Vol.66: e23220257, 2023
https://doi.org/10.1590/1678-4324-2023220257
ISSN 1678-4324 Online Edition

Article - Engineering, Technology and Techniques

Evaluating How the Social Restriction, the Government Response, the Health, and Economic Indices Affected the Prediction of the Number of Deaths Provoked by COVID-19 in Brazil Using Classical Statistical and Machine Learning Models

Marcello Montillo Provenza[1]*
https://orcid.org/0000-0003-0147-0795

Aderval Severino Luna[1]
https://orcid.org/0000-0001-6772-0182

Vinicius Layter Xavier[2]
https://orcid.org/0000-0002-7290-0652

[1]Universidade do Estado do Rio de Janeiro, Instituto de Química, Departamento de Química Analítica, Programa de Pós-Graduação em Engenharia Química, Rio de Janeiro, Brasil; [2]Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística, Departamento de Estatística, Programa de Pós-Graduação em Ciências Computacionais, Rio de Janeiro, Brasil.

Editor-in-Chief: Alexandre Rasi Aoki
Associate Editor: Fabio Alessandro Guerra

Received: 06-Apr-2022; Accepted: 22-Jun-2022.

*Correspondence: mprovenza@gmail.com; Tel.: +55-21-22340144 (M.M.P.).

HIGHLIGHTS

- Thirty statistical and machine learning models have been used to predict COVID-19 in Brazil.
- Each model has been trained and tested 266 times.
- The time series of accumulated deaths produces better estimates than daily deaths.
- The Cubist nonlinear regression model provides better predictions of accumulated deaths.

Abstract: The COVID-19 death predictions are helpful for the formulation of public policies, allowing the use of more effective social isolation strategies with less economic and social impact. This article evaluates a wide range of forecasting methods to identify the best models for predicting cumulative and daily deaths caused by COVID-19 in Brazil, considering a forecast for a seven-day horizon. With the seven-day horizon, the predictions have more accuracy. The dataset is from Oxford Covid-19 Government Response Tracker. The jackknife resampling technique was implemented, thus providing an accurate estimate for evaluating the predictive capacity of the models. Each model was fitted with 266 jackknife samples considering 30-day training bases. The comparison between predictions was made using the average results, considering R^2, MAPE, RMSE, and MAE. Models from different classes were adopted: 1 ETS, 4 ARIMA, 18 regression models, and 7 machine learning algorithms. The cumulative death models produce better results than daily deaths, as the cumulative death models are less influenced by time series components: cycle and seasonality. The best results for predicting daily deaths were attained by the Ridge regression method. The best results for predicting cumulative deaths were obtained by the Cubist regression method.

Keywords: COVID-19 Deaths; Jackknife; Statistical models; Regression models; Machine learning.

INTRODUCTION

Several cases of pneumonia patients have been associated with new human coronavirus disease (SARS-CoV-2, COVID-19) starting in December 2019 [1]. The virus demonstrated a large capacity for inter-human transmission spread rapidly worldwide, and became a pandemic with millions of deaths. Infected patients showed significantly varied symptoms, and their cases ranged from asymptomatic individuals to death. Studies and research on this new disease have become constant. Reports and articles with the prediction of cases and deaths emerged daily.

Virus outbreak forecasts for different countries are helpful for effectively allocating health resources and acting as an early warning system for government policymakers. The forecasts improve epidemiological surveillance and help stakeholders make timely decisions, allowing more specific social isolation strategies with less economic and social impact. Such responses can lead to correct political decisions that generate action protocols to contain future pandemics. Thus, the objective of this article is to predict deaths from COVID-19 in Brazil without the influence of vaccinating people and to guide public policy agents for future decision-making in other possible epidemics or pandemics.

Between 1927 and 1933, William Kermack and Anderson McKendrick created a model in which a fixed population (N) is considered with only three classes of individuals: Susceptible (S), Infected (I), and Recovered/Removed (R) [2-4]. Even though it was developed a long time ago, the SIR model is still widely used today in scientific articles, research, works, and studies in general about viruses and epidemics.

Each outbreak of viral infectious diseases exhibits specific patterns that need to be identified based on transmission dynamics. Machine learning algorithms have emerged as a new paradigm in recent scientific research due to their flexibility with data specifics. The versatility of this approach allowed its application in different contexts, from the forecast of financial variables to the analysis of sentiment in texts and medical applications [5].

Due to the highly complex nature of COVID-19, machine learning emerges as an effective tool to predict the outbreak been an alternative to the SIR model. An absolute novelty in forecasting outbreaks can be achieved by integrating machine learning and the SIR model [6]. Within a short period since the outbreak of COVID-19, advanced machine learning techniques have been used in the taxonomic classification of genomes, virus detection assay, and survival prediction in critically ill patients [7]. Machine learning methods are fundamental in screening, prediction, contact tracking, and drug development for epidemics [8].

Predictive models for the propagation of COVID-19 were developed using different characteristics such as climatic conditions (temperature and humidity), demographic data, health center data, et al. The experimental results show that climate variables are more relevant in predicting the mortality rate when compared to other census variables such as population, age, and urbanization. Thus, it could be concluded that temperature and humidity are essential characteristics for predicting the virus mortality rate. Furthermore, it is indicated that the higher the temperature value, the lower the number of infectious cases [9].

A COVID-19 outbreak prediction model was developed in Canada using state-of-the-art deep learning based on public datasets provided by the John Hopkins University and the Canadian Health Authority. Data patterns reveal that prompt and effective approaches taken by Canadian public health authorities to minimize human exposure have shown a positive impact compared to other countries. Based on the results, provinces that implemented social distancing guidelines before the pandemic had fewer confirmed cases [10].

Prediction models based on genetic algorithms have been developed for confirmed cases and deaths in India's three most affected states and across the country. The proposed models were developed from COVID-19 daily case reports published by the Government of India since the first lockdown in the country, which took place on March 24, 2020. The results found that the proposed genetic algorithm-based models are highly reliable for predicting COVID-19 time series in India. Models satisfy all external validation requirements and, therefore, can be used to predict future cases. Another relevant feature of genetic algorithm modeling is that it can work with less time series data and still provide reliable results [11].

The published works use a fixed training set and a fixed test set in a specific period in time series [11-13]. However, for a better assessment of the predictive capacity of the models, it is necessary to use many training and test sets. Thus, this work has as its differential the use of multiple sets of training and testing.

The jackknife method was programmed, generating a total of 266 training and test sets for each model, thus providing a more accurate estimate of the predictive capacity of the models.

Furthermore, the other works published for the time series prediction of COVID-19 use a minimal scope of predictive methods [6,9-13]. This work has a differential in using a broad scope of forecasting methods, a total of 30 methods, making it possible to identify the methods that produce better predictions of daily and accumulated deaths. This work also differs from the others by using, in addition to the time series, the exogenous variables: cases; restriction index; government response index; health index; and economic index.

MATERIAL AND METHODS

In addition to non-pharmacological measures, containment measures were very used to contain the spread of the virus. Most countries recommended their population stay at home. Other measures used were: to do extra investments in the health and the economy, closed schools, airports, workplaces, etc. The database used was Oxford COVID-19 Government Response Tracker (OxCGRT) [14]. This database systematically collects daily information on various public policies that governments have adopted to respond to the pandemic, such as lockdown, travel restrictions, etc. This work was done with nine variables (one dependent and eight independent):

- Dependent variable (Y): deaths.
- Independent variables: cases (X1); restriction index (X2); government response index (X3); health index (X4); economic index (X5); lag 7, Y values at prior time steps (X6 = Y(t-7)); lag 8, Y values at prior time steps (X7 = Y(t-8)); and lag 9, Y values at prior time steps (X8 = Y(t-9)).

The restriction index is the restriction adopted by the region, such as closing airports and other places. The government index is the response of the government to the pandemic of COVID-19. The health index is the investment done by the government in health policies. The economic index is the financial support that the government made during the pandemic, such as financial aid to people who have lost their jobs or cannot work. All variables are completely described in a global panel of the pandemic policies [14].

These variables were chosen because they are measures used by Oxford to analyze countries worldwide concerning the level of restriction against COVID-19. The lags of variables were selected according to the seven-day horizon, which offers greater precision. Furthermore, in an application of a time series study, it is interesting to initially use the period to 7 days forecast because of the small number of observations.

The period analyzed was from March 17, 2020, to January 20, 2021. It is because the first death by COVID-19 in Brazil was notified on March 17, 2020, and the vaccination started on January 17, 2021. All data analysis, modeling, and programming were conducted using the R [15] program, with the packages: stats [15], ggcorrplot [16], forecast [17], and caret [18].

Time series

The primary purpose of time series analysis is forecasting. This methodology allows forecasting future values through the present and past values [19]. Therefore, models are essential to provide the necessary support for statistical inference [20]. A time series can have four components: trend, cycle, seasonality, and residual. The trend describes the behavior of the variable over time. The cycle is a periodic fluctuation concerning the trend. Seasonality is a change that occurs in specific periods of a time series. The residual is represented by random fluctuations resulting from unexpected facts [19].

Exploratory data analysis

The first analysis of time series is visual. Some inferences and hypotheses can be suggested. The boxplot is a standardized way of displaying data distribution based on the five values: minimum, first quartile, median, third quartile, and maximum. Outliers can be displayed as individual points. This technique makes no assumptions about the statistical distribution involved in the data. The spaces between the different parts of the box indicate the degree of dispersion, the asymmetry in the data, and the outliers [21]. Scatter plots are used to observe relationships between two variables. Pearson's correlation measures the degree of linear association between variables [22].

Many statistical methods make assumptions that the data are from a population with a specific probability distribution. The characteristic of this distribution can be one of the purposes of the analysis. There are statistical tests responsible for determining the theoretical distribution of data. The following tests were used in this work: Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises, Lilliefors, Pearson, and Shapiro-Francia to test whether the data follow the Gaussian distribution. In these tests, the null hypothesis (H_0) is that data follow a normal distribution. The alternative hypothesis (H_1) is that data do not follow a normal distribution [23].

ETS and Box-Jenkins models

The ETS model is a "special" class of exponential smoothing. The point forecasts produced by the models are identical if they use the same smoothing parameter values [24]. The characterization of the model following the terminology of [25] and [26] is done using a three-character string. The first letter indicates the type of error (A or M); the second letter indicates the type of trend (N, A, or M); and the third letter indicates the type of seasonality (N, A, or M). In all cases, N = none, A = additive and M = multiplicative.

The Box-Jenkins methodology consists of adjusting Autoregressive Integrated Moving Averages (ARIMA) models. The strategy for building the model is based on an iterative cycle. The stages of the interactive cycle are specification, identification, estimation, and diagnosis. In general, the postulated models are parsimonious, as they contain a small number of parameters, and the predictions obtained are pretty accurate [15]. The ARIMA models assume that the values of a time series have a dependency relationship where each value can be explained by the previous value of the series data [20]. The purpose of the Box-Jenkins methodology is to determine the three components that make up the structure: p (autoregressive parameters), d (differentiation processes), and q (moving average parameters), thus forming the ARIMA (p,d,q) [19].

Autocorrelation is the correlation of the variable X(t) with itself at the last instant X(t-k), which is called the time lag k. The Autocorrelation Function (ACF) measures the dynamics of the correlation between a variable and its lags. The Partial Autocorrelation Function (PACF) is a measure of the correlation between observations of a time series that are separated by k time units (X(t) and X(t-k)) [19,20].

Regression models

Seek to identify the relationship between the dependent and independent variables. This relationship can be linear or nonlinear. In regression models with cross-sectional data, the order of observations is irrelevant for the analysis. In time series, the order of the data is fundamental. A significant feature of this type of data is that neighboring observations are generally dependent over time, so it is interesting to analyze and model this dependence [20].

Typically, the data sets have many independent variables, so it is necessary to know which are relevant to explain the dependent variable. In these cases, mechanisms are needed to choose the best subset of independent variables to explain the dependent variable. For this, Regularization Methods are recommended. These methods incorporate a constraint into the model, limiting the model's coefficients and therefore selecting the most important independent variables [27].

Dynamic regression models are also called ARIMA models with exogenous variables (ARIMAX). In linear regression models, it is assumed that noise has zero mean, constant variance, normal distribution, and independence, thus having no serial correlation [20]. ARIMAX combines the dynamics of time series and the effect of explanatory variables. The dependent variable is explained by its lagged values and current and past values of exogenous variables.

In addition to ARIMAX, the following linear regression models were used in this work: Multiple (MLR), Stepwise, Stepwise with lower Akaike value (Stepwise AIC), Lasso, Ridge, Elastic Net, Boosted, Boosted Tree, and Robust. The nonlinear models were: Cubist, Multivariate Adaptive Regression Splines (MARS), and MARS with cross-validation pruning (MARS gCV).

Machine learning

Explores the study and construction of algorithms that can learn from data and make predictions [28]. Studies with Support Vector Machines (SVM) started to be developed in the 60s, in Russia, by Vapnik, Lemer, and Chervonenkis. However, it can be said that the SVM had its starting point along with the development of the theory of statistical learning by Vapnik in 1979. The current form was developed by Vapnik in the late 1990s and aimed to find a hyperplane that maximizes the margin between classes. Support Vector Regression (SVR) maintains the same characteristics as SVM [29].

Random Forests (RF) are formed by several decision trees. All trees are used, each of which provides an estimate. The final classification is given by the most frequent result in all trees [27,28,30].

Artificial Neural Networks (ANN) are computational techniques that present a mathematical model inspired by the human brain [28]. The most crucial property of ANNs is their ability to learn from their environment and improve their performance. It is done through an iterative process of adjusting the weights of the network [30]. The Autoregressive Neural Networks technique combines the autoregressive statistical model and neural networks, resulting in an AR-NN(p) model. In the AR-NNX model, exogenous variables are included, providing new data to improve prediction performance. In addition to the lagged values of the dependent variable, independent variables can be added that will also be used [31].

Boosting belongs to the machine learning category called an ensemble. Ensemble techniques involve groups of predictive models to achieve better model accuracy and stability. Boosting refers to a family of algorithms that convert weak learning into strong learning. The prediction of each learning is combined to convert it into strong learning [27,30]. The eXtreme Gradient Boosting (XGBoost) algorithm combines the Boosting and tree models.

Assessment metrics

The Jackknife resampling method was implemented, in which the entire database was used. The strategy removes a sample from the total observed set, recalculating the estimator from the remaining values. The use of this technique promotes the reduction of uncertainties, thus having an accurate estimate for evaluating the predictive capacity of the models. Cross-validation uses 30 values (days) for training, and the forecast is given considering a 7-day horizon, where forecast metrics are applied. Each model has been trained and tested 266 times.

Forecast metrics are averaged to evaluate the models. In this work, the following figures of merit have used the Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE).

RESULTS

Between March 17, 2020, and January 20, 2021, the average number of deaths from COVID-19 was 687 ± 391. The mean of cases was 27,865 ± 18,461. The first death occurred when there were 51 reported cases (Table 1).

Table 1. Statistics of deaths and daily cases.

Statistics	Deaths	Cases
Average	687	27,865
Standard Deviation	391	18,461
Minimum Value	1	51
Q1	376	13,381
Median	676	26,017
Q3	1,018	42,144
Maximum Value	1,595	87,843

Both time series show a cycle, with trends of growth and reduction. The boxplot reveals no outliers for deaths and one outlier for cases (Figure 1). All p-values were below 0.05, indicating that the data are not normally distributed considering the seven tests performed (Table 2).

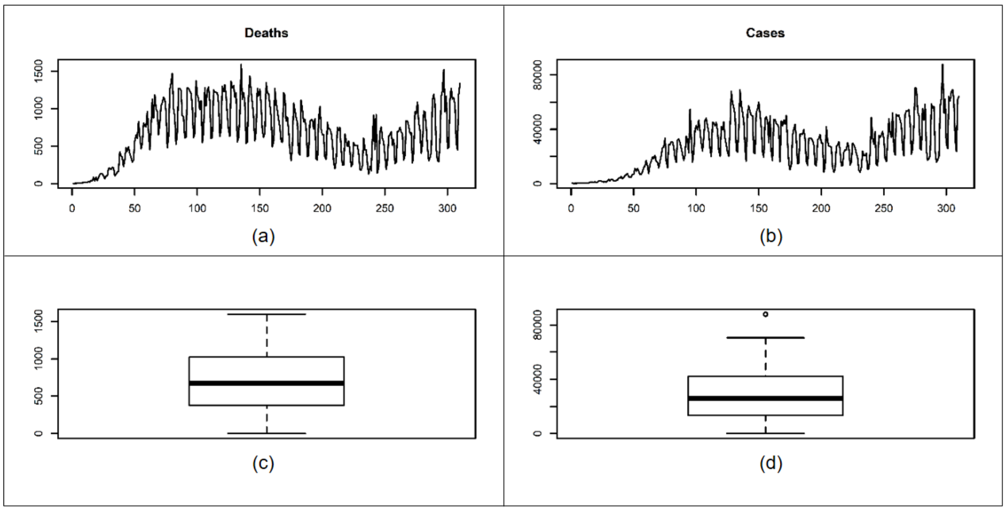

Figure 1. (a) Time series of deaths COVID-19; (b) Time series of cases COVID-19; (c) Box-plot of deaths COVID-19; (d) Box-plot of cases COVID-19.

Table 2. p-value of tests.

Tests	Deaths	Cases
Shapiro-Wilk	6.02e^{-6}	2.06e^{-6}
Kolmogorov-Smirnov	2.2e^{-16}	2.2e^{-16}
Anderson-Darling	2.30e^{-5}	2.35e^{-5}
Cramer-von Mises	0.0009	0.0007
Lilliefors	0.0022	0.0024
Pearson	0.0014	1.809e^{-7}
Shapiro-Francia	4.54e^{-5}	1.28e^{-5}

The scatter plot reveals that deaths and cases have a positive correlation - the greater the number of cases, the greater the number of deaths (Figure 2). Deaths and cases showed a coefficient of correlation equal to 0.78.

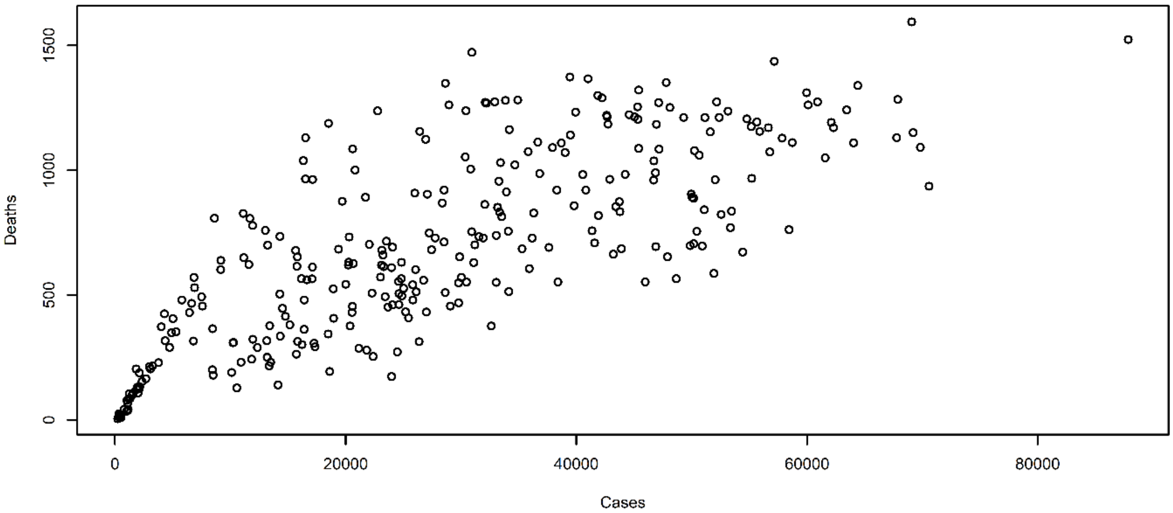

Figure 2. Scatter plot between cases/deaths.

Concerning the correlogram, which shows the correlation between all the variables considered in this study. Only the health index and the government index have a coefficient of correlation greater than 0.9 (Figure 3).

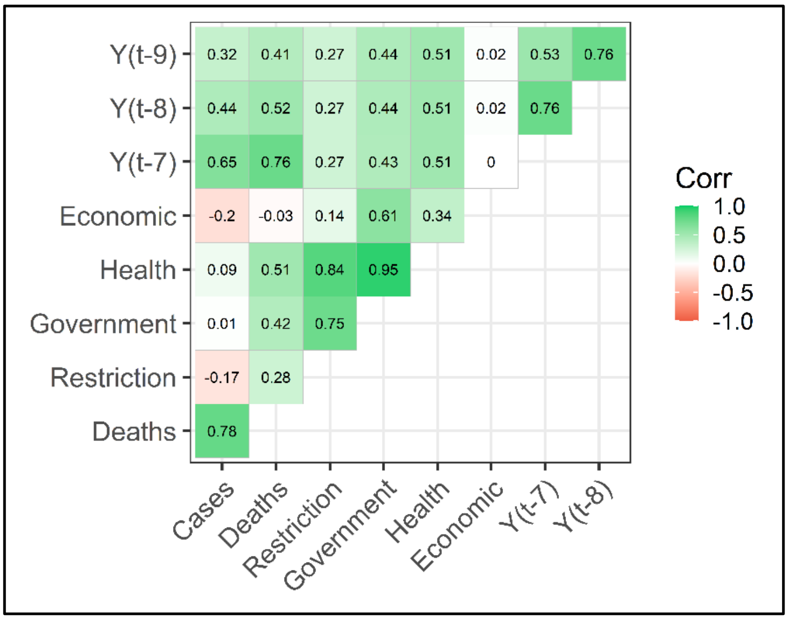

Figure 3. Correlogram.

ACF and PACF indicate a need to carry out at least one differentiation for the ARIMA model for the time series of daily and accumulated deaths (Figure 4).

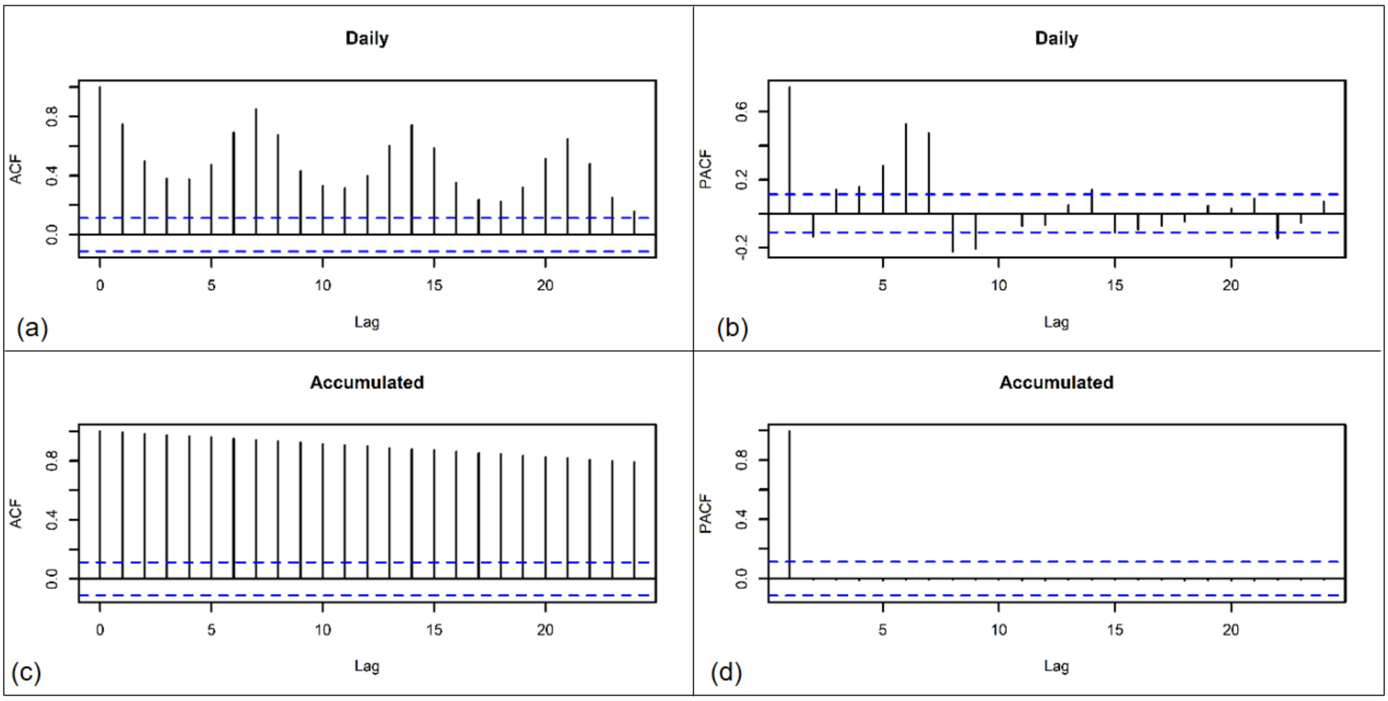

Figure 4. (a) ACF of daily deaths COVID-19; (b) PACF of daily deaths COVID-19; (c) ACF of accumulated deaths COVID-19; (d) PACF of accumulated deaths COVID-19.

After 266 training and test sets for each model, the averages of the prediction metrics of the test bases were calculated. Tables 3 and 4 show the evaluations for the regression models, Random Forest, Support Vector Regression, Artificial Neural Networks, and eXtreme Gradient Boosting. R^2 shows that accumulated deaths produced better estimates because their values are more significant than daily.

Table 3. Results of regression models, RF, SVR, Neural Networks, and XGBoost, to predict daily deaths.

Models - Daily Deaths	**R^2**	**RMSE**	**MAE**
Multiple Linear Regression	0.708	1,033.572	827.929
Stepwise	0.750	153.575	127.229
Stepwise AIC	0.738	330.231	263.424
Lasso	0.790	149.178	126.194
Ridge	0.772	136.081	112.700
Elastic Net	0.752	149.289	123.593
Boosted	0.715	185.547	157.310
Boosted Tree	0.689	189.967	160.042
Robust	0.796	145.080	119.805
Cubist	0.745	151.647	122.799
Multivariate Adaptive Regression Splines	0.731	151.048	121.064
Multivariate Adaptive Regression Splines gCV	0.724	153.773	123.253
Random Forest	0.669	157.836	126.500
Support Vector Regression	0.287	305.573	260.210
Artificial Neural Network	0.690	803.796	764.776
Artificial Neural Network Average	0.580	803.796	764.776
eXtreme Gradient Boosting	0.640	170.257	136.794

Table 4. Results of regression models, RF, SVR, Neural Networks, and XGBoost, to predict accumulated deaths.

Models - Accumulated Deaths	**R^2**	**RMSE**	**MAE**
Multiple Linear Regression	0.895	6,464.189	5,086.783
Stepwise	0.986	1,230.472	1,100.752
Stepwise AIC	0.970	1,918.301	1,615.088
Lasso	0.989	717.445	652.330
Ridge	0.987	1,104.163	1,036.507
Elastic Net	0.983	898.344	833.856
Boosted	0.873	9,810.700	9,741.453
Boosted Tree	0.550	6,172.956	5,967.626
Robust	0.986	789.590	717.162
Cubist	0.993	467.774	408.664
Multivariate Adaptive Regression Splines	0.995	523.655	468.949
Multivariate Adaptive Regression Splines gCV	0.995	520.716	466.455
Random Forest	0.575	4,621.139	4,327.940
Support Vector Regression	0.735	13,823.770	13,675.770
Artificial Neural Network	0.424	114,545.600	114,522.100
Artificial Neural Network Average	0.541	114,545.600	114,522.100
eXtreme Gradient Boosting	0.614	3,462.960	3,082.320

Robust (considering R^2) and Ridge (considering RMSE and MAE) regressions had the best fit to predict COVID-19 daily deaths (Table 3). Robust regression has a set of procedures that resist minor violations of a model's requirements, and Ridge regression has a penalty that decreases the complexity of a model.

MARS (considering R^2), and Cubist (considering RMSE and MAE) regressions had the best fit to predict COVID-19 accumulated deaths (Table 4). These methods use a rules-based approach, like decision trees.

Tables 5 and 6 present the evaluations for the ARIMA, ETS, ARIMAX, AR-NN, and AR-NNX models. The ARIMA and ARIMAX models were fitted with different p and q components and were up to 5 each (i.e., at each one of fitted 266 training bases, the p and q values can be changed). The differences (d) were up to 4 for ARIMA and up to 5 for ARIMAX. The MAPE shows that the accumulated deaths produced better estimates because their values are more significant than daily.

Table 5. Results of the ARIMA, ETS, ARIMAX, AR-NN, and AR-NNX models for predicting daily deaths.

Models - Daily Deaths	**MAPE (%)**	**RMSE**	**MAE**
ARIMA(p,1,q)	30.940	218.335	181.531
ARIMA(p,2,q)	97.416	685.901	621.078
ARIMA(p,3,q)	322.537	2,442.444	2,060.772
ARIMA(p,4,q)	493.150	4,033.814	3,168.591
ETS	42.142	293.087	247.299
ARIMAX(p,0,q)	6,266.804	64,337.909	27,326.906
ARIMAX(p,1,q)	234.694	1,976.351	854.554
ARIMAX(p,2,q)	70.643	459.269	267.681
ARIMAX(p,3,q)	687.674	3,709.207	2,219.037
ARIMAX(p,4,q)	3,149.418	18,092.704	10,388.141
ARIMAX(p,5,q)	104,905.200	480,294.400	240,536.700
AR-NN(p)	37.656	285.724	229.528
AR-NNX	37.627	286.041	229.761

Table 6. Results of the ARIMA, ETS, ARIMAX, AR-NN, and AR-NNX models for predicting accumulated deaths.

Models - Accumulated Deaths	MAPE (%)	RMSE	MAE
ARIMA(p,1,q)	1.105	619.431	541.510
ARIMA(p,2,q)	0.961	551.289	486.093
ARIMA(p,3,q)	3.172	2,732.814	2,272.388
ARIMA(p,4,q)	5.858	6,245.084	4,808.998
ETS	0.881	629.884	557.747
ARIMAX(p,0,q)	23.201	1,787.892	1,440.029
ARIMAX(p,1,q)	22.648	1,614.937	1,247.126
ARIMAX(p,2,q)	21.834	1,559.284	1,202.490
ARIMAX(p,3,q)	22.639	1,636.814	1,254.545
ARIMAX(p,4,q)	14.621	1,041.308	805.586
ARIMAX(p,5,q)	32.233	2,479.165	1,807.406
AR-NN(p)	2.826	1,870.246	1,587.506
AR-NNX	2.820	1,872.006	1,589.541

ARIMA(p,1,q) (considering MAPE, RMSE, and MAE) had the best fit to predict COVID-19 daily deaths (Table 5). ARIMA modeling contains a small number of parameters, and predictions are pretty accurate. In this case, one differentiation was necessary to obtain the best model.

ETS (considering MAPE) and ARIMA(p,2,q) (considering RMSE and MAE) had the best fit to predict COVID-19 daily deaths (Table 6). ETS modeling gives greater weight to past observations over recent ones. This model also considers elements such as trend and seasonality. In ARIMA(p,2,q), two differences were necessary to obtain the best model.

DISCUSSION

The COVID-19 pandemic has killed millions of people since the end of 2019. With the evolution of data systems and the high contagion, information began to be published daily on the number of cases and deaths caused by the virus [14]. Thus, it became essential to use forecasting techniques and models to project this count in regions and countries with high indexes [9-12]. In Brazil, there have already been more than 650,000 deaths (it is about 0.3% of the population). Vaccination in Brazil began on January 17, 2021, but few individuals were vaccinated during this work. It strengthens the hypothesis that there was no external influence on the predictions and certifies the results obtained. The Jackknife methodology was applied to reduce possible temporal variations, with many training and test sets.

Several prediction models were observed for daily and accumulated deaths. Time series do not have a normal distribution. The scatter plot and the correlogram show that cases and deaths have a positive correlation (0.78). Health and restriction (0.84) indices and government and restriction (0.75) indices also have a positive correlation. However, when these variables are tested with others, only health and government indices have a strong correlation (0.95).

The predicted numbers at the end of the epidemic are highly dependent on the length of the time series used in the predictive models [32]. Therefore, was used a 7-day prediction. The exogenous variables cases, restriction index, government response index, health index, and the economic index were used to get more precision. To get more accurate, the averages of the 266 prediction metrics of the test bases revealed the best fit of the models. The forecast for accumulated deaths produced better estimates than the daily ones (this result corroborates with [33]), as seen by R^2 and MAPE. Possibly, this is because temporal elements such as cycle, seasonality, and randomness have less influence on accumulated deaths than on daily deaths.

Considering R^2 for the prediction of daily deaths, the best models were the Robust (0.796), Lasso (0.790), Ridge (0.772), and Elastic Net (0.752) regressions. For accumulated deaths, the best models were the MARS gCV (0.995), MARS (0.995), Cubist (0.993), and Lasso (0.989) regressions. Looking at the MAPE, the models that showed better results for daily deaths were ARIMA(p,1,q) (30.940), AR-NNX (37.627), AR-NN (37.656), and ETS (42.142). For accumulated deaths were ETS (0.881), ARIMA(p,2,q) (0.961), ARIMA(p,1,q) (1.105) and AR-NNX (2.820). Analyzing the RMSE, the best models for daily deaths were the Ridge (136.081), Robust (145.080), Lasso (149.178), and Elastic Net (149.289). For the accumulated deaths were the Cubist (467.774), MARS gCV (520.716), MARS (523.655) regressions, and ARIMA(p,2,q) (521.289). Looking at the

MAE, the best models for the daily deaths were the Ridge (112.700), Robust (119.805), MARS (121.064), and Cubist (122.799) regressions. For the accumulated deaths were the Cubist (408.664), MARS gCV (466.455), MARS (468.949) regressions, and ARIMA(p,2,q) (486.093).

Considering deaths by COVID-19 in Brazil, as the time series does not have a normal distribution, there is no outlier, and the variables do not have a high correlation, nonlinear regressions had the best fit for predicting accumulated deaths.

The proposed models were used for the COVID-19 pandemic. However, they can be used for other epidemic or pandemic situations with possible good results, because in epidemiology the historical context is extremely important [34]. The change in daily numbers of COVID-19 is affected by many factors, such as the population's adherence to prevention measures, vaccination, social isolation, and new variants of the virus. Analysis suggests that COVID-19 shows chaotic behavior, like in previous epidemics [35]. Government campaigns are very important to avoid the possible underreporting of cases and deaths. Delays in notifications can also bring biased results.

Due to these reasons, to have better accuracy of the predictions, it is necessary to use many training and test bases. It is important to emphasize that this study was designed without the influence of vaccination of people. Therefore, these models may be interesting for use at the beginning of an epidemic or a pandemic. In practice, this work proposes that at the beginning of an epidemic, the forecast is made by the non-linear model. In addition, predictions can be made daily, as long as the data used is accumulated.

The results presented in this study differ from some previously published works [36-39]. The Jackknife resampling method used here has better accuracy, because used 266 training and testing bases. So as data on the pandemic are published daily, the forecasts must also be updated periodically. For future works, it is recommended to include some other variables like the vaccination of people and the virus reproduction rate. In addition, to have the best fit of the models, it may be interesting to consider smaller regions such as some districts.

CONCLUSION

This work used multiple sets of training and testing to predict the number of deaths from COVID-19 in Brazil. Furthermore, exogenous variables were used. This procedure helps to produce a more accurate estimate of the predictive capacity of the models. A total of 30 forecasting methods were used, making it possible to identify the methods that produce better predictions of daily and accumulated deaths. Therefore, this work showed that the time series of accumulated deaths produced better estimates than daily deaths. The cubist regression had the best fit for cumulative deaths, and ridge regression had the best fit for daily deaths. The contribution of this work revealed that nonlinear regressions are the best methodology to predict the number of accumulated deaths from COVID-19 in Brazil.

Funding: This research received no external funding.

Acknowledgments: The third author thanks the Research Support Foundation of the State of Rio de Janeiro (FAPERJ) for their financial support (E_26/2021 – Basic Research Assistance, APQ1).

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020.
2. Kermack WO, Mckendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of London. Series A, Containing Papers of a Math and Phys Character. 1927;115(772):700-21.
3. Kermack WO, Mckendrick AG. Contributions to the mathematical theory of epidemics. II.-The problem of endemicity. Proceedings of the Royal Society of London. Series A, Containing Papers of a Math and Phys Character, 1932;138(834):55-83.
4. Kermack WO, Mckendrick AG. Contributions to the mathematical theory of epidemics. III.-Further studies of the problem of endemicity. Proceedings of the Royal Society of London. Series A, Containing Papers of a Math and Phys Character. 1933;141(843):94-122.
5. Peng Y, Nagata MH. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. Chaos, Solitons & Fractals. 2020;139.

6. Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Atkinson PM. Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188. 2020.
7. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. 2020.
8. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals. 2020;138.
9. Malki Z, Atlam ES, Hassanien AE, Dagnew G, Elhosseini MA, Gad I. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons & Fractals. 2020;138.
10. Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals. 2020;138.
11. Salgotra R, Gandomi M, Gandomi AH. Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. Chaos, Solitons & Fractals. 2020;138.
12. Melin P, Monica JC, Sanchez D, Castillo O. Multiple ensembles neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico. In: Healthcare. Multidisciplinary Digital Publishing Institute. 2020;181.
13. Khan FM, Gupta R. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. J Saf Sci Res. 2020;1(1):12-8.
14. Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, Webster S, Cameron-Blake E, Hallas L, Majumdar S, Tatlow H. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour. 2021;5(4):529-38.
15. R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019.
16. Kassambara A. ggcorrplot: Visualization of a Correlation. Matrix using 'ggplot2'. R package version 0.1.3. 2019.
17. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, et al. _forecast: Forecasting functions for time series and linear models_. R package version 8.12. 2020.
18. Kuhn M. caret: Classification and Regression Training. R package version 6.0-86. 2020.
19. Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. John Wiley & Sons, 2015.
20. Gujarati DN, Porter DC. Econometria básica. 5th ed. Amgh, 2011.
21. Ross SM. Introduction to probability and statistics for engineers and scientists. Academic Press, 2020.
22. Kloke J, Mckean JW. Nonparametric statistical methods using R. Boca Raton: CRC Press, 2015.
23. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J of End and Met. 2012;10(2).
24. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts, 2018.
25. Hyndman RJ, Athanasopoulos G. A state space framework for automatic forecasting using exponential smoothing methods. Int. J. Forecast. 2002;18(3):439-454.
26. Hyndman RJ, Athanasopoulos G. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media, 2008.
27. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer, 2013.
28. Mueller JP, Massaron L. Aprendizado de Máquina para Leigos. Alta Books, 2019.
29. Naguib IA, Darwish HW. Support vector regression and artificial neural network models for the stability-indicating analysis of mebeverine hydrochloride and sulpiride mixtures in pharmaceutical preparation: A comparative study. Spectrochim. Acta A Mol. Biomol. Spectrosc. 2012;86:515-26.
30. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.
31. Haddoun A, Benbouzid MEH, Diallo D, Abdessemed R, Ghouili J, Srairi K. Modeling, analysis, and neural network control of an EV electrical differential. IEEE Trans on Ind Electr, v. 55, n. 6, p. 2286-2294, 2008.
32. Martinez EZ, Aragon DC, Nunes AA. Long-term forecasts of the COVID-19 epidemic: a dangerous idea.Rev. Soc. Bras. Med. Trop. 2020;53.
33. Braga MDB, Fernandes RDS, Souza Jr GND, Rocha JECD, Dolácio CJF, Tavares Jr IDS, et al. Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon.PloS one. 2021;16(3).
34. Beaglehole R, Bonita R, Kjellström T. Basic epidemiology. World Health Organization, 2006.
35. Jones A, Strigul N. Is spread of COVID-19 a chaotic epidemic? Chaos, Solitons & Fractals. 2021;142.
36. Divino F, Ciccozzi M, Farcomeni A, Jona-Lasinio G, Lovison G, Maruotti A, et al. Unreliable predictions about COVID-19 infections and hospitalizations make people worry: The case of Italy. J. Med. Vir. 2022;94(1):26-28.
37. Maleki M, Mahmoudi MR, Wraith D, Pho KH. Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel Med and Infec Disease. 2020;37.
38. Masum M, Masud MA, Adnan MI, Shahriar H, Kim S. Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for COVID-19 forecasting and management. Socio-Econ Plann Sciences. 2022;80.

39. Mohan S, Solanki AK, Taluja HK, Singh A. Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. Comp Bio Med. 2022;144.