vizinhos sobre um dado município, além do efeito de outras variáveis. O desempenho do modelo proposto para previsões 1 dia à frente mostrou-se superior em todos os casos analisados em relação a um conjunto *baseline* de modelos *naive forecaster*, alisamento exponencial e ARIMA (autorregressivo integrado de média móvel) segundo duas das três métricas adotadas.

A proposta desse trabalho é ajustar o número acumulado de infectados confirmados da Covid-19 nos estados do Sul e Sudeste do Brasil, relativizados para cada cem mil habitantes, utilizando um modelo de regressão *spline* cúbica com efeitos mistos, uma vez que os estados apresentam perfis com formatos semelhantes e curvas de crescimento distintas. Como forma de minimizar violações de suposições do modelo, considerou-se que os erros são autorregressivos de médias móveis (*autoregressive-moving average* - ARMA) e estacionários (PINHEIRO; BATES, 2000).

O objetivo do presente estudo é, com base em uma metodologia estatística semiparamétrica, modelar o número de casos confirmados de Covid-19, nos estados do Sul e Sudeste do Brasil, desde a décima semana epideomiológica de 2020 até o início da sexta semana epideomiológica de 2021. Há de se considerar que o período de análise é anterior ao início da vacinação em massa no Brasil e nos estados em questão, o que permite utilizar o tempo, em dias, como a única variável explicativa, sem a necessidade de considerar variáveis que capturem o efeito da vacinação sobre o número de casos

confirmados. Com esse estudo é possível modelar o número de casos de doença com propagação viral e permitir o planejamento de estratégias para prevenir e/ou atenuar os impactos sociais e econômicos e, assim, restringir o seu avanço. Assim, é possível proporcionar estimativas do número acumulado de infectados, otimizando os recursos.

Além dessa Introdução, a seção 2 desse artigo apresenta a metodologia utilizada, baseada no modelo de regressão *spline* com efeitos mistos e erros ARMA e na seção 3 discute-se a aplicação dessa metodologia aos dados da Covid-19 nos estados do Sul e Sudeste do Brasil. As considerações finais desse trabalho são apresentadas na seção 4.

2 Modelo de regressão *spline* com efeitos mistos e erros ARMA

Os modelos de regressão *spline* com efeitos mistos permitem ajustar dados considerando a correlação intragrupo. Um modelo de regressão *spline* com efeitos fixos e aleatórios, com o vetor de respostas observadas $y_i = (y_{i1}, \dots, y_{im})^\top$, pode ser escrito como

$$y_i = f_i(x_i, \beta) + Z_i b_i + \epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

em que a função $f_i(x_i, \beta)$ é conhecida e pode assumir algumas representações, $\beta = (\beta_0, \beta_1, \dots, \beta_{N+1})^\top$ é o vetor de parâmetros desconhecidos, ou efeitos fixos, x_i, de dimensão $m \times 1$, é um vetor de variáveis explanatórias, Z_i, de dimensão $m \times (N+2)$, é uma matriz de constantes conhecidas, os efeitos aleatórios $b_i = (b_{i0}, b_{i1}, \dots, b_{iN+1})^\top$ e os erros $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{im})^\top$ são variáveis aleatórias não correlacionadas e seguem distribuições normais multivariadas centradas no vetor nulo, com o vetor de erros modelado por processo ARMA. Os dados foram ajustados com o *software* R (R CORE TEAM, 2022), por meio da biblioteca *nlme* (PINHEIRO *et al*., 2022).

2.1 Spline linear

No modelo de regressão *spline* linear, a função $f_i(x_i, \beta)$ é definida por um conjunto de funções do primeiro grau (polinômio de grau um) unidos por meio de nós continuamente. Segundo Harrell Junior (2015), as *splines* têm sido usados, principalmente nas ciências físicas, para aproximar uma ampla variedade de funções. Dessa forma, considerando x_i um vetor ordenado de tamanho $m \times 1$, e definindo N nós, $N < m$, $\delta_1, \delta_2, \dots, \delta_N$, tem-se a representação

$$f_{ij}(x_{ij}, \beta) = \beta_0 + \beta_1 x_{ij} + \beta_2 (x_{ij} - \delta_1)_+ + \cdots + \beta_{N+1} (x_{ij} - \delta_N)_+, \tag{2}$$

em que

$$(x_{ij} - \delta_l)_+ = \begin{cases} (x_{ij} - \delta_l), & (x_{ij} - \delta_l) \geq 0 \\ 0, & (x_{ij} - \delta_l) < 0, \end{cases}$$

com $i = 1, \dots, n, j = 1, \dots, m, l = 1, \dots, N$.

2.2 Spline cúbica

No modelo de regressão *spline* cúbica, a função $f_i(x_i, \beta)$ é definida por uma curva constituída de polinômios do terceiro grau, que são unidos continuamente por meio de nós. De forma análoga dos *splines* lineares, tem-se a representação

$$f_{ij}(x_{ij}, \beta) = \beta_0 + \beta_1 x_{ij} + \beta_2 (x_{ij} - \delta_1)_+^3 + \cdots + \beta_{N+1} (x_{ij} - \delta_N)_+^3, \tag{3}$$

em que

$$(x_{ij} - \delta_l)_+^3 = \begin{cases} (x_{ij} - \delta_l)^3, & (x_{ij} - \delta_l) \geq 0 \\ 0, & (x_{ij} - \delta_l) < 0. \end{cases}$$

Na análise de dados da Covid-19 no Brasil, verificou-se a ausência de algumas premissas para os resíduos, como a suposição de normalidade e de independência ao longo do tempo. Dessa forma, apresentamos um modelo de regressão *spline* com efeitos mistos em que os erros são modelados por processos ARMA e estacionários. Conforme descrito por Box, Jenkins e Reinsel (1994) e Pinheiro e Bates (2000), esses processos são uma mistura de modelo autorregressivo e modelo de média móvel. Eles também são chamados de modelos *Box* e *Jenkins*.

2.3 Processo autorregressivo e de médias móveis p e q (ARMA(p, q))

Os modelos autorregressivos baseiam-se na ideia que o valor corrente da variável aleatória ϵ_k, observação em pontos de tempo inteiros obtida no tempo k, pode ser explicado por seus valores passados $\epsilon_{k-1}, \epsilon_{k-2,} \dots, \epsilon_{k-p}$, onde p denota o número de passos em direção ao passado necessários para se prever o valor da variável em questão. Um processo AR(p) pode ser representado por:

$$\epsilon_k = \sum_{r=1}^{p} \phi_r \epsilon_{k-r} + a_k, \tag{4}$$

onde ϕ_r, $r = 1, \dots, p$, são parâmetros reais do modelo, e o termo a_k adicionado é um ruído homocedástico, centrado em zero e independente das observações anteriores, ou seja, $\mathbb{E}(a_k) = 0$ e $\text{var}(a_k) = \sigma_a^2$.

Uma alternativa à descrição apresentada em (4), é o processo média móvel, onde a variável aleatória ϵ_k é função de uma combinação linear de ruídos brancos $a_{k-1}, a_{k-2,} \dots, a_{k-q}$, caracterizando então um processo MA(q). O mesmo pode ser representado como

$$\epsilon_k = \sum_{s=1}^{q} \theta_s a_{k-s} + a_k, \tag{5}$$

onde $\theta_s, s = 1, \dots, q$, são parâmetros reais do modelo.

O vetor de parâmetros de correlação $\rho = (\phi^\top, \theta^\top)^\top$, de dimensão $p + q$, é formado pela combinação de p parâmetros autorregressivos, $\phi = (\phi_1, \dots, \phi_p)^\top$, e q parâmetros de médias móveis, $\theta = (\theta_1, \dots, \theta_q)^\top$. Por convenção são denominados ARMA (p, q), assim como, por convenção, ARMA$(p, 0)$ = AR(p) e ARMA$(0, q)$ = MA(q), que são casos particulares do processo ARMA(p, q). Mais detalhes sobre famílias de estruturas de correlação estão descritos em Box Jenkins e Reinsel (1994).

Dessa forma, um modelo ARMA(p, q) é obtido da combinação de processo autorregressivo com um processo média móvel, sendo representado por:

$$\epsilon_k = \sum_{r=1}^{p} \phi_r \epsilon_{k-r} + \sum_{s=1}^{q} \theta_s a_{k-s} + a_k. \tag{6}$$

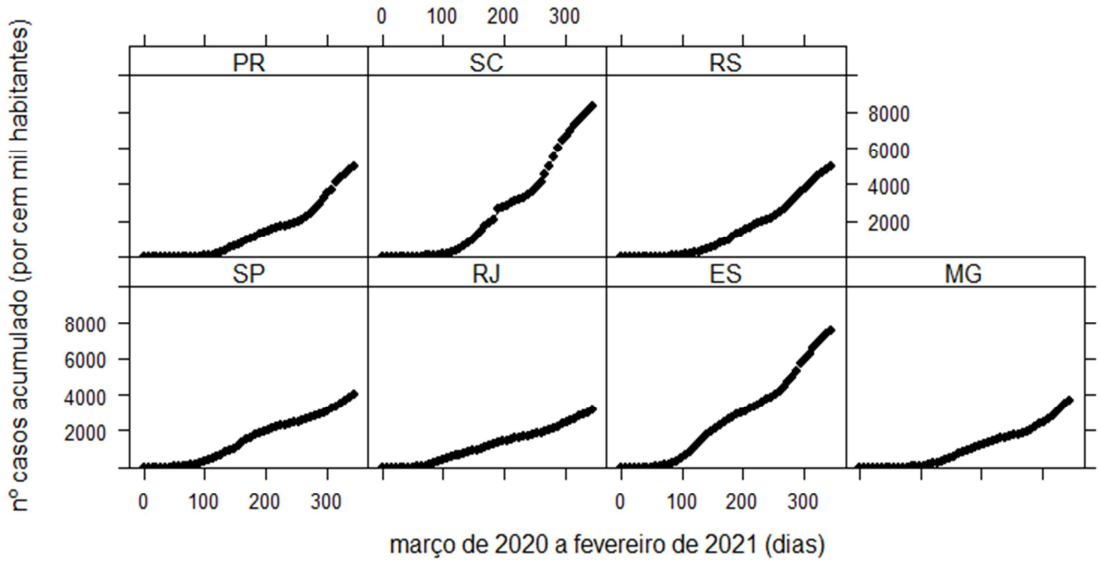
3 Aplicação do modelo proposto aos dados da Covid-19 nos estados do Sul e Sudeste do Brasil

O modelo que propomos é o modelo de regressão *spline* cúbica com efeitos mistos e erros ARMA foi aplicado aos dados da Covid-19 no Brasil obtidos no Guia de Vigilância Epidemiológica da Covid-19[1] do Ministério da Saúde e Secretaria de Vigilância Sanitária (SVS), assim como os dados das populações dos estados. Ajustou-se o modelo proposto com sete estados ($n = 7$), considerando o número acumulado de casos confirmados da infecção a cada cem mil habitantes, para cada semana desde 1º de março de 2020 até 07 de fevereiro de 2021, totalizando um período 344 dias ou 50 semanas de observação ($m = 50$). Os estados analisados foram Minas Gerais (MG), São Paulo (SP),

[1] Painel coronavírus: Disponível em: https://covid.saude.gov.br

Espírito Santo (ES), Rio de Janeiro (RJ), Paraná (PR), Santa Catarina (SC) e Rio Grande do Sul (RS), compreendendo, portanto, os estados das regiões Sudeste e Sul. Os dados observados estão apresentados na Figura 1.

Figura 1 – Gráfico do número acumulado de infectados confirmados da Covid-19 a cada cem mil habitantes, contra o tempo, para os estados do Sul e Sudeste do Brasil



Fonte: dados da pesquisa

Nota-se na Figura 1 que os estados com mais casos confirmados proporcionalmente ao número de habitantes são os estados de Santa Catarina e Espírito Santo, e o estado com menos casos a cada cem mil habitantes é o Rio de Janeiro.

No ajuste, o vetor de respostas observadas $y_i = (y_1, \dots, y_{50})^\top$, é o número acumulado de infectados confirmados a cada cem mil habitantes, e a variável explanatória é o número de dias após o início da contagem do tempo, que é a mesma para cada estado. Assim, tem-se que $t_i = t = (t_1, \dots, t_{50})^\top, t_{ij} = t_j$ e $f_{ij} = f_j$, $i = 1, \dots, 7$, $j = 1, \dots, 50$. A matriz de constantes Z_i também é a mesma para cada um dos sete estados, ou seja, $Z_i = Z$. Dessa forma, o modelo de regressão *spline* cúbica misto com erros ARMA pode ser escrito, conforme as Equações 1, 3 e 6, como:

$$y_{ij} = f_j(t_j, \beta) + Z_j b_i + \epsilon_{ij}, \quad i = 1, \dots, 7, \quad j = 1, \dots, 50, \tag{7}$$

$$\epsilon_{ij} = \sum_{r=1}^{p} \phi_r \epsilon_{i,j-r} + \sum_{s=1}^{q} \theta_s a_{i,j-s} + a_{ij}, \tag{8}$$

em que Z_j é a linha j da matriz Z.

Inicialmente ajustou-se um modelo com *spline* linear e o comparamos como modelo com *spline* cúbica. Como modelo linear, houve dificuldade em encontrar nós que apresentassem estatísticas *t* significativas e problemas para se obter convergência com a utilização da biblioteca *nlme* (PINHEIRO *et al*., 2022), além de valores elevados para o critério de informação de Akaike (AIC) (AKAIKE, 1974), dado pela Equação 9.

$$\text{AIC} = 2\,K - 2\ln(L) \tag{9}$$

em que *K* é o número de parâmetros do modelo e *L* é a verossimilhança estimada.

Assim, optou-se pelo modelo com *spline* cúbica por apresentar menor valor do AIC e de nós com estatísticas *t* significativas, além de não apresentar problemas de convergência computacional.

Para escolha dos nós, o conjunto de dados com 344 observações dos sete estados em 50 semanas foi dividido em *decis*, ou seja, dividido em dez partes de tamanhos iguais e utilizou-se a estatística *t* para selecionar os nós significativos. Isso resultou no modelo de regressão *spline* cúbica com 4 nós, em que as estimativas do intercepto (β_0) e do parâmetro β_1 não se mostraram significativas, segundo a estatística *t*, ou seja, $\beta_0 = \beta_1 = 0$. Assim, tem-se o modelo (Equações 7 e 8) em que

$$f_j(t_j, \beta) = \beta_2(t_j - \delta_1)_+^3 + \beta_3(t_j - \delta_2)_+^3 + \beta_4(t_j - \delta_3)_+^3 + \beta_5(t_j - \delta_4)_+^3. \quad (10)$$

Os efeitos aleatórios foram associados aos 4 nós, ao intercepto e a variável tempo, da seguinte forma

$$Z_j b_i = b_{i0} + b_{i1}t_j + b_{i2}(t_j - \delta_1)_+^3 + b_{i3}(t_j - \delta_2)_+^3 + b_{i4}(t_j - \delta_3)_+^3 + b_{i5}(t_j - \delta_4)_+^3. \quad (11)$$

No Quadro 1 estão apresentadas algumas estimativas de parâmetros do modelo (Equação 7), em que não foi considerada a estrutura de modelagem ARMA para os erros.

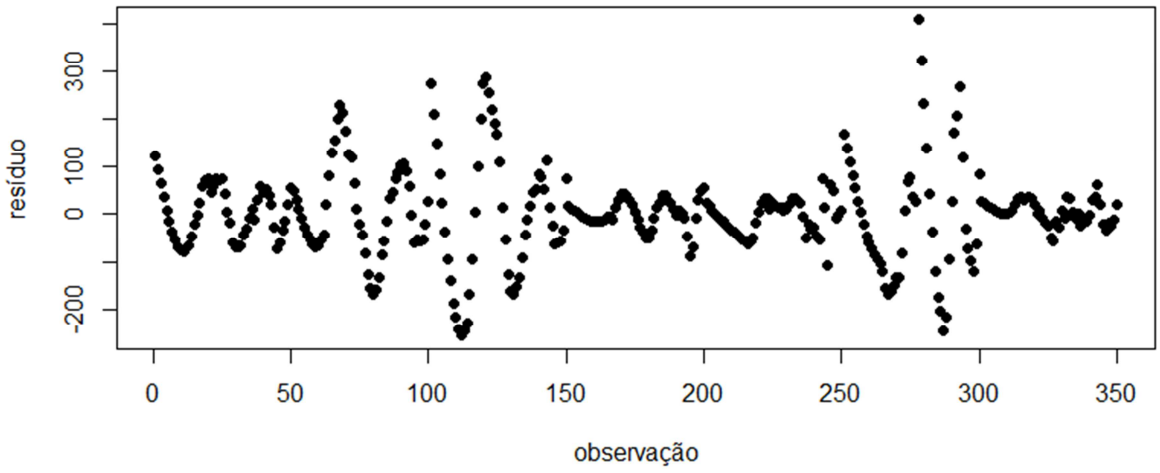
Quadro 1 – Parâmetros estimados, AIC e log-verossimilhança do modelo de regressão *spline* cúbica misto aplicado aos dados da Covid-19

Parâmetro	**Estimativa**	**Erro padrão**	***p*-valor**
β_2	0,001	< 0,001	< 0,001
β_3	-0,001	< 0,001	< 0,001
β_4	0,003	< 0,001	< 0,001
β_5	-0,006	0,001	< 0,001
AIC	4436,34		
Log-verossimilhança	-2192,17		

Fonte: dados da pesquisa

Uma forma de verificar a adequabilidade do modelo ajustado é por meio da análise dos resíduos. No gráfico da Figura 2 são apresentados os resíduos para cada observação do conjunto de dados.

Figura 2 – Gráfico dos resíduos do modelo de regressão *spline* cúbica misto ajustado aos dados da Covid-19



Fonte: dados da pesquisa

Observa-se no gráfico da Figura 2 a presença de padrões que sugerem autocorrelação nos resíduos. Dessa forma, para identificar a ordem do modelo ARMA que melhor ajusta os resíduos, utilizou-se a biblioteca *forecast* (HYNDMAN *et al.*, 2000), e segundo o critério AIC, obteve-se $p = 3$ e $q = 2$. Além disso, foi feito o teste de *Dickey-Fuller Aumentado*, por meio da biblioteca *tseries* (TRAPLETTI; HORNIK; LEBARON, 2019), concluindo-se que os resíduos são estacionários. Assim, o modelo (Equações 7 e 8) foi ajustado considerando o processo ARMA(3,2) para modelar os erros. Os resultados são apresentados no Quadro 2.

Quadro 2 – Parâmetros estimados, AIC e log-verossimilhança do modelo de regressão *spline* cúbica misto e erros ARMA(3,2) aplicado aos dados da Covid-19

Parâmetro	**Estimativa**	**Erro padrão**	***p*-valor**
β_2	0,001	< 0,001	< 0,001
β_3	-0,002	< 0,001	< 0,001
β_4	0,003	< 0,001	< 0,001
β_5	-0,005	0,001	< 0,001
ϕ_1	0,9257	-	-
ϕ_2	0,7895	-	-
ϕ_3	-0,8785	-	-
θ_1	0,0863	-	-
θ_2	-0,9052	-	-
AIC	3742,25		
Log-verossimilhança	-1840,12		

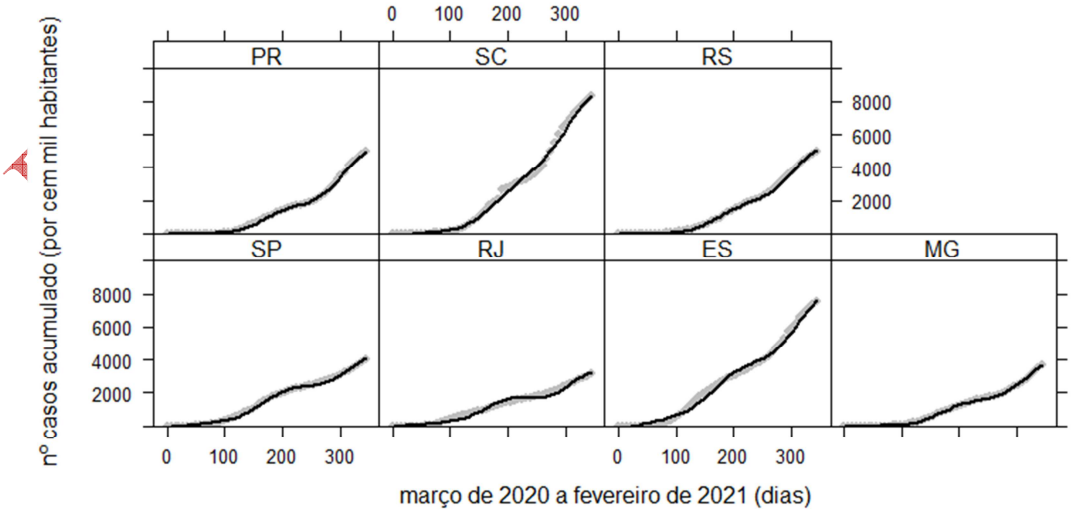
Fonte: dados da pesquisa

Nota-se uma melhora significativa no AIC, confirmando que um modelo mais adequado deve considerar uma estrutura para modelar os erros.

Segundo Schuster *et al.* (2022), o uso de modelos de *splines*, ao dividir a variável independente em múltiplos intervalos distintos, implica que o efeito de cada um desses intervalos sobre a variável resposta será estimado separadamente. No caso de *splines* cúbicas, o efeito médio da variável independente (tempo) sobre a variável resposta (número acumulado de infectados). As estimativas dos parâmetros β_2, β_3, β_4 e β_5 não possuem uma interpretação direta, devendo ser entendidas como parte da soma dos efeitos dos intervalos distintos do tempo sobre o número acumulado de infectados.

Os perfis ajustados com o modelo de regressão *spline* cúbica misto com erros ARMA(3,2), assim como os perfis observados, estão apresentados nos gráficos da Figura 3.

Figura 3 – Perfis ajustados e observados sob o modelo de regressão *spline* cúbica misto e erros ARMA(3,2) aplicado aos dados da Covid-19



Fonte: dados da pesquisa

Os gráficos da Figura 3 mostram os perfis ajustados em preto para cada estado estudado, além dos perfis observados em cinza. Observa-se uma boa concordância entre os dados ajustados e observados para o modelo apresentado.

Com os resultados do ajuste, pode-se calcular taxa de crescimento estimada do número acumulado de infectados usando a primeira derivada da curva ajustada para cada estado em relação ao tempo (PAINE *et al.*, 2012). No Quadro 3 estão apresentadas as estimativas das taxas de crescimento do número acumulado de infectados no último dia de observação, ou seja, estimativas para o dia 07 de fevereiro de 2021.

Quadro 3 – Estimativas da taxa de crescimento no número acumulado de infectados pela Covid-19 para 07 de fevereiro de 2021 obtidas com o modelo de regressão *spline* cúbica misto e erros ARMA(3,2)

Estado	**SP**	**RJ**	**ES**	**MG**	**PR**	**SC**	**RS**
Taxa de crescimento (%)	22,32	14,42	36,27	22,47	23,89	35,59	18,59

Fonte: dados da pesquisa

Observa-se que os estados do Espírito Santo e Santa Catarina possuem proporcionalmente as maiores taxas de crescimento no número de casos confirmados no dia 07 de fevereiro de 2021, e que o Rio de Janeiro apresenta a menor taxa. Nota-se também que há uma semelhança nas taxas de crescimento para os estados de São Paulo, Minas Gerais e Paraná.

3.1 Predição de observações futuras da variável resposta

Para predição de uma observação futura da variável resposta (y_i^+) no modelo de regressão *spline* com efeitos mistos e erros ARMA utilizou-se o estimador empírico de Bayes obtido pela esperança condicional de y_i^+ dado y_i para distribuições simétricas, conforme Lachos, Ghosh e Arellano-Valle (2010) e Pereira e Russo (2019). Dessa forma, foi avalidada a adequabilidade da metodologia proposta, considerando as observações $y_i' = (y_{51}, y_{52}, y_{53})^\top$, $i = 1, \ldots, 7$, referentes aos dias 14, 21 e 28 de fevereiro de 2021 como observações desconhecidas e, por isso, tratadas como observações futuras. As predições de $\widehat{y_i^+} = (\widehat{y_{51}^+}, \widehat{y_{52}^+}, \widehat{y_{53}^+})^\top, i = 1, \ldots, 7$, foram obtidas e, assim, foi calculado o valor absoluto do desvio relativo (VADR) por meio da expressão

$$\text{VADR} = \left| \frac{y_{ij} - \widehat{y_{ij}^+}}{y_{ij}} \right|, \tag{12}$$

com $i = 1, \ldots, 7$, $j = 51, 52, 53$, entre os valores observados e preditos das observações futuras. Os resultados estão apresentados na Quadro 4.

Quadro 4 - Valores absolutos dos desvios relativos entre valores observados e preditos para observações futuras no modelo de regressão *spline* cúbica misto com erros ARMA(3,2) aplicado aos dados da Covid-19 no Brasil

Medida	**Estado**						
	SP	**RJ**	**ES**	**MG**	**PR**	**SC**	**RS**
14 de fevereiro de 2021	0,008	0,024	0,002	0,016	0,001	0,013	0,018
21 de fevereiro de 2021	0,013	0,059	0,009	0,018	0,027	0,051	0,044
28 de fevereiro de 2021	0,024	0,097	0,037	0,031	0,071	0,105	0,104

Fonte: dados da pesquisa

Pode-se notar que os desvios são crescentes para todos os estados e que para o dia 28 de fevereiro, terceira medição predita, o estado de São Paulo apresenta o menor desvio, 2,4%, enquanto os estados que apresentam maiores desvios são os estados de Santa Catarina e Rio Grande do Sul, em torno de 10,5%. De maneira geral, os estados de São Paulo, Espírito Santo e Minas Gerais apresentaram valores preditos próximos dos valores observados. Em relação ao dia 14 de fevereiro de 2021, primeira medição, o estado do Paraná apresentou o menor desvio, 0,1%. Essas estimativas

sugerem que as predições de observações futuras mais próximas são mais realistas, com valores absolutos dos desvios relativos menores, devido a dinamicidade dos dados de contágio da Covid-19.

4 Considerações finais

Esse estudo contribui, de forma prática, para ajuste e predição do número acumulado de infectados pela Covid-19 nos estados das regiões Sul e Sudeste do Brasil, como também de outras doenças, por meio de modelos de regressão *spline* cúbica com efeitos mistos e erros ARMA. Com o estudo proposto, é possível dar suporte a gestores públicos e privados em tomadas de decisões para implementação de estratégias voltadas para a contenção de doenças com propagação viral, como a antecipação de ações em relação a gestão de recursos da saúde para o bem-estar da população.

O modelo apresentado faz boas previsões de observações futuras para a variável resposta, apesar de algumas predições com desvios razoavelmente elevados para um intervalo de três semanas, como os desvios computados para os estados de Santa Catarina e Rio Grande do Sul, conforme Quadro 4. Porém, para os estados de São Paulo, Espírito Santo e Minas Gerais, as predições foram mais realistas, com valores pequenos para os desvios relativos. Como uma forma de minimizar o surgimento desvios relativos elevados, ou seja, para melhorar a acurácia das previsões, o modelo, assim como proposto, pode ser ajustado, por exemplo, semanalmente com dados atualizados sobre a pandemia de Covid-19. Vale ressaltar que os dados de contágio da Covid-19 apresentam uma certa dinamicidade e, para o período analisado, essa dinâmica dependia de ações da população e de medidas tomadas pelos agentes públicos, por exemplo, no isolamento social e na disponibilização de exames de diagnóstico, causando impacto na qualidade das previsões. As taxas de crescimento do número de casos da doença, apresentadas no Quadro 3, são também ferramentas úteis para auxiliar na tomada de decisões, uma vez que é possível obter taxas de crescimento em momentos específicos.

Vale ressaltar que o presente estudo não considerou dados de todos os estados brasileiros, bem como o número acumulado de óbitos por Covid-19 como variável resposta, alternativa ao número acumulado de infectados registrados. Além disso, o modelo proposto fornece apenas estimações pontuais, e não estimações por intervalos, assim como em Schumacher *et al.* (2020), e também o modelo apresenta desvios relativos crescentes com o aumento do horizonte de previsão.

Como sugestão para futuros estudos para ajuste de dados da Covid-19, podem ser buscados modelos de séries temporais, como por exemplo, um modelo com a adição da componente de sazonalidade (SARIMA) e o modelo autorregressivo integrado de média móvel e entradas exógenas (ARIMAX), que se distingue do modelo proposto que faz uso da modelagem ARMA apenas para os resíduos. Assim como ajustes com modelos não lineares com efeitos mistos que considerem funções adequadas nos ajustes. Nesses modelos, podem ser consideradas também distribuições assimétricas com caudas leves e pesadas para as componentes aleatórias.

Financiamento

Esta pesquisa não recebeu financiamento externo.

Conflito de interesses

Os autores declaram não haver conflito de interesses.

Referências

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716-723, 1974. DOI: https://doi.org/10.1109/TAC.1974.1100705.

ASHOUR, H. M.; ELKHATIB, W. F.; RAHMAN, M.; ELSHABRAWY, H. A. Insights into the recent 2019 novel coronavirus (SARS-CoV-2) in light of past human coronavirus outbreaks. **Pathogens**, v. 9, n. 3, 186, 2020. DOI: https://doi.org/10.3390/pathogens9030186.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time series analysis:** forecasting and control, 3. ed. San Francisco: Holden-Day, 1994.

CRODA, J. H. R.; GARCIA, L. P. Resposta imediata da vigilância em saúde à epidemia da COVID-19. **Epidemiologia e Serviços de Saúde**, v. 29, n. 1, p. 1-3, 2020. DOI: https://doi.org/10.5123/S1679-49742020000100021.

DEMERTZIS, K.; TSIOTAS, D.; MAGAFAS, L. Modeling and forecasting the COVID-19 temporal spread in Greece: An exploratory approach based on complex network defined splines. **International Journal of Environmental Research and Public Health**, v. 17, n. 13, 4693, 2020. https://doi.org/10.3390/ijerph17134693.

FRANÇA, E. B.; ISHITANI, L. H.; TEIXEIRA, R. A.; ABREU, D. M. X. D.; CORRÊA, P. R. L.; MARINHO, F.; VASCONCELOS, A. M. N. Deaths due to Covid-19 in Brazil: how many are there and which are being identified? **Revista Brasileira de Epidemiologia**, v. 23, e200053, 2020. https://doi.org/10.1590/1980-549720200053.

GARCIA, L. P.; DUARTE, E. Nonpharmaceutical interventions for tackling the Covid-1 epidemic in Brazil. **Epidemiologia e Serviços de Saúde**, v. 29, n. 2, e2020222, 2020. https://doi.org/10.5123/S1679-49742020000200009.

GOMES, S. C. P.; MONTEIRO, I. O.; ROCHA, C. R. Modelagem dinâmica da COVID-19 com aplicação a algumas cidades brasileiras. **Revista Thema**, v. 18, p. 1-25, 2020. https://doi.org/10.15536/thema.V18.Especial.2020.1-25.1793.

GRAJEDA, L. M.; IVANESCU, A.; SAITO, M.; CRAINICEANU, C.; JAGANATH, D.; GILMAN, R. H.; CRABTREE, J. E.; KELLEHER, D.; CABRERA, L.; CAMA, ,V.; CHECKLEY, W. Modelling subject-specific childhood growth using linear mixed-effect models with cubic regression splines. **Emerging Themes in Epidemiology**, v. 13, 1, 2016. https://doi.org/10.1186/s12982-015-0038-3.

HARRELL JUNIOR, F. E. **Regression modeling strategies:** with applications to linear models, logistic and ordinal regression, and survival analysis. New York: Springer-Verlag, 2015. DOI: https://doi.org/10.1007/978-3-319-19425-7.

HYNDMAN, R.; ATHANASOPOULOS G.; BERGMEIR C.; CACERES G.; CHHAY, L.; KUROPTEV, K.; O'HARA-WILD, M.; PETROPOULOS, F., RAZBASH, S.; WANG, E.; YASMEEN, F. **forecast.** R package version 8.12, http://pkg.robjhyndman.com/forecast. (2023). Acesso em: 02 mai. 2023.

LACHOS, V. H.; GHOSH, P.; ARELLANO-VALLE, R. B. Likelihood based inference for skew-normal independent linear mixed models. **Statistica Sinica**, v. 20, n. 1, p. 303-322, 2010. https://www.jstor.org/stable/24308993. Acesso em: 26 jun. 2023.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, v. 38, n. 4, p. 963-974, 1982. DOI: https://doi.org/10.2307/2529876.

NORDHAUSEN, K.; OJA, H.; PARSSINEN, O. Mixed-effects regression splines to model myopia data. **Journal of Biomeetrics & Biostatistics**, v. 6, n. 1, e100023, 2015. http://dx.doi.org/10.4172/2155-6180.1000239.

OLIVEIRA, L. C.; OLIVIA, J. T; RIBEIRO, M. H. D; TEIXEIRA, M.; CASANOVA, D. Forecasting the COVID-19 space-time dynamics in Brazil with convolutional graph neural networks and transport modals. **IEEE Access**, v. 10, p. 85064-85079, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3195535.

PAINE, C. E. T.; MARTHEWS, T. R.; VOGT, D. R.; PURVES, D.; REES, M.; HECTOR, A.; TURNBULL, L. A. How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists. **Methods in Ecology and Evolution**, v. 3, n. 2, p. 245-256, 2012. DOI: https://doi.org/10.1111/j.2041-210X.2011.00155.x.

PARRO, V. C.; LAFETÁ, M. L. M..; PAIT, F,; IPÓLITO, F. B.; TOPORCOV, T. N. Predicting COVID-19 in very large countries: **The case of Brazil. PLoS ONE**, v. 16, n. 7, 2021. DOI: https://doi.org/10.1371/journal.pone.0253146.

PEREIRA, I. G.; GUERIN, J. M.; SILVA JÚNIOR, A. G.; GARCIA, G. S.; PISCITELLI, P.; MIANI, A.; DISTANTE, C.; GONÇALVES, L. M. G. Forecasting Covid-19 dynamics in Brazil: a data driven approach. **International Journal of Environmental Research and Public Health**, v. 17, n. 14, 5115, 2020. DOI: https://doi.org/10.3390/ijerph17145115.

PEREIRA, M. A. A.; RUSSO, C. M. Nonlinear mixed-effects models with scale mixture of skew-normal distributions. **Journal of Applied Statistics**, v. 46 n. 9, p. 1602-1620, 2019. DOI: https://doi.org/10.1080/02664763.2018.1557122.

PINHEIRO, J. C.; BATES, D. M. **Mixed-effect models in S and S-PLUS**. New York: Springer, 2000. DOI: https://doi.org/10.1007/b98882.

PINHEIRO, J.; BATES, D.;DEBROY, S.; SARKAR, D. **nlme**: linear and nonlinear mixed-effects models. R package version 3.1-147, 2023. Disponível em: https://CRAN.R-project.org/package=nlme. (2023). Acesso em: 02 mai. 2023.

R CORE TEAM. **R: Language and environment for statistical computing**. R Foundation for Statistical Computing. Vienna, Austria. Disponível em: https://www.R-project.org. (2023). Acesso em: 02 mai. 2023.

RIBEIRO, M. H. D. M.; SILVA, R. G.; MARIANI, V. C.; COELHO, L. S. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. **Chaos, Solitons & Fractals**, v. 135, 109853, 2020. DOI: https://doi.org/10.1016/j.chaos.2020.109853.

RODRIGUEZ-MORALES, A. J.; GALLEGO, V.; ESCALERA-ANTEZANA, J. P.; MENDEZ, C. A.; ZAMBRANO, L. I., FRANCO-PAREDES, C.; SUÁREZ, J. A.; RODRIGUEZ-ENCISO, H. D.; BALBIN-RAMON, G. J.; SAVIO-LARRIERA, E.; RISQUEZ, A.; CIMERMAN, S. COVID-19 in Latin America: the implications of the first confirmed case in Brazil. **Travel Medicine and Infectious Disease**, v. 35, 101613, 2020. DOI: https://doi.org/10.1016/j.tmaid.2020.101613.

SALGOTRA, R.; GANDOMI, M.; GANDOMI, A. H. Time series analysis and forecast of the covid-19 pandemic in india using genetic programming. **Chaos, Solitons & Fractals**, v. 138, 109945, 2020. https://doi.org/10.1016/j.chaos.2020.109945.

SARAIVA, E. F.; SAUER, L. Modelagem e predição das quantidades de casos confirmados da Covid-19 no estado do Mato Grosso do Sul. **Revista Brasileira de Estatística**, v. 78, n. 245, p. 42-68, 2020. Disponível em: https://rbes.ibge.gov.br/images/doc/rbe_245jul_dez2020.pdf. Acesso em: 23 jun. 2023.

SCHUMACHER, F. L.; FERREIRA, C. S.; PRATES, M. O.; LACHOS, A.; LACHOS, V. H. A robust nonlinear mixed-effects model for COVID-19 death data. **Statistics and Its Interface**, v. 14, n. 1, p. 49-57, 2020. DOI: https://dx.doi.org/10.4310/20-SII637.

SCHUSTER, N. A.; RIINHART, J. J. M.; TWISK, J. W. R.; HEYMANS, M. W. Modeling non-linear relationships in epidemiological data: the application and interpretation of spline models. **Frontiers in Epidemiology**, v. 2, 975380, 2022. DOI: https://doi.org/10.3389/fepid.2022.975380.

SILVA, C. C.; LIMA, C. L.; SILVA, A. C. G.; SILVA, E. L.; MARQUES, G. S.; ARAÚJO, L. J. B.; ALBUQUERQUE JUNIOR., L. A.; SOUZA, S. B. J.; SANTANA, M. A.; GOMES, J. C.; BARBOSA, V. A. F.; MUSAH, A.; KOSTKOVA, P.; SANTOS, W. P.; SILVA FILHO, A. G. Covid-19 dynamic monitoring and real-Time spatio-temporal forecasting, **Frontiers in Public Health** v. 9, 2021. DOI: https://doi.org/10.3389/fpubh.2021.641253.

TRAPLETTI, A.; HORNIK, K.; LEBARON, B. **tseries**: time series analysis and computational finance. R package version 0.10-47, 2023. Disponível em: https://CRAN.R-project.org/package=tseries. (2023). Acesso em: 02 mai. 2023.

TSALLIS, C.; TIRNAKLI, U. Predicting COVID-19 peaks around the world. **Frontiers in Physics**, v. 8, 217, 2020. DOI: https://doi.org/10.3389/fphy.2020.00217.