An Acad Bras Cienc (2022) 94(Suppl. 3): e20210921 DOI 10.1590/0001-3765202220210921
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

MATHEMATICAL SCIENCES

Predicting COVID-19 cases in various scenarios using RNN-LSTM models aided by adaptive linear regression to identify data anomalies

LUIS RICARDO ARANTES FILHO, MARCOS L. RODRIGUES, REINALDO R. ROSA & LAMARTINE N.F. GUIMARÃES

Abstract: The evolution of the Sars-CoV-2 (COVID-19) virus pandemic has revealed that the problems of social inequality, poverty, public and private health systems guided by controversial public policies are much more complex than was conceived before the pandemic. Therefore, understanding how COVID-19 evolves in society and looking at the infection spread is a critical task to support efficient epidemiological actions capable of suppressing the rates of infections and deaths. In this article, we analyze daily COVID-19 infection data with two objectives: (i) to test the predictive power of a Recurrent Neural Network - Long Short Term Memory (RNN-LSTM) on the daily stochastic fluctuation in different scenarios, and (ii) analyze, through adaptive linear regression, possible anomalies in the reported data to provide a more realistic and reliable scenario to support epidemic control actions. Our results show that the approach is even more suitable for countries, states or cities where the rate of testing, diagnosis and prevention were low during the virus dissemination. In this sense, we focused on investigating countries and regions where the disease evolved in a severe and poorly controlled way, as in Brazil, highlighting the favelas in Rio de Janeiro as a regional scenario.

Key words: COVID-19, machine learning, RNN-LSTM, Brazil, favelas, Rio de Janeiro.

INTRODUCTION

The pandemic caused by the SARS-CoV-2 virus (COVID-19), has severely affected humanity in various segments, such as the public health, economy, and political sectors. Moreover, the disease raises severe infections and has a high spreading capability, which makes it lethal in many cases. By August 2021, about 200 million cases have been counted, with approximately 4.5 million deaths, and as indicated by various studies the normality recovery, even with the available vaccines (Sinovac, Pfizer, Astrazeneca, Moderna, Janssen), will be slow and difficult (Tian et al. 2020, Lauer et al. 2020, Anjos 2020, The Lancet 2020, Sharma et al. 2021).

Thus, actions to monitor and control the virus spreading play a fundamental role in reducing the number of infected people and, consequently, the number of related deaths. In this sense, this work brings two complementary machine learning approaches to understanding the stochastic nature of COVID-19. Firstly, we test a prediction strategy for COVID-19 based on Long-Short-Term Memory (LSTM), a Recurrent Neural Network (RNN) model. LSTM (Hochreiter & Schmidhuber 1997a,b) attempt to establish relationships in data that have long-term temporal dependencies through memory cells (memory neurons, context nodes) that retain information for long periods during training epochs. Such a deep learning model is able to learn intrinsic information in time series data, allowing learning and inferring predictions with good accuracy and precision[1]. The concept of Deep Learning (DL) corresponds to techniques directed at artificial neural networks to explore and amplify the power of nonlinear data analysis using a large number of intermediate layers (Deng et al. 2014, Goodfellow et al. 2016, Vasilev et al. 2019). The LSTM-RNN algorithms were implemented in Python using Keras[2].

We consider COVID-19 infection reported daily from January to August 2020 (Voz das Comunidades 2020, Worldometer 2020), analyzing the daily transmission fluctuation in a period where the main suppression factor would be social isolation (about six months before vaccination starts). In this study, we focus on observing the evolution of confirmed cases in some countries, especially in Brazil, including data related to favelas in Rio de Janeiro. Such scenarios allow us to infer how robust the LSTM models are to improve the understanding, monitoring and prediction of the infection transmission rate. Therefore, we have applied LSTM models based on specific cause and effect criteria, strictly related to each COVID-19 time series behavior, verifying the effectiveness of these criteria on its fluctuation and evolution in different scenarios. We provide a background of events in the countries that had their first contact with the virus and how the authorities and society have dealt with the pandemic. We looked at certain criteria linked to the disease's behavior and its impacts, such as how symptoms occur, the average time to diagnosis, time to treatment, and the effects of epidemiological control interventions made by the authorities, such as lockdown actions (Tian et al. 2020, Lauer et al. 2020, Zhou et al. 2020, Chen & Yu 2020).

We have analysed data from 12 different countries on a grid of parameters resulting in about 288 predictive time series models. The application of these models in specific regional scenario, as Rio's favelas, helps to identify how the disease evolves in communities that lack basic health structure and conditions that make both social distance and lockdown actions extremely difficult (Voz das Comunidades 2020, Worldometer 2020). In most of these virus transmission scenarios an important aspect to be discussed is the quality of data collection, subject to outliers such as underreporting. Therefore, as a complementary approach we use Adaptive Linear Regression to detect data anomalies (*outliers*) in the daily cases of COVID-19 infections. This approach allows the data to be treated with more confidence by getting closer to the actual pandemic behavior after removing the outliers values. In this way, the models can come closer to identifying the actual scenario caused by human influence on pandemic control (Taylor & Letham 2017). At this stage, the anomaly detection approach was focused on two specific countries, Brazil and the USA, which, in addition to presenting the anomalies in accounting for daily infection cases, have the worst scenarios related to a large number of infected people and deaths.

The rest of the paper is organized as follows: We describe, in section Related Works, some published results related to the approaches of this paper. In section Materials and Methods, we explain the methodology that indicates the main steps in selecting and collecting the data to feed the LSTM models, explaining how these models can be parameterized and trained, and how the predictions are validated by handling the data to remove anomalies in the time series. In the section Results,

[1] This work include an Appendix with a brief description of LSTM-RNN into the context of DL.

[2] The Keras tool is a framework designed in Python programming language, which supports the development of DL applications (Chollet 2021, Vasilev et al. 2019).

we present the main results related to both approaches. The section Conclusions, touches on the critical interpretation of these approaches, as well as a discussion of our work in progress capable of improving results by updating the data.

RELATED WORKS

To contextualize the research presented in this article, we describe below some published works whose approaches are in the same analytical context as those discussed in the previous section.

The work presented by Kırbaş et al. (2020) indicates a comparison of several approaches for predicting COVID-19 cases. The researched models correspond to ARIMA, NARNN, and LSTM models. The analyzed data were the total number of confirmed cases in countries such as Denmark, Belgium, Germany, France, the UK, Finland, Switzerland, and Turkey. According to Kırbaş et al. (2020), the model with the best prediction performance was the LSTM model, hence the authors present the prediction of the total number of cases in 14 days for the selected countries. In the same line, the authors Chimmula & Zhang (2020) present a DL model to predict the outbreak of COVID-19 in Canada. These authors evaluated LSTM models to predict trends in the data of COVID-19 cases and, so, they indicate by these models the moment when the cases start to decrease.

Recently, there have been many studies on the prediction of COVID-19 growth using DL techniques. However, few of them consider the end of the epidemic. At the forefront of this goal Yan et al. (2020), shows the first LSTM model to predict daily cases considering the end of the epidemic and compares with two other approaches based on Logistics and Hill mathematics models. The main result achieved for this study was the best forecast model concerning real data on the number of cases in Tianjin, China, Hong Kong, South Korea, and Italy. The novelty method proposed, uses Fully connected layers to emulate the disease incubation period combined with an LSTM model to find factors that influence potential cases. The authors pointed out that their approach can better handle multiple characteristics, such as the number of cumulative diagnoses, the number of deaths, city lockdown, the number of new diagnoses, the growth rate of deaths, etc. On the other hand, the authors point out that the logistic regression model loses its performance when there are multiple features in the model. The improved method used 21 days of input data to predict 7 days. For all regions studied, this method had performed better and more accurate results than traditional logistic and hill equation prediction algorithms.

Arora et al. (2020) present a model for predicting positive cases of COVID-19 for 32 states in India by evaluating the performance of recurrent neural network models and alternative LSTM models. In the same paper, the authors also present an approach that combines DL techniques such as Convolutional LSTM and Bidirectional LSTM to predict the number of positive cases on a daily and weekly frequency. Arora et al. (2020) also points out that, because of their performance, the same models may be suitable for analysis in other countries. A similar approach has been explored by Yudistira (2020), who uses LSTM models to check the disease expansion rate.

The approach proposed by Tomar & Gupta (2020) tries to make it easier to count COVID-19 cases to avoid the burden on the agents involved in identifying the case and controlling the infected people. Thus, the authors developed a LSTM network model to predict the total number of COVID-19 cases in India, indicating models that can predict the next 30 days of disease progression. Highlighting the importance of applying ML methods to different problems Lalmuanawma et al. (2020) have developed a study exploring different approaches related to the use of Artificial Intelligence (AI) applied to problems arising from COVID-19, which include both prediction and clinical analysis tasks, medication and the monitoring of infected people.

Regarding the performance of prediction models for COVID-19 cases Anirudh (2020) points out the challenges in the development of models with an epidemiological nature discussing approaches related to models such as SIR, SEIR, SEIRU, SIRD, SLIAR, ARIMA, SIDARTHE, and their performance in predicting the peak of the disease, the spreading evolution, and transmission rates.

In relation to the field of analysis of time series dynamics Pereira et al. (2020) presents a model to identify the behavior of the disease in several countries and the particular Brazilian scenario. The model uses clustering algorithms to verify the epidemic similarities in various regions where the disease is at an advanced stage.

In summary, all the works mentioned above demonstrate the importance and effectiveness of the computational nonlinear models to improve the data analysis related to the COVID-19 dynamics. Although better than other approaches, the DL techniques still does not allow good predictions about the emergence of second and third waves, as well as the end of the pandemic. However, such works helped us to choose better settings for our LSTM hyperparameters and encouraged us to consider the outliers, which are still little explored so far by COVID-19's monitoring and forecasting models.

Basically, inspired by the previous published work discussed in this section, we define our new analytical approach as follows: (i) we consider as input the daily measure of new COVID-19 cases from different scenarios (country level and also more regional levels , with emphasis on the favelas of Rio de Janeiro); (ii) we tested, adjusted and validated for each scenario an RNN-LSTM prediction model; (iii) we identify and discuss several anomalies that may result in scenario-specific outliers; (iv) we refined the analysis using the Linear Regression technique to impute possible values neglected by public policies. Based on the *Prophet* tool (Taylor & Letham 2017), we combine linear regression and statistical data imputation within a machine learning paradigm that we call here Adaptive Linear Regression. Materials and analytical methodology related to these approaches are detailed in the next section.

MATERIALS AND METHODS

To present the materials, methods and analysis used in this work, we defined the main stages of our study. First, we developed all data analysis and prediction models from the identification of COVID-19 outbreaks in scenarios with extreme fluctuations. Such patterns reflect a lack of control over infection suppression in competing scenarios (countries, regions, states, cities and local communities). Thus, this work seeks to address the daily COVID-19 complexity at different population scales with the aim of generating information that helps to reduce the rate of infections and, consequently, deaths caused by the disease in specific spatiotemporal scales.

Thus, based on identified cause and effect characteristics for each time series, we build LSTM models for each scenario. After modeling and parameterization, we train and validate each LSTM model. Validation occurs through statistical analysis of each series and forecast evaluation considering anomalies and outliers. If there is any anomaly in the time series, the data will be

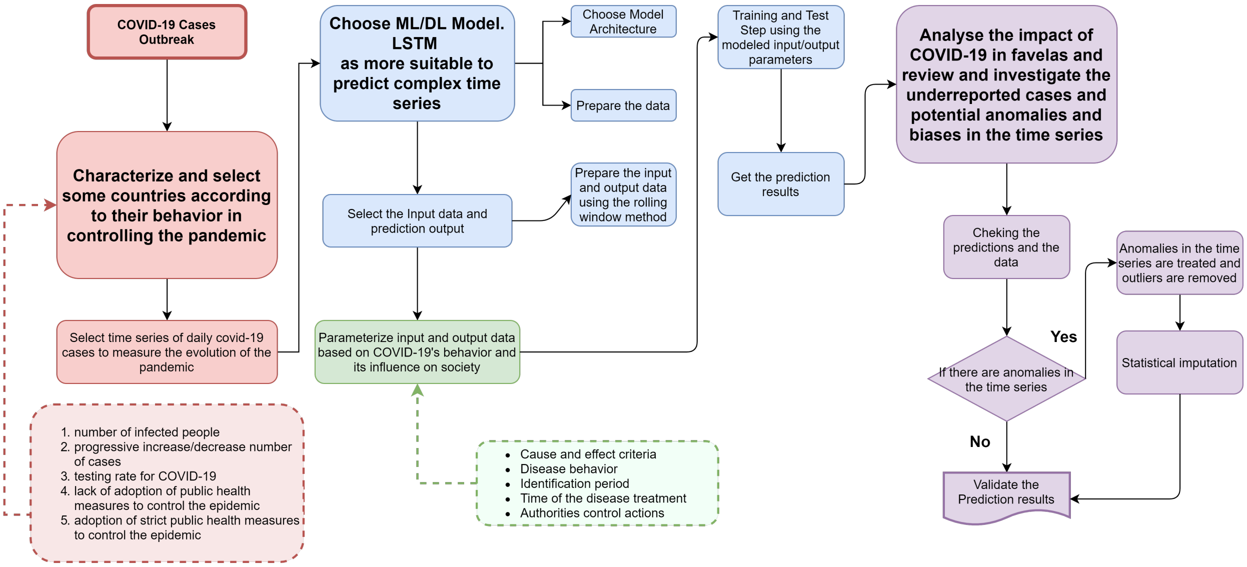

COVID-19 Cases Outbreak
Characterize and select some countries according to their behavior in controlling the pandemic
Select time series of daily covid-19 cases to measure the evolution of the pandemic
1. number of infected people
2. progressive increase/decrease number of cases
3. testing rate for COVID-19
4. lack of adoption of public health measures to control the epidemic
5. adoption of strict public health measures to control the epidemic
Choose ML/DL Model. LSTM as more suitable to predict complex time series
Choose Model Architecture
Prepare the data
Select the Input data and prediction output
Prepare the input and output data using the rolling window method
Parameterize input and output data based on COVID-19's behavior and its influence on society
Cause and effect criteria
Disease behavior
Identification period
Time of the disease treatment
Authorities control actions
Training and Test Step using the modeled input/output parameters
Get the prediction results
Analyse the impact of COVID-19 in favelas and review and investigate the underreported cases and potential anomalies and biases in the time series
Cheking the predictions and the data
If there are anomalies in the time series
Yes
No
Anomalies in the time series are treated and outliers are removed
Statistical imputation
Validate the Prediction results

Figure 1. Methodology overview diagram. In this figure, we focus on the main steps to treat the data, choose the model parameters and validate the model with an statistical imputation method.

submitted to the Statistical Imputation process, imposing an adaptive regression to the analysis. Figure 1 shows how the processes occur for the development of the models and data analysis. The methodology diagram aims to briefly indicate how each process is done to cover all the proposed methods and analysis of this approach, thus indicating to the readers all the main steps of the methodology to guide them in the manuscript reading.

The two approaches described in the first section are complemented by four methodological principles (different colors in the methodological diagram (Figure 1)), which constitute our analytical approach. The methods and parameters introduced in the diagram in Figure 1 are described in detail following the methodological order of approaching the data. It starts with the description of the input data considering the LSTM model, indicating the inputs to obtain predictions through a subsampling method. All necessary introductory material on the fundamental concepts of LSTM neural networks is described in detail in the Appendix. It then proceeds to describe the behavior of the disease (fluctuation of daily cases of SARS-Cov-2 infection) in the context of input data adopted by the LSTM models. We then proceed to select the parameters for training the models, including the number of models generated and the computational cost described as the processing time required for a predefined hardware. Methods related to the statistical treatment of fluctuations found in the daily data of reported COVID-19 cases are also explicitly considered in the approach. The main materials and methods related to adaptive linear regression models to find anomalies in data are based on the Prophet package. Data abnormalities are then considered and identified. Finally, the data with the anomaly are treated and normalized for insertion into LSTM models.

Data Scenario in the Context of LSTM Models

The effects of the pandemic differ in many aspects in different countries and their internal regions. In Brazil, such effects become at some point more diverse and unpredictable, since each state, region and city has had different strategies to face the pandemic. Therefore, our LSTM models are designed to understand the spread of the disease in distinct scenarios, with more attention to the case of Brazil and in communities, such as favelas, where basic hygienic and sanitary conditions are limited. In a such complex scenario, where long-term effects get easily out of control, daily infection cases monitoring, when possible, is most desirable.

A further point of interest concerns the personal diagnosis of SARS-CoV-2 infection. It is well known that each studied scenario had a different approach to testing for COVID-19. According to Worldometer[3] Brazil has a low level of testing 6.94 per million (1M) inhabitants compared to other countries Italy 75.48 / 1M, USA 73.28 / 1M and South Korea 21.35 / 1M. Figure 2 shows testing data from 4 countries in completely different situations regarding mass testing of their population, with Brazil being the one with low levels on the moving average of 7 days for testing per million inhabitants. Thus, methods that can predict the evolution of the disease in places with lower testing rates become essential.

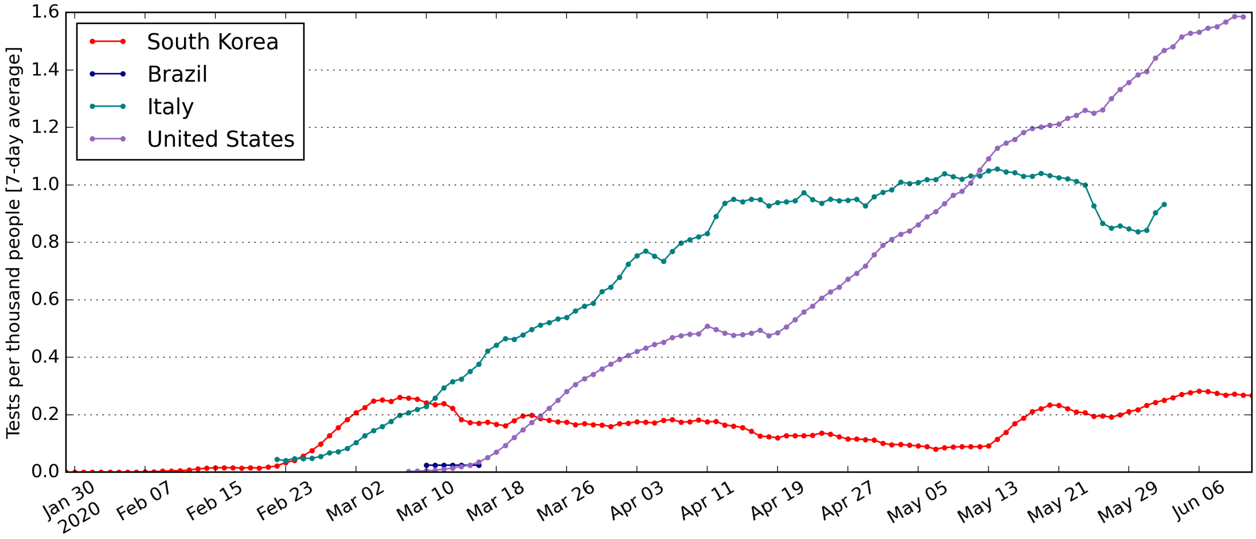


Figure 2. Brazil has a low level of daily COVID tests according to Coronavirus Pandemic (COVID-19) publication data (Ritchie et al. 2020).

For this purpose, we have developed LSTM models to deal with each country and region aside. Data used to feed the predictive models refer to the Our World in Data (Ritchie et al. 2020) database from January to July 2020. For particular regions, we have applied a database concerning the occurrence of cases in communities (favelas) in Rio de Janeiro/Brazil, provided by Voz das Comunidades (2020) from April to August 2020.

As each county exhibited different behaviors in dealing with the pandemic, we looked at the main features that distinguish them, such as the decrease in their infection rate or a significant evolution in the number of daily cases of COVID-19. Such observations provided a basis for predictive models that learn from different and simultaneous fluctuations in the time series data. In this way, we have

[3] Information about tests of COVID-19 by countries, source Worldometer available at https://www.worldometers.info/coronavirus. Accessed on June 12, 2020.

selected and grouped each country into different control situations corresponding to the pandemic stage in each nation. Countries have been grouped as follows:

- Critical Group: Brazil, USA, India, South Africa and Russia;
- Attention and Impact Group: Australia, Portugal, China, Italy and Switzerland;
- Successful Group: South Korea and New Zealand.

The **Critical Group** corresponds to the countries where the COVID-19 pandemic grew out of control, leading to a large number of daily cases and deaths. In these countries, the politicization of the pandemic has caused important divergences in society, undermining the measures to control it. In these countries, the number of daily cases still evolves in many days, and it is difficult to decrease the number of infections and deaths. Figure 3 shows the evolution of COVID-19 cases on the logarithmic scale for the critical group.

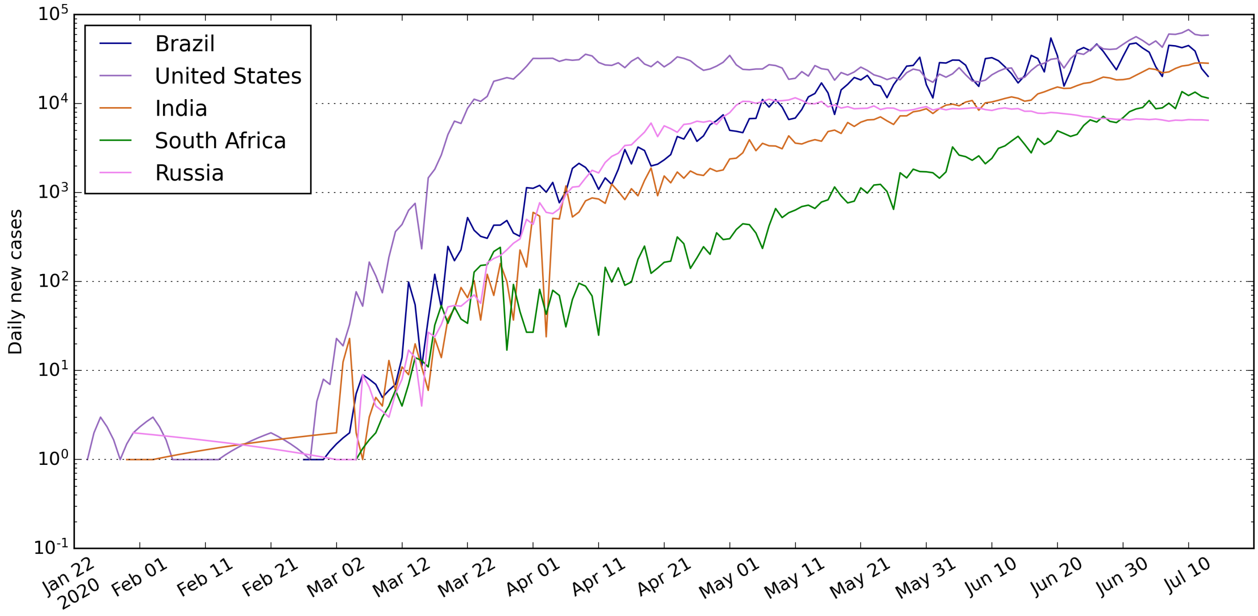


Figure 3. COVID-19 daily cases evolution for critical group countries in log scale. Source Ritchie et al. (2020).

The **Impact and Attention Group** consists of those countries which at the onset of the outbreak have not taken effective epidemiological control policies and restrictions, which has resulted in a high number of cases and deaths over a short time. In such cases, the restrictive control epidemiological policies have been controlling the growth of the epidemic. This group also corresponds to countries that have adopted epidemiological control policies which are becoming less effective and are on the verge of showing growth in the number of daily cases. Figure 4 shows the evolution of daily cases on the logarithmic scale for the attention and impact group.

The **Successful Group** consists of those countries that have adopted severe control measures and, thus, these countries have been able to contain the pandemic. Such countries adopted strong restrictive measures on people's circulation, a high testing rate per million inhabitants, and a high potential to track and control the infected people. It should be pointed out that China has adopted

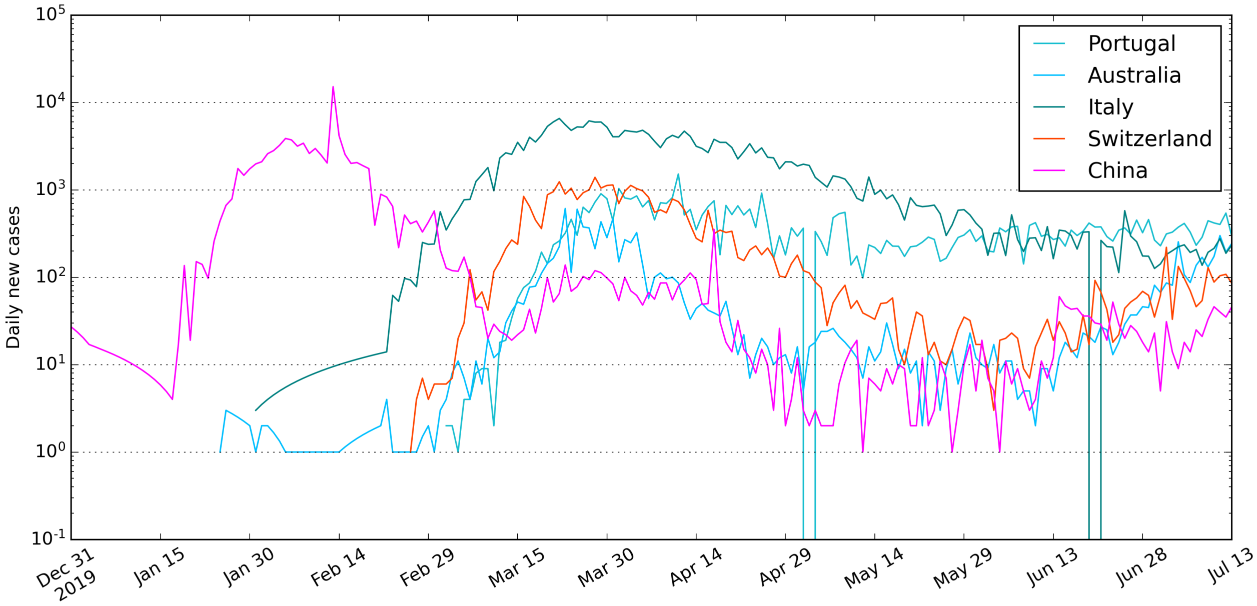


Figure 4. COVID-19 daily cases evolution for attention and impact group countries in log scale. Source Ritchie et al. (2020).

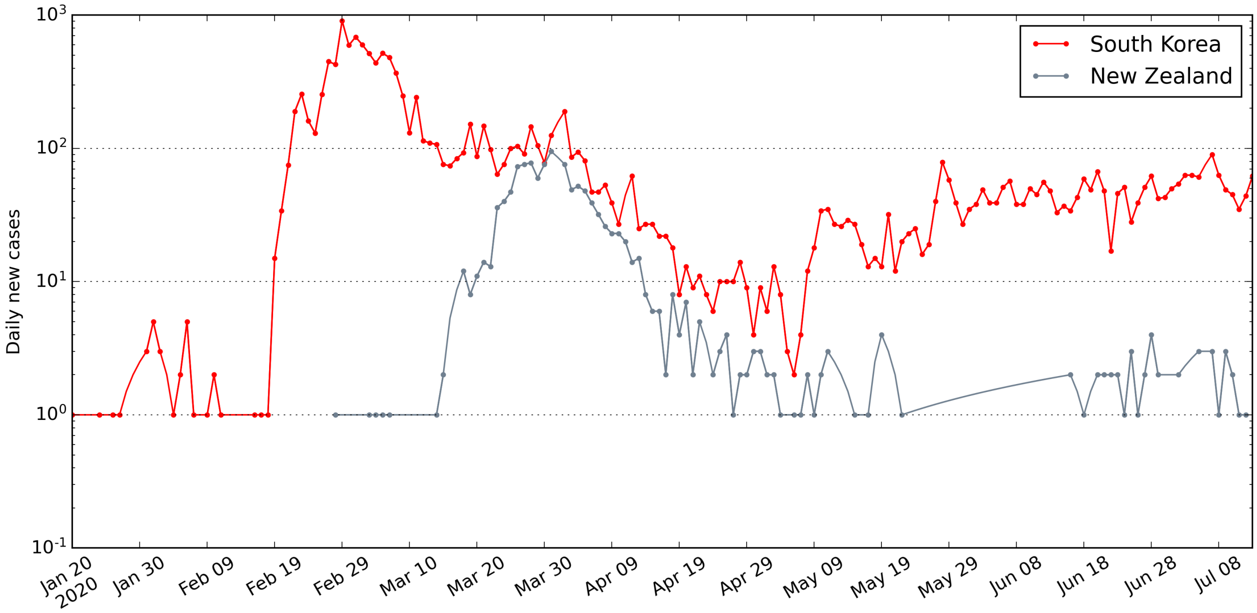


Figure 5. COVID-19 daily cases evolution for successful group countries in log scale. Source Ritchie et al. (2020).

severe epidemiological policies, however, these actions took place after the outbreak of the disease. Figure 5 shows the evolution of cases on the logarithmic scale for the successful group.

By looking at these different scenarios (shown in the figures 3, 4, and 5), we avoided a global model to predict daily cases, so we considered a different model for each country or region. That is, we conducted different training steps for each country and region.

Thus, we present a methodology to develop prediction models for daily COVID-19 cases based on the observation of different epidemic scenarios. As input, the prediction models receive the number of days in which the time-series learning strategy must be done to estimate future points, i.e., the number of predicted cases. Different types of models indicate possible ways to feed a time-series forecasting model that can observe N points in time $\{t = n, t = n-1, t = n-2, t = n-3, ..., t = 1\}$ to estimate the correspondent next values $\{t = n+1, t = n+2, t = n+3...\}$. An output y in time $\{t = n+1\}$ may be dependent on the correspondent inputs X $\{t = n, t = n-1, t = n-2, ..., t = 1\}$.

In the artificial neural networks context, the definition of inputs and outputs is highly important to the behavior and performance of regression, classification, and predictive models. For example, in a time series, every input $X[i]$ corresponds to the features that must be learned and mapped to their respective outputs $y[i]$. Machine learning models, which are the models described in this paper, work as learning a target function f that maps the points of the time series (x) to their respective outputs y, that is, the next points in the time series (prediction points), so that $y = f(x)$.

Our prediction models require the correct declaration of inputs and outputs for their proper use. For time series, the input X and output y data must be arranged according to the framework and definitions of the LSTM models $\{1..1, 1..N, N..1, N..N\}$. We have arranged the data for the prediction of the LSTM models by sub-sampling through a rolling window method (Zivot & Wang 2007).

The Rolling window methodology consists in choosing some parameters that will be key factors for the analysis and forecast sensitiveness of a model, such as window size, m, that is, the size of an input that corresponds to the number of consecutive window observations. The size of the rolling windows depends on the size of the time series, T. An additional parameter is the prediction horizon, h, i.e., the output prediction size of the models. This prediction horizon depends on the features of the data and the intended application. Figure 6 shows how the inputs and outputs pass through the time series creating rolling windows through the data. In this paper, each time series of daily COVID-19 infections has been applied to the rolling window sub-sampling method where each window is subjected to pattern learning by LSTM networks that estimate the prediction horizon, i.e., the model outputs to predict new daily cases.

The architecture of LSTM models was defined after a set of test and training approaches where the parameters were varied for each training epoch. Those parameters were set according to the RMSE value in the tests, i.e., we chose the architecture that provided the best RMSE value for the predictions. Table I summarizes the model architecture setting the main described LSTM hyper-parameters.

Table I. LSTM models overview.

Input Layer	**Hidden Layer 1 (LSTM cells)**	**Hidden Layer 2 (LSTM cells)**	**Output Layer**	**Activation function**	**return sequences**	**Optimizer**
N	100	200	N	ReLU	True	Adam/Stochastic Gradient Descent

The models were built with four layers: i) the first layer containing corresponds to the number of input points; ii) LSTM memory cells (100 neurons); iii) LSTM memory cells (200 neurons) and iv) the output layer corresponds to the prediction horizon. These two hidden layers (memory cells)

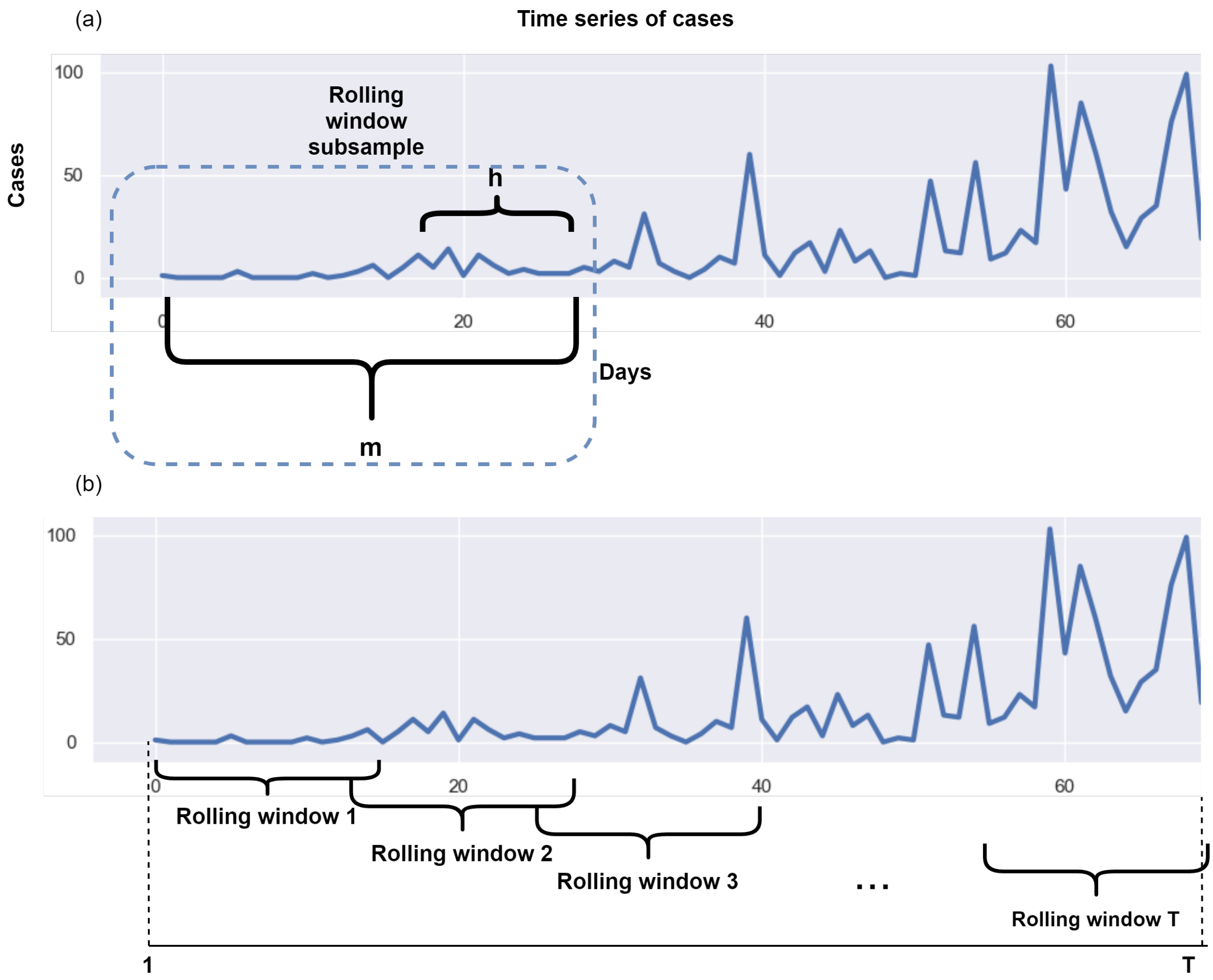


Figure 6. Rolling window sub-sampling method. In (a) we show the rolling window sub-sampling, this method allows us to choose the window size (m), that is, the size of input data to get the prediction horizon (h), that is, the prediction output. In (b) we show how this sub-sampling method runs through the entire time series, setting the inputs and prediction outputs.

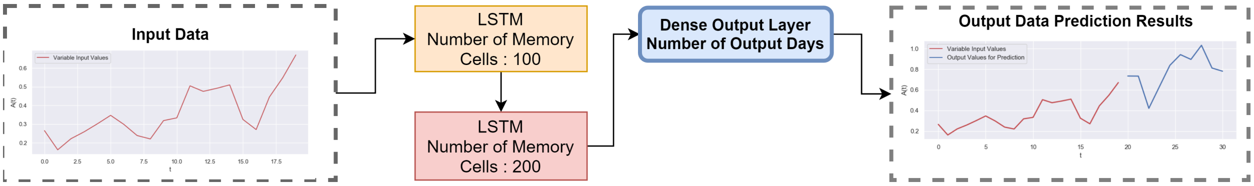

Input Data
Variable Input Values
LSTM
Number of Memory Cells : 100
LSTM
Number of Memory Cells : 200
Dense Output Layer
Number of Output Days
Output Data Prediction Results
Variable Input Values
Output Values for Prediction

Figure 7. LSTM Model for Daily Coronavirus Cases Predictions.

are activated by the ReLU function. The developed LSTM models are based on the so-called *N*...*N* architecture (many input points allow the prediction of many output points) (Figure 7).

Parameterization of LSTM Models Based on Cause and Effect Criteria

We have attempted to contextualize the dynamics of the disease corresponding to how the infections spread and the time over which an infected person is diagnosed with the disease, thus resulting in a new case. The goal of this contextualization is to establish scientific criteria in the parameterization of LSTM models for the prediction of new daily cases of COVID-19 and, consequently, observing the eventual progression or stability of the pandemic.

These LSTM models are highly sensitive to parameter values set for the chosen input and output days. Thus, it is important to understand the dynamics of the infection and how new cases are reported to establish strong criteria for choosing these days. This characteristic of RNNs can lead to some inconsistencies in replicating the prediction results. Hence, we have defined the following rules for choosing the parameters for the input and output days of the model:

- **Disease behavior**: How soon are the first symptoms noticed?
- **Identification period**: After the appearance of the first symptoms, for how long are healthcare actions taken? Such as seeking medical attention or tests to identify the COVID-19 infection.
- **Time of disease treatment**: How long does a person identified with COVID-19 reduce their viral load and recover from hospital treatments?
- **Authorities Control Actions**: After identifying a large number of infected people and establishing an epidemic state, how soon are epidemiological control actions, such as lockdown, taken?

Disease Behavior. According to Tian et al. (2020), the infections related to COVID-19 can present different scenarios, such as severe cases, mild cases, cases without pneumonia, and asymptomatic cases. This study covered about 262 patients in several hospitals in Beijing in February 2020, when COVID-19 cases in China became internationally notable. Several cases and symptoms affect the way people seek medical assistance to identify what is the cause of the infection. The common symptoms observed in the study conducted by Tian et al. (2020) include fever, cough, fatigue, loss of olfactory and taste sensors, and headache, the combination of one or more symptoms characterizing the severe condition. The important observation made by Tian et al. (2020) was that the average period of virus incubation in the patient, that is, the period until the first symptoms of the disease appear, was 6.7 days. The same behavior was observed by another study by Lauer et al. (2020) that considered the characteristics of the disease and its spread in 181 patients who had symptoms. Lauer et al. (2020) observed that the estimated virus incubation period varied between 4.5 to 5.8 days, and the development of symptoms occurs between 8.2 to 15.6 days.

Identification Period. According to studies that pointed out the time when the infected person develops the first symptoms of the disease, the identification period criterion is related to the time when a person, aware of infection symptoms (not necessarily aware of the infection by COVID-19), seeks medical care. According to Tian et al. (2020) the average time these patients sought medical assistance was about 4.5 days after the occurrence of first symptoms. Regarding the study made by Lauer et al.

(2020) 97.5% of the people sample took an average of 11.5 days to show the illness symptoms and consequently seek medical care.

Time of disease treatment. After the initial symptoms identification and the seeking of medical care, a relevant measure for controlling LSTM models is connected with the time necessary for a patient to be recovered of the disease. This pertinent remark has been discussed by Zhou et al. (2020) and turns out to be a crucial factor that implies decisions to isolate patients, make resources available for treatment, and then make projections regarding the availability of intensive care unit beds specifically for severe respiratory illness. Thus, Zhou et al. (2020) observed the average time from infection discovery to the end of hospital treatment in a group of 41 patients hospitalized in the Western District, Union Hospital of Tongji Medical College, from February to March 2020. They discovered that viral load reduction in hospitalized patients occurs in a period of 24 to 40 days after identifying COVID-19 infection. The shortest period to eliminate and reduce viral load was 18 days, while the longest was 48 days. There were no significant differences in this period for men and women and considering the ages in people over 65 and under 65. Regarding the time a patient was discharged from the hospital, the average period was 32 to 46 days. The shortest period observed for a patient to be considered recovered was 24 days, while the longest was 56 days. No major differences in time were observed with respect to the age of the patients or with respect to the gender of the patients.

Authorities Control Actions. The epidemiological control action criterion adopted by authorities (governments, public officials, scientists, etc.) is related to the first months in which the COVID-19 event was observed in a country. Due to the observation of the days when the disease affects the population and how long people seek medical assistance (Lauer et al. 2020, Tian et al. 2020, Zhou et al. 2020) taking control actions become difficult and must be done in an emergency scenario. The first countries to experience an outbreak with the Sars-CoV-2 virus were China and Italy. According to Chen & Yu (2020), in China, the actions has taken followed a total emergency criterion, as there was no knowledge of the disease COVID-19 and the number of hospitalizations and the number of cases could collapse the health system. In Italy, the epidemiological actions were slowly, although already known the epidemic in China. Chen & Yu (2020) consider that strict actions should be taken as outlined below:

1. identify those infected and take them to hospital treatment for infectious diseases;
2. locate and quarantine all those who have had contact with those infected;
3. sterilize environmental pathogens;
4. promoting the use of masks and disclosing to the public the number of infected, suspected, under treatment and daily deaths and cases;
5. provide for closure of all public and private activities that promote crowds of people. Establish a closure of all non-essential activities for some time.

Observing these control actions, in the Chinese case provided an epidemic regression in about 14 days (Tian et al. 2020). According to Chen and Yu (2020), after the detection of the COVID-19 outbreak, between January to February 2020, a considerable and accelerated decline in the number of cases from February 18 onwards. This was highlighted when the daily number of infected people dropped from 44 to 2 in a period of 11 days, evidencing the efficiency of the actions adopted by this country. In cases where control actions were slow, the number of infections and deaths have an expressive growth, it was observed in Italy, the USA, and Brazil. In these countries, the epidemic of COVID-19 has been politicized and underestimated, the lack of adopting rigorous epidemiological actions has resulted several cases and deaths. Looking at the overall pandemic scenario, by August 2021 only in Brazil and the USA about 63 million cases and approximately 1.3 million deaths were counted. We cannot underestimate the pandemic's severity, since around the world more than 200 million cases of infections and approximately 4.5 million deaths have already been recorded (Ritchie et al. 2020, Worldometer 2020, Voz das Comunidades 2020).

Cause and Effect criteria

Based on the scenarios described above where some typical time windows, of cause and effect, were empirically found, we decided to make the LSTM model more realistic from them. Therefore, the choice of parameters for the LSTM models summarizes this contextualization in a pandemic **Cause-Effect** principle.

As **Cause** object we have chosen the criteria of **Disease Behavior** and **Identification Period**, which are consistent with the initial perception of symptoms and the identification of the disease, observing both the clinical evaluation and testing methods. The **Cause** object represents the number of input days for the prediction in the LSTM models.

As **Effect** object we have chosen the criteria of **Time of Disease Treatment** and **Authorities Control Actions** that are consistent with the actions to combat the pandemic itself. It represents the intensive action of health professionals as the adoption of severe control actions to decrease the number of infected people. The **Effect** object represents the number of output days for the prediction in the LSTM models. According to the established criteria, the input values of the models have been chosen as:

- Days when the first symptoms appear: 5, 7 and 10 days on average;
- Days when symptoms are detected and medical diagnosis occurs: 10 and 15 days on average;

In this way, the cause object comprises the range of 5 to 15 days as input parameters for the forecasting models. According to the established criteria, the output values of the models have been chosen as:

- Considering the period in which the identification of the disease occurs (5 to 15 days) and the period which the treatment lasts from 32 to 40 days average, the range of this criterion is between 15 to 25 days after disease detection;
- Considering the period of action occurring as soon as the COVID-19 outbreak is detected, implying the crowding of hospitals and an increasing number of deaths, the effect of control actions is observed from 10 to 15 days. We can estimate that the range of this criterion occurs between 25 to 40 days after the identification of the epidemic.

In this way, the effect object comprises the range of 15 to 40 days as output parameters for the forecasting models. Thus, the Table II indicates the parameters of input and output days for the LSTM models to estimate the daily infection cases of COVID-19. We set all the LSTM models with 1000 training epochs, and the processing time of each one has varied according to the input/output parameters.

Table II. **Input and output parameters for LSTM models.**

Input days	Output Days Range 1	Output Days Range 2	Output Days Range 3	Output Days Range 4	Output Days Range 5	Output Days Range 6
5	15	20	25	30	35	40
7	15	20	25	30	35	40
10	15	20	25	30	35	40
15	15	20	25	30	35	40

Thus, we have developed 288 prediction models to forecast new daily cases of infection for each country in the groups, estimating at least 15 days of the decrease or progression of COVID-19 cases. In this paper, we only selected the LSTM models that achieved the best validation result based on RMSE score. Processing time for all 288 models took 3 days on a 6 Gigabyte NVIDIA Ge-Force GTX-1060 platform.

After training the models and choosing the best prediction results, we perform a careful analysis of the time series to identify the quality of the data regarding the daily cases of COVID-19. In this process, we seek to identify if the data are being correctly computed, if there are inconsistencies in the time series, if there are anomalies coming from the sources themselves, such as data collected from health agencies, and finally, if there are unexpected seasonalities that do not match the actual behavior of COVID-19. After going through this analysis the prediction models can be validated.

Thus, we discuss these issues and approaches in the next subsections dealing with statistical analysis and time series treatment.

Dealing with the under-reporting case problem

The fast spread of the COVID-19 epidemic around the world has had an important impact on the health systems of several countries. Although some have shown greater resilience, such as New Zealand and South Korea, due to mass testing and restrictive measures such as lockdown, others have not been sufficiently prepared, for political and ideological choices (The Lancet 2020). In this context, countries like Brazil and USA have presented high rates of COVID-19 cases and deaths. In addition, technical and structural problems have repeatedly shown cases of underreporting on weekends in Brazil, even when the trend indicates that the number of cases is growing or remaining at a high level above 30,000 daily cases. In order to present a closer picture of the real number of cases throughout the epidemic in the country, we conducted a study to identify anomalous values[4] in the historical series of case records and applied techniques to input values where the data present behavior of outlier in relation to the real dynamics of the disease in Brazil.

[4] The definition for anomaly depends on the domain of the problem, but in general we can assume that an anomaly is an outlier point. “An outlier is an observation that appears to deviate markedly from other observations in the sample NIST/SEMATECH (2013).”

Abnormal weekly seasonality in Brazil daily cases

The Brazil and the USA are experiencing similar situations in this pandemic, with high numbers of new cases, above 30,000 on average, and more than 1,000 daily deaths. However, as we can see in Figure 8, notifications about the disease show different behaviors in both. The number of cases in Brazil reported by state health agencies and compiled by Worldometer specialists[5] show recurrent decreases at the end of the week and the beginning of the week, while in the USA[6] this weekly seasonality is not demonstrated.

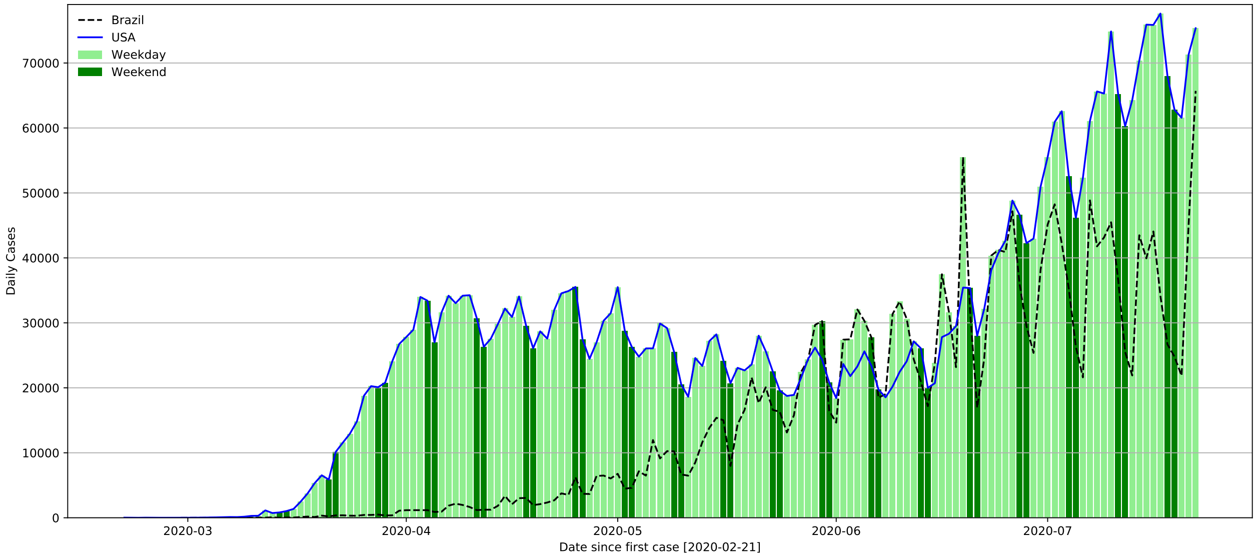


Figure 8. Number of daily cases in Brazil and USA.

In order to deal with the seasonality effects of the data in both countries and to identify anomalies in the time series, we used *Prophet*, an open source framework for time series forecasting proposed for Taylor & Letham (2017) and released by Facebook's Core Data Science team[7]. This framework was initially developed for business forecasts and enable estimate non-linear trends fitted with yearly, weekly, and daily seasonality, plus holiday effects to predict future and detect anomalies. It has been employed in COVID forecasting, for example, Wang et al. (2020) using the characteristics of historical data to predict the number of accumulated confirmed cases, recovered, death, and active confirmed cases of COVID in the global and regional context for some countries (Brazil, Russia, India, Peru, and Indonesia).

The Prophet is based on assumptions of an additive regression model $y(t)$:

$$y(t) = g(t) + s(t) + h(t) + \mathcal{E}_t, \tag{1}$$

where $g(t)$ is a trend function to model non-periodic changes of time series, $s(t)$ is a component to model seasonality representing periodic changes, such as the periodicity a week or a year, $h(t)$

[5] Worldometer homepage with information about the COVID spread in Brazil available at https://www.worldometers.info/coronavirus/country/brazil. Accessed on July, 22, 2020.

[6] Worldometer homepage with information about the COVID spread in the USA available at https://www.worldometers.info/coronavirus/country/us. Accessed on July, 22, 2020.

[7] The Prophet framework available at https://facebook.github.io/prophet. Accessed on 01 July 2020.

contributes information about holidays and recurrent events, and finally the error term $\mathcal{E}_t$ represents any idiosyncratic changes which are not accommodated by the model (Taylor & Letham 2017).

Prophet was created to permit further customization of the model using intuitive and easily interpretative options for improving the quality of the forecast. These parameters can be adjusted by non-experts in time series analysis, enabling optimizations and better results. Because the framework handles the main problems of scale such as computational and infrastructure, this way it is possible making forecasts without specific knowledge about time series training and methods only with a focus in the data generating process. Moreover, the Prophet have automated means of evaluating and comparing ensemble models to detecting when they performing poorly (Taylor & Letham 2017). Then, we use the Prophet in a context we call Adaptive Linear Regression (ALR). In this context, the algorithm admits incorporation of prior knowledge to find the best regression curve, thus behaving as a machine learning solution.

In this study, we have used Prophet to show that in Brazil there are anomalous values related to delays in the transmission of daily cases and the under-reporting of COVID-19 cases on weekends and Mondays. Although many countries have experienced difficulties in estimating the real number of infected people, in Brazil this is aggravated due to a small testing rates. According to Ribeiro & Bernardes (2020), the main factors which contribute for this are "the absence of adequate laboratory infrastructure and qualified people, which are often not available in the appropriate quantity; difficulty in buying tests due to high international demand and low availability of suppliers; logistical difficulty in the national distribution of tests in a country of continental dimensions such as Brazil".

We known that it is very difficult to model perfectly the COVID-19 dynamics because there are several factors can influence dissemination speed, for example, when countries close borders or determine lock-down and as a result of human behaviors like as social distancing. It changes all the time as the infection spreads, for this reason, the main goal of our approach was to demonstrate that in Brazil the disease cases show abnormal periodic changes (weekly) in time series. For that, we create an instance of Prophet to fit a model using data of Brazil and the USA simulating this seasonality, because both countries present similar characteristics in this pandemic.

Figures 9 and 10 show the results of fitted model using the number of daily cases (black circles) in Brazil and USA respectively, the band with 80% of uncertainty intervals (green band) and anomaly points (red circles).The importance size was determined based on the distance between the value and the upper or lower limit of the uncertainty band.

We perform trends analyses for Brazil and the USA using the options to identify weekly seasonality. Brazil presents an abnormal seasonality (weekly) that can be corroborated by the fact of presenting fewer anomalous points than the USA in this analysis. This behavior (weekly) does not reflect the real dynamics of the disease anywhere in the world. This evidence can be confirmed by analyzing of the components of weekly seasonality for both countries (Figures 11(a) and 11(b)). This demonstrates that in Brazil there is a great decrease in the registration of cases on weekends and Mondays.

Identifying anomalies in the data

This way is very important to identify and handle anomalies (outliers) in time series data of cases in Brazil to reflect better the true situation of COVID spread in the country. Outlier identification depends

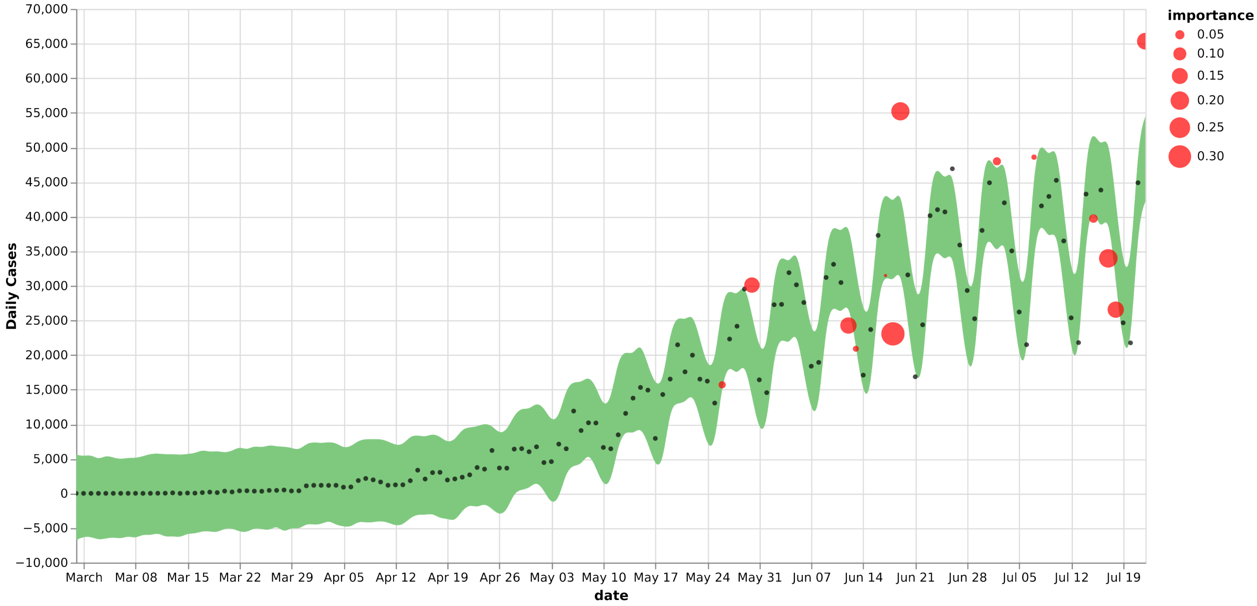


Figure 9. The confirmed cases reported in Brazil (black dots), line with an 80% confidence interval (green band), and the importance of outliers is annotated based on how far the dot is from the boundary of the confidence interval.

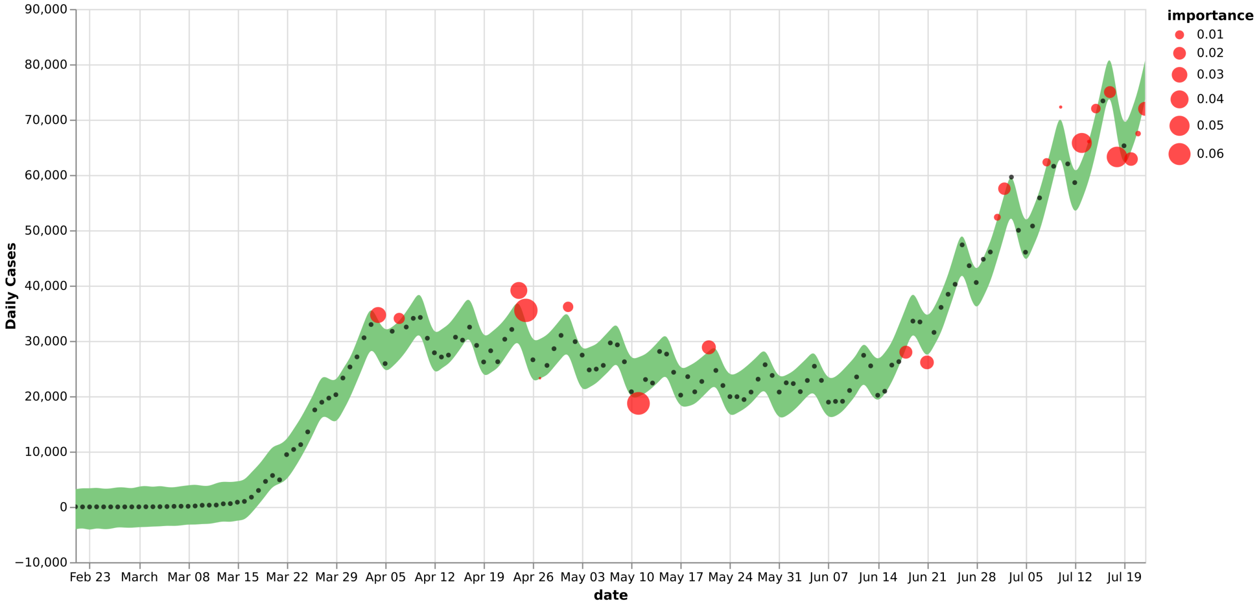
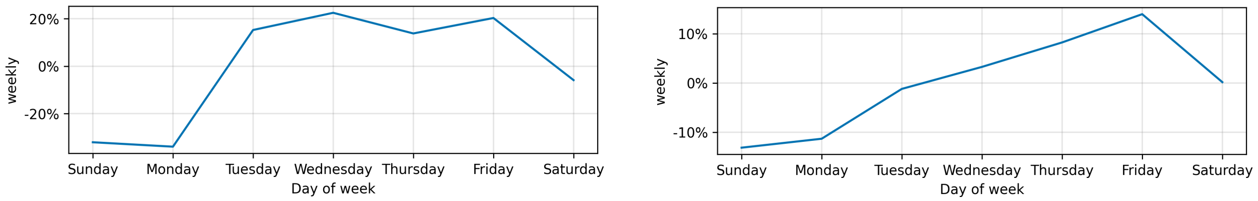


Figure 10. The confirmed cases reported in the USA (black dots), line with an 80% confidence interval (green band), and the importance of outliers is annotated based on how far the dot is from the boundary of the confidence interval.

on how you able to define the unnatural behavior of the data. The methods of general-purpose for this are based on robust statistics, such as least-squares techniques, residual variance ratios among others (Tsay 1988).

Basically, our motivation was to initially identify the points in the data series that correspond to very high numbers of the cases (upper outliers) linked to notification delays, so that these points



(a) Brazil **(b) USA**

Figure 11. The components plot of weekly seasonality for both countries.

were removed to did not influence the identification of cases of underreporting on weekends (lower outliers). Thus, it was possible through a statistical method of data imputation to correct the time series of daily cases with the objective to demonstrate a scenario closer to the reality of the COVID epidemic in Brazil.

Detecting outliers in time series is particularly challenging because the process is based on old data. In the cases of diseases like COVID where the outbreak revealed nonlinear and chaotic characteristics (Chen & Yu 2020), this process can be even more complex. Normally, three issues with regards to outliers can be specified: Outlier labeling, when we need to flag potential outliers for further investigation; Outlier accommodation, when we cannot determine potential outliers and more robust statistical techniques can be employed to take into account these observations; Outlier identification, when we apply formally test to reveal whether observations are outliers (NIST/SEMATECH 2013).

Most of the methods for detecting outliers are based on the notion that the data have a normal distribution, but our data with Skewness = 0.7962709 and Kurtosis = -0.4209722 clearly don't present this expected distribution[8]. This way, as the normality assumption for the data, is not valid, we decided to use a method based on a robust moving window test over data to detect outliers. The Median Absolute Deviation (MAD) is a method that defines the variation of the data related to the median, this is a stronger measure than the standard deviation. Because the median used as a measure of central tendency shows the advantage of being very insensitive to the presence of outliers. Additionally, the MAD is independent of the sample size (Leys et al. 2013). This method enables detect outliers continuously based on old data without the necessity to repeat tests, showing a low rate of false positives.

The calculating of MAD involves finding the median of absolute deviations from the median and can be expressed as:

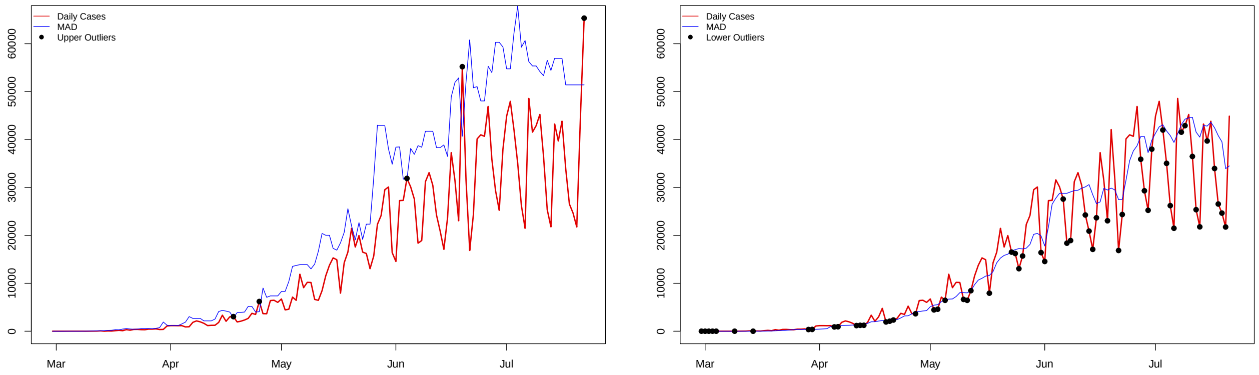
$$MAD = b * median(abs(x - m)) \tag{2}$$

where m is the median of values from vector x, the dimension of this vector is defined from the data days window used. The b constant is linked to the assumption of normality of the data, but disregarding the abnormality induced by outliers, the standard value defined for the normal distribution is $b = 1.4826$ Leys et al. (2013). We assumed $b = 1$ related our data distribution.

Our approach is based on the use of graphs to define the best adjustment of the time window size to be considered for the calculation of the median and the threshold value to remove the

[8] Skewness and Kurtosis are statistics of data distribution, the first measures lack of symmetry and the second is a measure of the heaviness of the distribution tails. A normal distribution (symmetric data) Skewness is zero, and Kurtosis equal three (NIST/SEMATECH 2013).

upper peaks relative to lag communication of daily cases, and in a second step identify lower peaks related to underreporting outliers. The elimination of upper peaks is necessary as they influence the determination of the expected variation in the data. Figure 12 shows results about outliers identification on Brazil's daily cases of COVID using MAD. For outliers related to lags on cases notification (upper peaks), the method applies a rolling window on data (red curve) with 10 days and median + 3 * MAD as a threshold (blue curve), the data values above this curve are labeled as outliers (black dots) (Figure 12(a)). It is important to mention that the threshold can only be computed until the final centralized window. For all data after this final window, the last computed threshold is used, that is why the blue curve is flat between x_{last-5} and x_{last}. After the identification of peaks, these anomalies points were modified using a median in a centered window with 15 days to correct the weekday trend and permit detect the points of underreporting cases. For outliers related to underreporting daily cases, the rolling window applied has 10 days and median + 0.5 * MAD as a threshold enable detecting valleys points (Lower Outliers) for all values below the blue curve (Figure 12b). It is important to note that this approach seeks to determine the bulk of the data in a way that is not influenced by any outliers, where the tuning of parameters (threshold and window) aims to identify how much deviation from the mass of the data should be considered an anomaly and how far in time should one look to characterize points that not fit well with the data. The main advantages of this method are: simple and robust moving window outlier test over the data; the data do not have to be normally distributed and periodic; they don't have necessary to be positive.



(a) Outliers from lags on cases notification

(b) Outliers from underreporting cases

Figure 12. Outliers identification in data of COVID daily cases in Brazil using the Median Absolute Deviation (MAD).

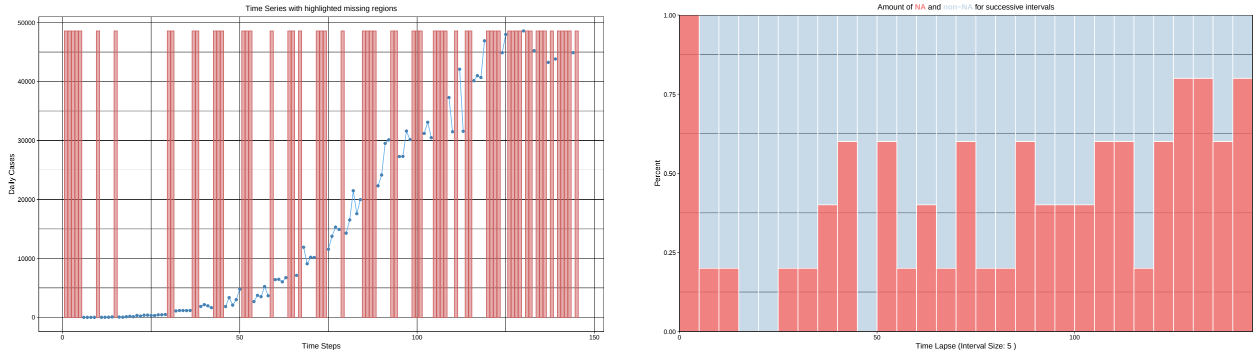
Applying statistical imputation to handle with outliers

Whatever the domain of data considered: industrial, biological, financial, social, and health. All of them potentially generate data series that suffer from the same problem, the missing values (Moritz & Bartz-Beielstein 2017). The use of statistical methods to replace unknown, unmeasured, or missing data with a certain value is called data imputation. In general, imputation can be performed by replacing these data with the mean of the variable considered, although better results can be achieved with the use of methods that take into account the correlation between values of data series. Its application is mostly used in surveys, but some computer experiment applications employ imputation to handle missing values NIST/SEMATECH (2013).

According to Moritz & Bartz-Beielstein (2017), time series imputation is considered a special sub-field of the imputation research area, where the main techniques like Multiple Imputation, Expectation-Maximization among others can't be applied directly, because they rely on inter-attribute correlations. While univariate time series (one attribute) use the inter-time correlations. An univariate time series can be defined as a sequence of single observations $o_1, o_2, o_3, ..., o_n$ on at successive points in time $t_1, t_2, t_3, ..., t_n$. The univariate time series is considered one column of observations and time is in fact an implicit variable (Moritz et al. 2015).

There are a lot of methods that can be employed to treat missing values using simple imputation like mean or learning approach like Self-Organizing Maps (SOM) Sorjamaa (2010). However, only since 2017 after imputeTS package of R proposed for Moritz & Bartz-Beielstein (2017), came into being an approach with algorithms considering time series aspects, because classical methods like mean imputation usually don't perform well in this context. It is practically impossible to point out a single method as the best in all cases, as performance generally depends on the intrinsic characteristics of the data. On one hand, data with a strong trend or seasonality perform better with methods based on Kalman Filter or Seasonally Decomposition/Splitting. On the other hand, sometimes imputation with mean values can be an appropriate method (Moritz & Bartz-Beielstein 2017).

In this context and based on the difficulty of evaluating the performance of the imputation algorithms for real missing data like ours (Figure 13), because a performance comparison can only be done for simulated missing data (Moritz et al. 2015). We adopted the StructTS[9] pointed out by Moritz et al. (2015) as one of the best general methods in their research comparing various methods for univariate time series imputation in R. Another reason which guided our choose was the fact that this method presented the best result in a dataset with behavior similar to our data (no seasonality/with trend).



(a) Distribution of missing values **(b) Missing values per interval**

Figure 13. Characterization of the distribution and percentage of the missing data in daily cases of Brazil.

We have adopted the following steps to handle the anomaly values of COVID daily cases in Brazil:

- Step 1 - Apply imputation by next observation carried backward to fill the gap at the beginning of the time series;

[9]StructTS function implement imputation by Structural Model & Kalman Smoothing (Moritz & Bartz-Beielstein 2017).

- Step 2 - Apply imputation by last observation carried forward to fill the gap at end of the time series;
- Step 3 - Perform imputation using interpolation based on seasonal Kalman filter on a structural model (StrucTS).

The Kalman filter provides the means to update the state of new observations, where the state space form (SSF) is the key to handling structural time series. Basically, the state represents various unobserved components such as trends and seasonals. Forecasting is made by extrapolating these components into the future using state-of-art smoothing algorithms related to the Kalman filter to obtain the best estimate for missing values. Important to note that both prediction and smoothing depend on parameters which model stochastic behavior of variables have been estimated, these parameters called hyperparameters, are also based on the Kalman filter. The reason for this is that the prediction errors expressed in terms of one-step-ahead predictions from the likelihood function emerge as a by-product of the filter. The space framework provides a way to estimate hyperparameters carried out in the frequency domain besides allowing to modeling of non-linear effects and structural change (Harvey 1990).

RESULTS

Forecast Results for Country Groups

The forecast results concern the models that achieved the best performance regarding the RMSE score. The RMSE metric is applied to check the differences between the predicted values ($Predicted_i$) and the true values ($True_i$) of the time series. It is calculated by the Equation 3 according to Willmott (1982).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Predicted_i - True_i)^2} \tag{3}$$

The obtained results are described in Table III. We explain the most expressive results obtained in the evaluation of 288 prediction models for daily cases of COVID-19 in 12 different countries. The table III shows the best validation result based on the analysis of the RMSE score, the input days, the output days (prediction), and the combination of parameters that reveal the best and the worst results of the validation.

Description of results for the Critical Group: The results for Brazil and India are similar concerning the input and output parameters that produced the best RMSE scores in the time series validation. Brazil and India obtained as best prediction models those that received 15 days as input (as an object of Cause) and produced 20 days of output (prediction) (as an object of Effect).

The LSTM models for Russia and South Africa demonstrate similar indices in the RMSE score and produce the best result in the prediction for the next 15 days receiving as input days a maximum of two weeks. The results observed for the USA indicate the worst performance in the group, the model that obtained the best result allowed the analysis of 15 days to predict the next 30 days. Figure 14 shows the validation results obtained with the best performing models for the countries of the critical group on a linear scale.

Table III. Best and worst validation results of LSTM models for daily COVID cases prediction.

Country	Best Validation Score [RMSE]	Input and Output days for Best Score	Worst Validation Score[RMSE]	Input and Output days for Worst Score
Critical Group				
Brazil	0.0059	(15,20)	0.1356	(5,40)
India	0.0072	(15,20)	3.9783	(15,25)
South Africa	0.0124	(7,15)	5.0142	(15,35)
Russia	0.0170	(15,15)	0.0747	(5,30)
USA	0.0231	(15,30)	0.6030	(15,25)
Impact Group				
China	0.0062	(5,15)	0.0930	(7,40)
Italy	0.0145	(10,15)	0.2186	(7,40)
Australia	0.0154	(15,20)	0.2236	(5,40)
Switzerland	0.0134	(15,15)	0.2208	(5,30)
Portugal	0.0250	(10,15)	0.0778	(5,40)
Successful Group				
New Zealand	0.0105	(7,15)	0.3098	(10,30)
South Korea	0.0121	(10,20)	0.1183	(5,15)

Description of results for the Impact and Attention Group: For this group, we point out the results obtained for China, because the models are more suitable to the chosen parameters of the LSTM models. As explained earlier, the disease behavior can be observed on average in 5 days, and the effectiveness of control actions can be observed on average in 15 days with respect to the initial identification of the pandemic. The parameterization with the criteria explained in subsection "Parameterization of LSTM Models Based on Cause and Effect Criteria", allowed the development of a model with good performance in validating the number of daily cases, enabling the possibility of replicating these results for specific locations, provided that the environmental variables are understood and modeled. Also, the obtained result for China indicates that the efficiency of severe control actions strongly influences the containment of disease spread. For other countries such as Italy, Australia, Switzerland and Portugal, the results were similar, after a minimum of 10 days as input from the models to a 15-day average forecast. Figure 15 shows the validation results obtained with the best performing models for the countries of Impact and Attention Group on a linear scale.

Description of Results for the Successful Group: In this group, we point out the similarities between the countries and the results obtained regarding the validation of the RMSE score. Results for this group indicate that the model could identify the peak of the disease and the progressive reduction in the number of cases. Both South Korea and New Zealand have developed severe control actions to reduce new infections in the population. Thus, for every 10 days observed on average, prediction with

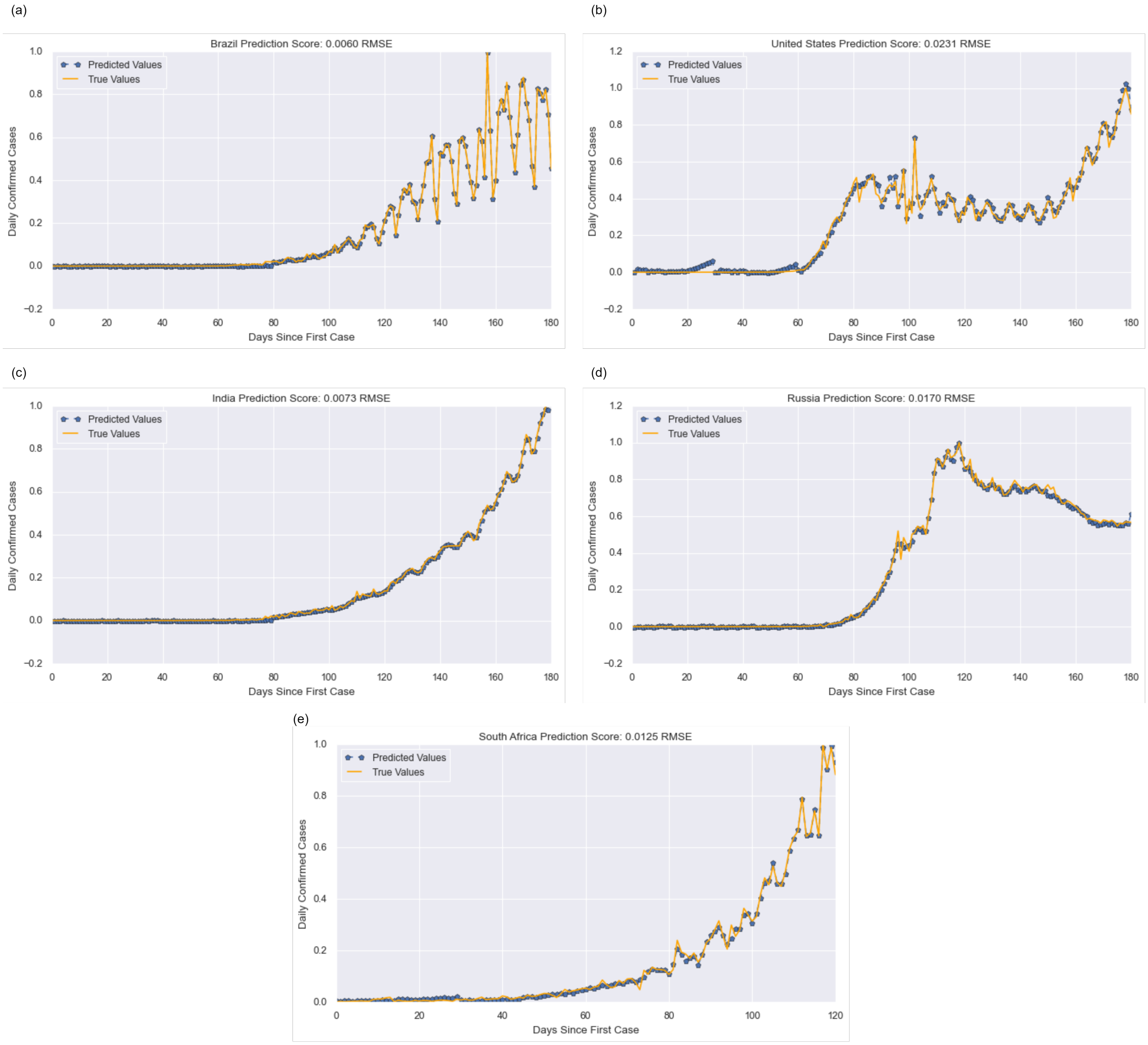


Figure 14. Validation results of LSTM models for daily COVID-19 cases predictions for countries (Brazil (a), USA (b), India (c), Russia (d) and South Africa (e)) of Critical group.

good model performance occurs between 15 and 20 days. Figure 16 illustrates the validation results obtained with the best performing models for countries in the successful group on a linear scale.

Observing these results indicates that the chosen LSTM models for daily case prediction are adequate when compared to the true values of the time series, as illustrated in figures (Figure 14, Figure 15, and Figure 16).

It is important to understand that each country presents a different behavior so that to obtain a better result for the validation of the models, a specific and exact parameterization of events in each country is necessary. Identifying the incubation period of the virus, the transmission rate, the time a person seeks medical care, and testing rate are all important tasks to modeling this kind of

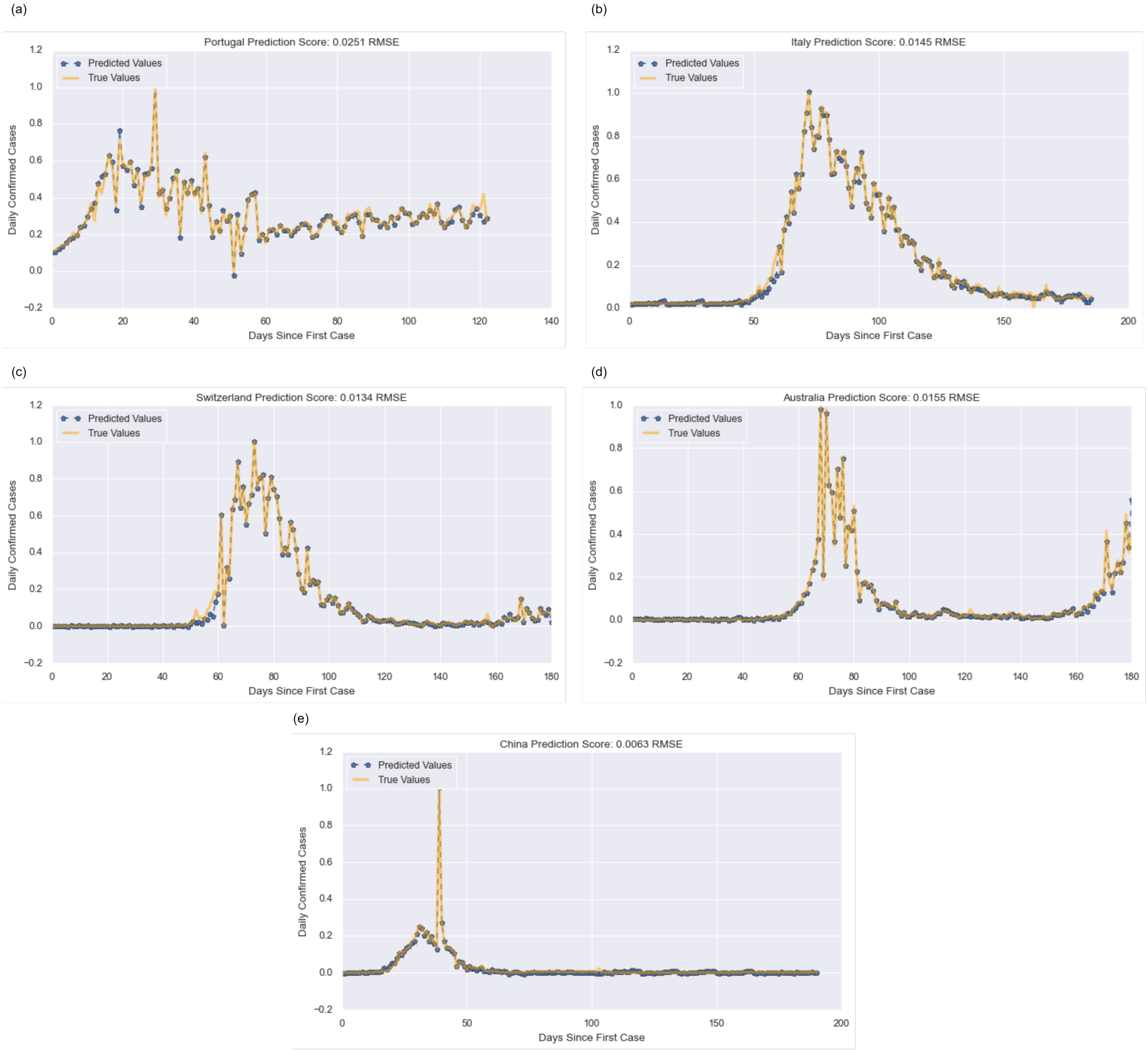


Figure 15. Validation results of LSTM models for daily COVID-19 cases predictions for countries (Portugal (a), Italy (b), Switzerland (c), Australia (d) and China (e)) of Impact group.

model, ensuring that the forecasting of daily COVID-19 cases came closer to the real behavior of the pandemic.

In order to report the performance of the models in forecasting new daily cases of COVID-19, we have selected the models that had the best performance concerning the RMSE index. The selected countries include Brazil, India, China, and New Zealand, thus covering all three control groups. The benchmark values (updated values) refer to cases collected after the development of the models, thus considering new information in the system. In this way, the obtained results of the respective models are presented. Figure 17 shows the predictions made for each country indicating the input days, the prediction itself, and the updated values of daily cases obtained from Our World in Data (Ritchie et al. 2020), updated on 13 August 2020.

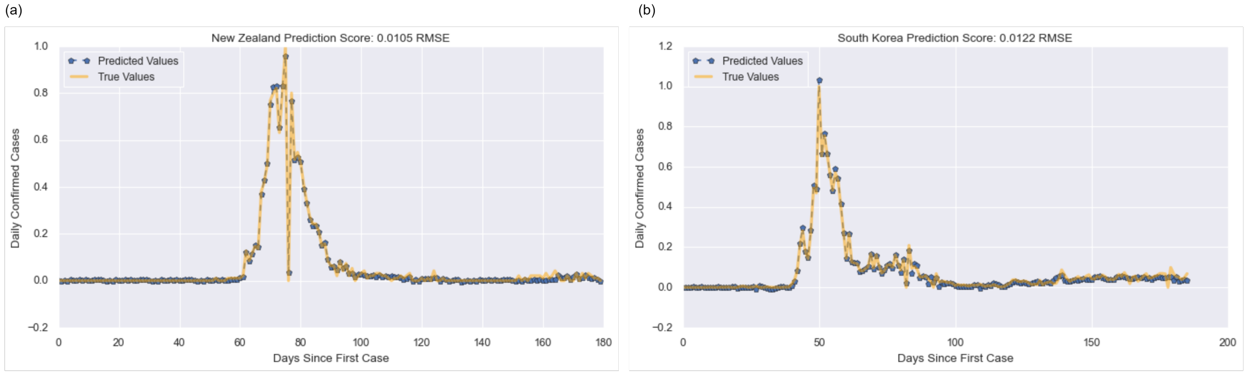


Figure 16. Validation results of LSTM models for daily COVID-19 cases predictions for countries (New Zealand (a) and South Korea (b)) of Successful group.

The further deployment of LSTM models for daily case prediction may contribute to the analysis of specific regions to understand the local evolution of infections. Easily interpreted and parameterized models allow several possibilities to develop and compare new prediction models.

In this paper, we seek to yield sufficient conditions to apply these models to regional data from cities, neighborhoods, villages, and regions to estimate accurately the epidemic behavior and new cases of COVID-19. The subsequent section indicates the application of this approach to several communities in Rio de Janeiro to observe model performance in predicting new daily cases of COVID-19.

LSTM modeling and forecasting results for favelas in Rio de Janeiro

By adopting the same setup made in the COVID-19 daily case forecast for countries, we have applied the methodology for regions and communities in Brazil. We have chosen a few regions, including the favelas of Rio de Janeiro. Applying the case prediction models for favelas becomes a critical instrument to help us to adopt actions or policies that can mitigate the effects of the pandemic in the communities. In Rio de Janeiro communities only, COVID-19 has killed about 850 people with 7385 cases of infection until August 2020. In Brazil, the COVID-19 epidemic has killed about 600,000 people until August 2021.

The daily case data comes from the local health units, and all the information is available on the website **Voz das Comunidades** Voz das Comunidades (2020), the data started to be updated on April 10, 2020. This data comes from the following sources: *Prefeitura do Rio de Janeiro, Governo Estadual do Rio de Janeiro, Clínica da Família Zilda Arns, Clínica da Família Pavão-Pavãozinho e Cantagalo, Centro de Saúde Escola Germano Sinval Faria - ENSP, Clínica da família Victor Valla, Clínica da Família Maria do Socorro Silva e Souza, Clínica da Família Valter Felisbino de Souza, Unidade de Saúde da Familia João Candido, Clínica da Família Anthídio Dias da Silveira, Clínica da Família Rinaldo De Lamare, Cms Dr Albert Sabin e Comitê SOS Providência.*

The obtained outcomes concern the LSTM model parameterization explained before. To simulate a closer scenario to reality, we have modeled the environmental parameters observed in the favelas to infer the number of input and output days (prediction).

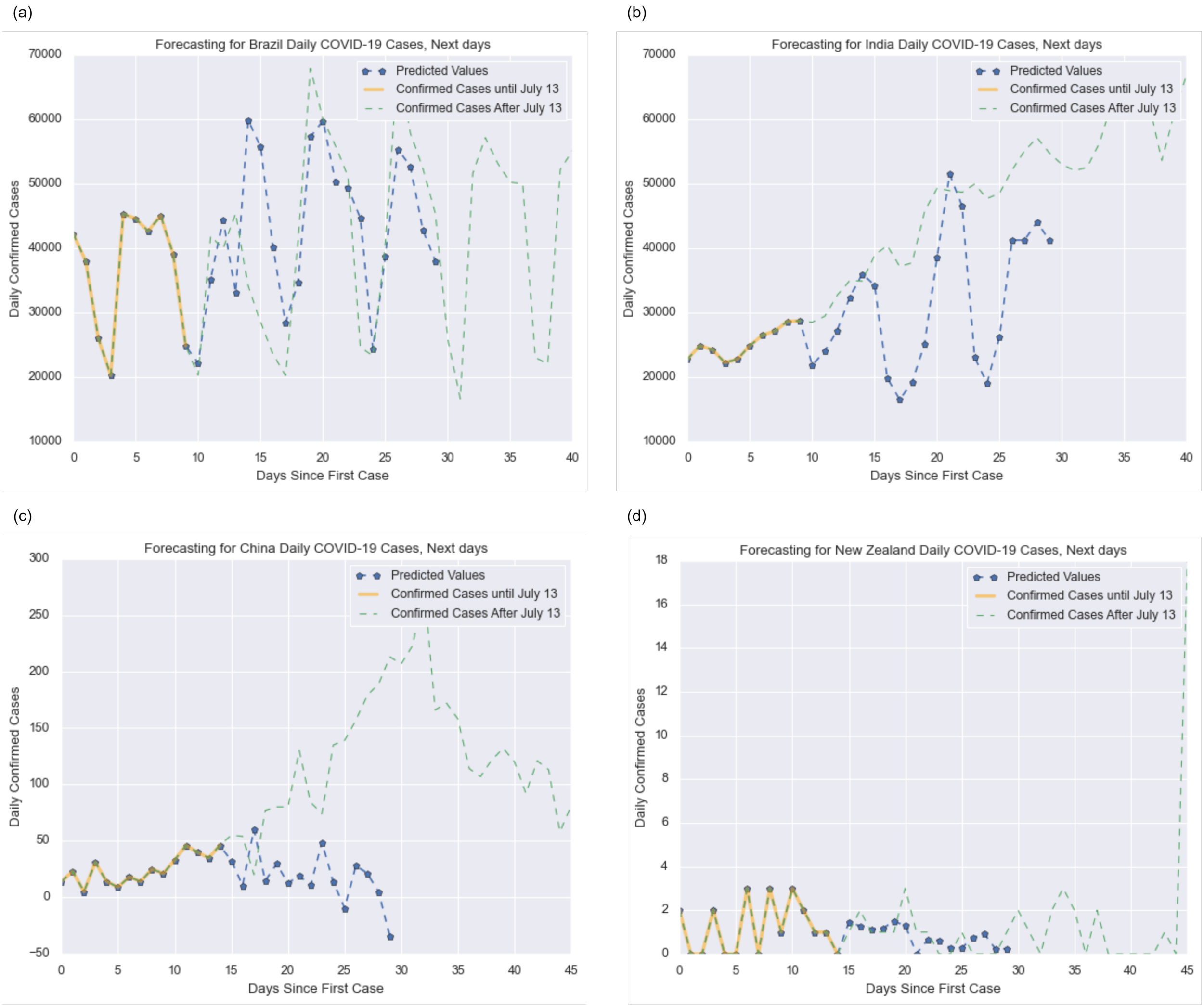


Figure 17. Forecasting Results for the Best LSTM Models. In (a) we show the forecasting results of the best LSTM model for Brazil daily COVID-19 cases. We show in (b) the forecasting results of the best LSTM model for India daily COVID-19 cases. For countries that have controlled the COVID-19 evolution, in (c) we show the forecasting China, and in (d) we show the results for New Zealand in both countries, the predictions show a progressive growth in the number of cases.

Favelas are regions with high population density and low-income inhabitants. This pandemic reveals a phenomenon in which social isolation measures, such as lockdown, are impractical due to the social structure established in these places (Anjos 2020). The effects of COVID-19 in Brazil go beyond infections and are related to social inequity associated with the lack of jobs and of conditions that ensure dignity, food, and health (Boehm 2020, Anjos 2020).

According to Anjos (2020), people who live in favelas and do not fit into essential services, cannot practice social isolation because they have an extreme need to work to ensure their survival and provide resources for their families. Figure 18 shows the evolution of daily cases in the Rio de Janeiro favelas according to the collected information.

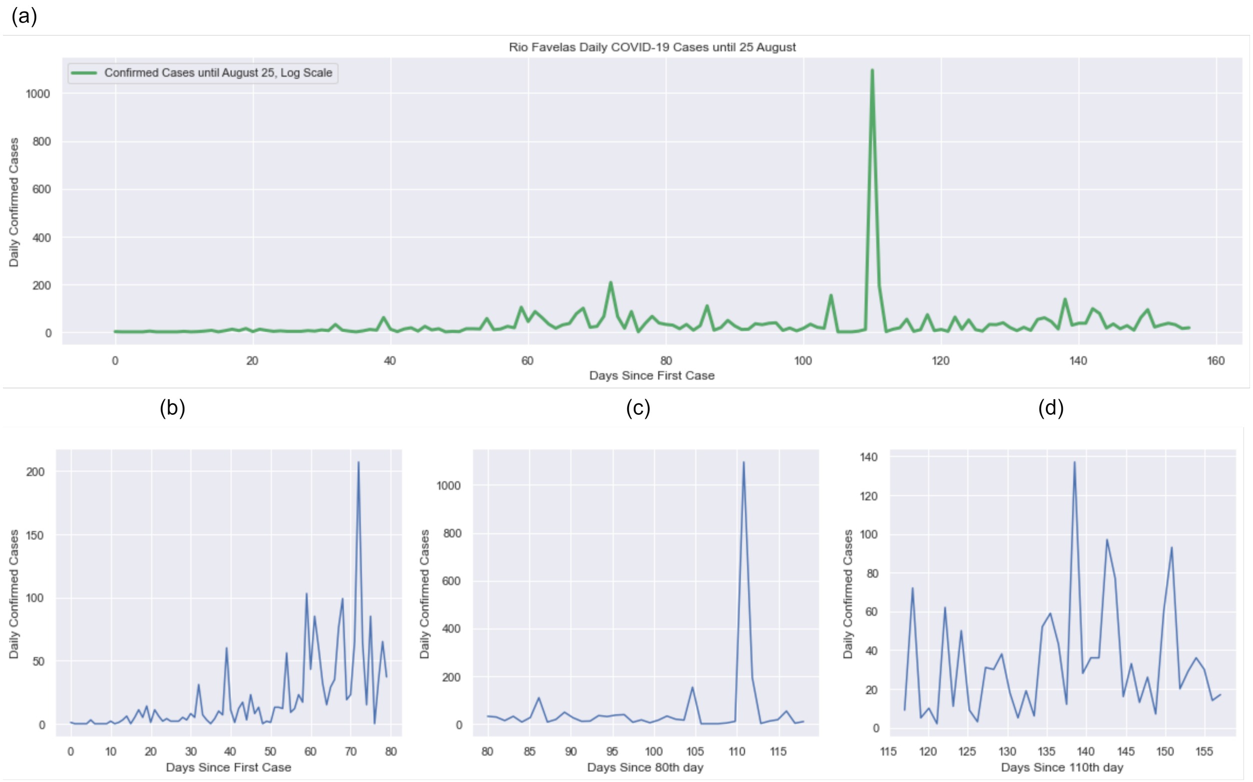


Figure 18. The different behaviors of COVID-19 evolution in Rio de Janeiro favelas, based on the time series of daily cases. (a) shows the entire series; (b) shows a window for the first 80 days since the first reported case, in this scenario it is possible to observe a progressive growth in the number of infections; (c) shows a window having an extreme peak of cases; and (d) shows a window that the time series presents similar behavior to the first 80 days, this shows that controlling the epidemic in the favelas is difficult, and this problem grows when there is an underreporting of cases and deaths.

According to Anjos (2020), 70% of favelas inhabitants had a significant reduction in their basic income due to the pandemic effects. Thus the need for survival surpassed any action to control the pandemic. According to (Boehm 2020) in Brazil 13.6 million people live in favelas, in a survey conducted by *Agencia Brasil* (Boehm 2020) these people do not feel confident about their safety or public health services.

Therefore, we adopt as input parameters a range from 5 days to 15 days and for the prediction a range from 5 days to 10 days. These parameters were adopted to allow a short prediction horizon so that the epidemic control actions can be made with more agility. Table IV indicates the results obtained in the prediction of daily cases of COVID-19 in Rio de Janeiro favelas.

The best results were observed by introducing seven days to predict the next five days. The results show that the epidemic has a faster evolution and observing these events in 7 days is enough to predict the cases in the favelas more accurately. Figure 19 shows the results of the model validation for predicting daily COVID-19 cases in Rio de Janeiro's favelas.

To evaluate the performance of the prediction, we established a data input counting up to August 9, 2020. Figure 20 shows the result of the prediction for the chosen model indicating the true values of daily COVID-19 infection cases, the predicted values, and the values updated until August 25, 2020.

Table IV. LSTM Models Results for Rio de Janeiro Favelas.

Number of Input Days	Number of Output Days	Number of Training Epochs	Prediction Score [RMSE] for Rio de Janeiro Communities
5	5	1000	0.0272
7	5	1000	**0.0152**
10	5	1000	0.0636
10	7	1000	0.0641
10	10	1000	0.1195
15	10	1000	0.1173

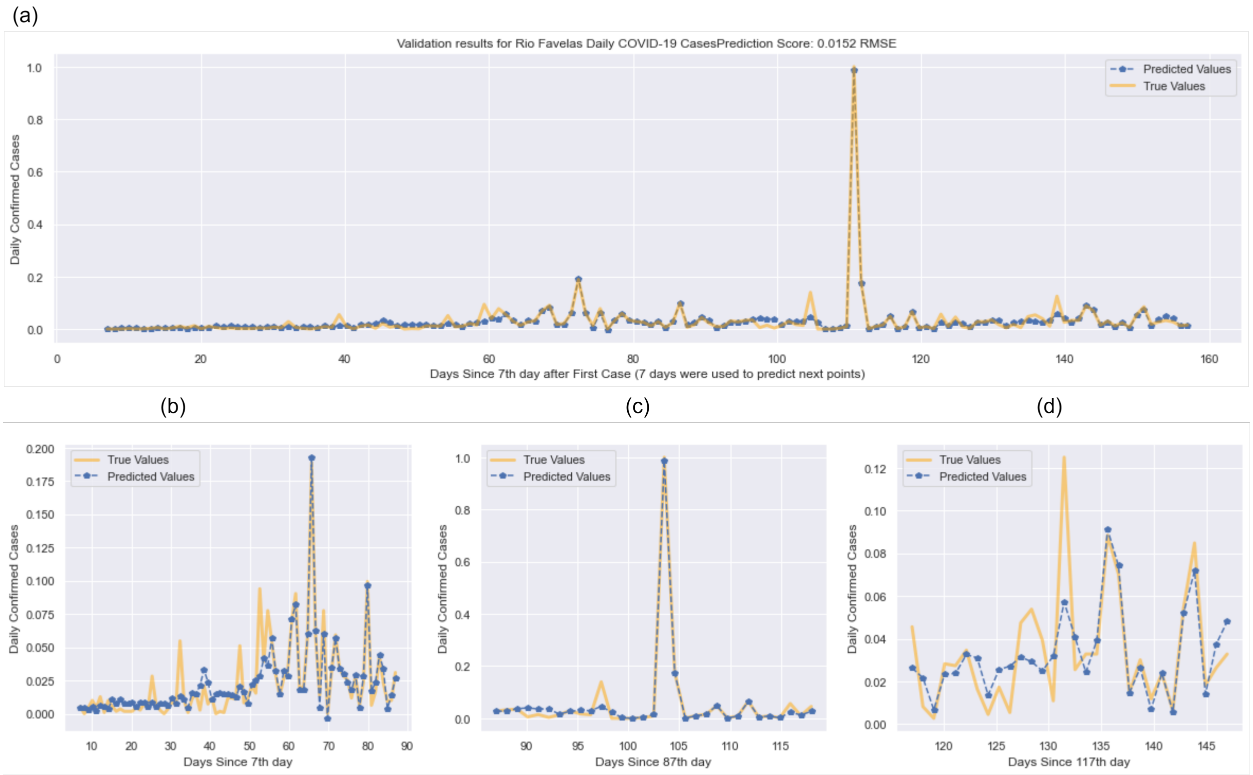


Figure 19. Validation results for Daily cases predictions in Rio de Janeiro favelas. The validation results show that the model can predict the different epidemic scenarios shown in (a,b,c, and d) in cities and regions.

Forecasting for Brazil With Treatment of Daily Cases Anomalies

The main objective of this work was to propose an approach capable of modeling the different dynamics presented by the COVID pandemic in some countries and also in particular areas such as poor communities. In this context, one of the aspects identified is that the data referring to this disease may present inconsistencies such as delays (impounded daily cases communication) or seasonality on weekends. In this way, we intend to present a scenario closer to reality in terms of the reaction to

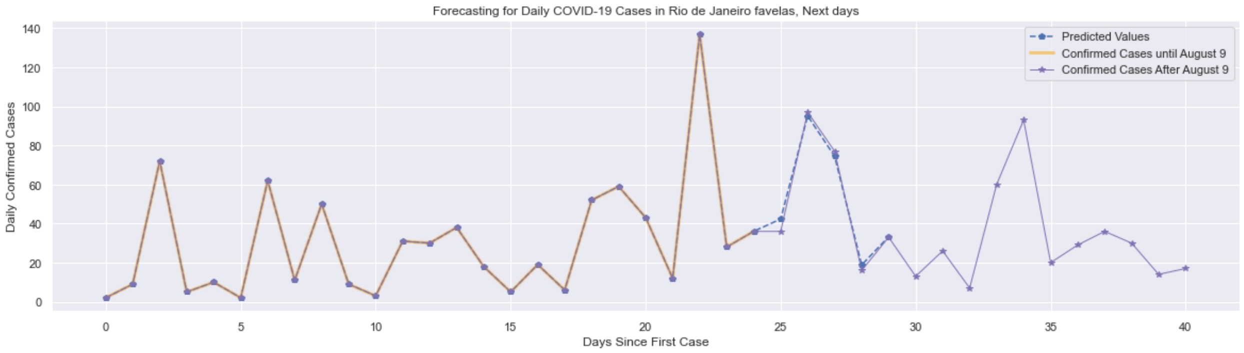


Figure 20. Forecasting Results for Daily COVID-19 cases in Rio favelas.

the situation of daily cases of COVID in Brazil using statistical methods for imputation of values, thus allowing a more reliable forecast.

Our data was handled with StructTS function from imputeTS package, this function employs the Kalman filter over a structural model (Figure 21), which enables considering trend and seasonal influences improving forecasting and even imputation results (Moritz et al. 2015). The methods based on the Kalman filter provide really good results for time series with a strong trend like our data of daily cases from COVID in Brazil similarly a common long-term observation for stock prices (Moritz et al. 2015, Moritz & Bartz-Beielstein 2017).

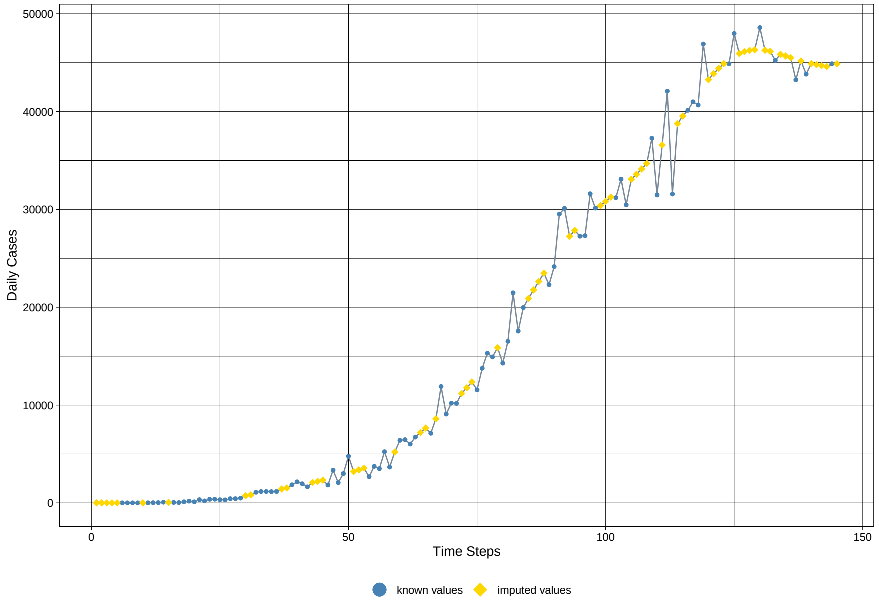


Figure 21. Result after statistical imputation for anomaly values identified at COVID daily cases from Brazil.

After statistical imputation, we use the adjusted LSTM model for forecasting daily cases in Brazil to compare with official data from the Brazilian Ministry of Health (Figure 22).

The results demonstrate that probably Brazil experienced a more dissemination of COVID than the official information presented, reaching a total of 3,572,551 at the end of July infected people against 2,992,600 official data. This situation can be even more serious since the data mentioned here

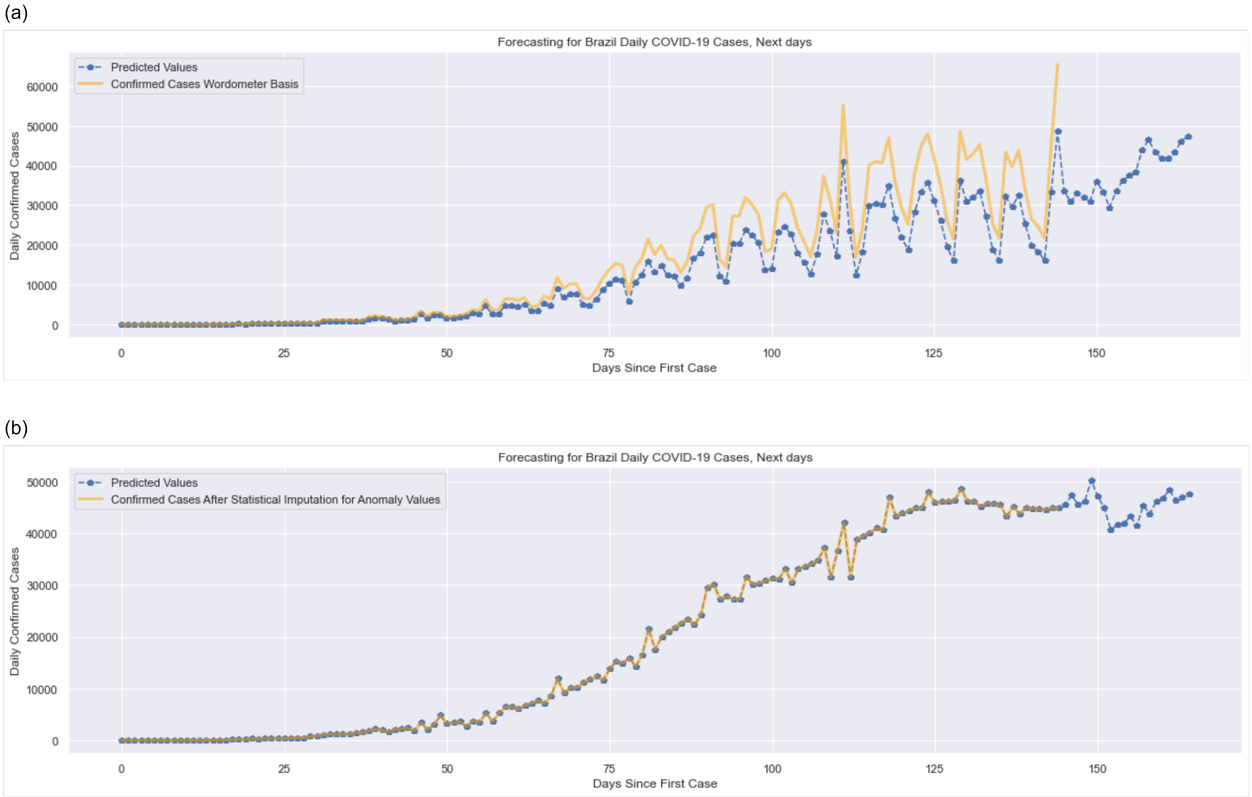


Figure 22. Prediction result for forecasting with treated data. In (a) we show the forecasting result for untreated data of daily COVID-19 cases in Brazil. In (b) we show the same prediction under-treated data using the statistical imputation method. Here it is observed that the prediction results obey the time series regime, since the unexpected seasonality was treated and removed from the time series. Thus, we observe that the application of this statistical imputation method allowed us to produce a more reliable prediction that obeys the natural evolution behavior of COVID-19.

report only the number of cases officially confirmed through exams, the problem is that Brazil tests proportionally much less than most countries in the world[10]. The underreporting of COVID-19 cases can be up to 7 times higher than the official numbers according Ribeiro & Bernardes (2020), further aggravating the situation of the population in a situation of vulnerability, especially in the needy communities of large cities (The Lancet 2020).

CONCLUSIONS

In this paper, we present an approach for applying RNN-LSTM models to predict COVID-19 in specific scenarios. We also incorporate adaptive linear regression analysis to detect anomalies and statistical imputation, with the aim of making the data more compatible with the complexity of each scenario. The forecasting models were investigated based on a set of criteria inherent to the stochastic dynamics of the pandemic (such as the behavior of society and government authorities in facing the crisis in

[10] Information about tests of COVID-19 by countries available at https://www.worldometers.info/coronavirus/\#countries. Accessed on June 12, 2020.

each scenario). Approximately 288 different prediction models were evaluated. We validate models in 12 countries, including Brazil and Favelas in Rio de Janeiro. We discussed the importance of this type of approach to help authorities with pandemic control actions, to more accurately and quickly predict new cases of COVID-19.

An important finding is that prediction models based on RNN-LSTM require careful parameterization in this context, avoiding random choice of parameters. Therefore, we have developed models that incorporate data quality assessment in each specific environment where the prediction of the daily number of COVID-19 cases is monitored (different countries and specific regions).

Eventually, we achieved the best forecasting results for the countries which showed the closest similarity to the established criteria of Cause vs. Effect. Regarding the prediction results obtained for Brazil, they also point to an evolution of the observed disease until August 2020, as can be seen by the predicted values and the updated values of the time series. Another important point is the application of these models to specific regions, as is the case of the Favelas of Rio de Janeiro. By developing the models, we have been able to verify the real situation of the favelas, providing support so that the competent authorities can act to control the epidemic. Our study also supports the prediction of cases of COVID-19 infection in small countries and regions with many social difficulties, poverty, and social conflicts in which vaccine distribution and humanitarian aid are extremely difficult.

We also emphasize that the parameterization and the models developed in this work can be applied to several problems that go beyond epidemiological situations and can also be fed by different sources of information, that is, two or more time series. Therefore, improvements to the methodology presented here are underway, also considering an update of the data for the different investigated scenarios. Given the stochastic nature of the daily cases analyzed, complementary techniques (e.g. *Spectral Fluctuation Analysis* (Veronese et al. 2011) are being incorporated to try to more robustly characterize the different scenarios. New results, including the high suppression rates achieved due to mass vaccination, which started in 2021, will be published further.

Although data collection and validation problems are common in the most diverse areas, this type of limitation and low quality of the information in health-related areas is extremely serious, as it directly impacts the safety and quality of life of each citizen. In the case of COVID data, incorrect reporting of the actual number of cases leads to a false impression that the disease cannot actually harm health in general. This denial view, partially imposed by low-quality data analysis, contributes to the growth of fatal cases as well as economic and geopolitical crises. Thus, our main purpose in analyzing the anomalous behavior of officially reported data was to shed light on a situation that potentially masked the true impact of the disease in the world, especially in Brazil, which has already been confirmed by other studies.

We provide all supplementary material on Github repositories. These repositories contain source code to predict daily cases of COVID-19 using LSTM network models. Basically, the package has a main module (core) and sub-modules with functions to tuning models and predict daily cases using time series data. The datasets are available at https://github.com/marcosmlr/docs_lstm_covid/blob/master/DatasetsCOVID.ipynb and approaches at https://github.com/marcosmlr/lstm_covid.

Finally, it is important to highlight that the new strains of COVID (eg Delta), vaccination programs and the emergence of third waves in each scenario still pose major challenges for COVID prediction based on mathematical and computational epidemiological models, given the nature of such factors are not yet fully known and incorporated into existing models. Highlighting scenarios in which the disease is expanding rapidly and its vaccination campaigns have barely begun. Such countries or regions often have outdated information and severe underreporting. Therefore, further studies are needed within a more specialist scope such as parameterization and the models developed in this work. This perspective must be considered within a scope that goes beyond epidemiological situations. In this context, the ecosystem that feeds the database to face the pandemic and its effects can and should improve substantially.

Acknowledgments

The authors thank CNPq, CAPES, ITA/IEAv-CTA and INPE-MCTI for the research support. This study was financed in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant 140473/2018-1, and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

ANIRUDH A. 2020. Mathematical modeling and the transmission dynamics in predicting the Covid-19-What next in combating the pandemic. Infectious Disease Modelling 5, 366–374. doi:10.1016/J.IDM. 2020.06.002

ANJOS AJ. 2020. A dificuldade do isolamento social nas favelas de Pernambuco (accessed 20 July 2020). Available at https://www.anf.org.br/o-isolamento-social-nas-favelas-de-pernambuco. Agência de Notícias das Favelas, Rio de Janeiro, Brasil.

ARORA P, KUMAR H & PANIGRAHI BK. 2020. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos Soliton Fract 139: 110017. doi:10.1016/j.chaos.2020.110017.

BOEHM C. 2020. Moradores de favelas movimentam R$ 119,8 bilhões por ano (accessed 20 July 2020). Available at https://agenciabrasil.ebc.com.br/geral/noticia/2020-01/moradores-de-favelas-movimentam-r-1198-bilhoes-por-ano. Agência Brasil, São Paulo, Brasil.

BUDUMA N & LOCASCIO N. 2017. Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. O'Reilly Media, Inc.

CHEN X & YU B. 2020. First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: real-time surveillance and evaluation with a second derivative model. Global Health Res Pol 5(1): 7. doi:10. 1186/s41256-020-00137-4.

CHIMMULA VKR & ZHANG L. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Soliton Fract 135. doi:10.1016/j.chaos.2020.109864.

CHOLLET F. 2021. Deep learning with Python. Simon and Schuster.

DENG L ET AL. 2014. Deep learning: methods and applications. Found Trends Signal Process 7(3-4): 197-387. doi:10.1561/2000000039.

GOODFELLOW I, BENGIO Y & COURVILLE A. 2016. Deep learning. MIT press.

GULLI A & PAL S. 2017. Deep Learning with Keras. Packt Publishing Ltd.

HARVEY AC. 1990. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

HOCHREITER S & SCHMIDHUBER J. 1997a. Long short-term memory. Neur Comput 9(8): 1735-1780.

HOCHREITER S & SCHMIDHUBER J. 1997b. LSTM can solve hard long time lag problems. In: Advances in neural information processing systems, p. 473-479.

KıRBAŞ İ, SÖZEN A, TUNCER AD & KAZANCıOĞLU FŞ. 2020. Comperative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos, Solitons & Fractals 138. doi:10. 1016/j.chaos.2020.110015.

LALMUANAWMA S, HUSSAIN J & CHHAKCHHUAK L. 2020. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals 139. doi:10.1016/j.chaos.2020.110059.

LAUER SA, GRANTZ KH, BI Q, JONES FK, ZHENG Q, MEREDITH HR, AZMAN AS, REICH NG & LESSLER J. 2020. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med 172(9): 577-582. doi:10.7326/ M20-0504.

LEYS C, LEY C, KLEIN O, BERNARD P & LICATA L. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J

Exp Soc Psychol 49(4): 764-766. doi:10.1016/j.jesp.2013.03.013.

MANASWI NK, MANASWI NK & JOHN S. 2018. Deep Learning with Applications Using Python. Springer.

MORITZ S & BARTZ-BEIELSTEIN T. 2017. imputeTS: time series missing value imputation in R. R Journal 9(1): 12. doi:10.32614/RJ-2017-009.

MORITZ S, SARDÁ A, BARTZ-BEIELSTEIN T, ZAEFFERER M & STORK J. 2015. Comparison of different methods for univariate time series imputation in R. arXiv preprint arXiv:151003924.

NIST/SEMATECH. 2013. e-Handbook of Statistical Methods. National Institute of Standards and Technology (NIST) & Semiconductor Manufacturing Technology (SEMATECH) consortium. doi:10.18434/M32189.

PEREIRA IG, GUERIN JM, JUNIOR AGS, DISTANTE C, GARCIA GS & GONCALVES LM. 2020. Forecasting Covid-19 dynamics in Brazil: a data driven approach. arXiv preprint arXiv:200509475.

RIBEIRO LC & BERNARDES AT. 2020. Estimate of underreporting of COVID-19 in Brazil by Acute Respiratory Syndrome hospitalization reports (accessed 15 May 2020). Notas Técnicas Cedeplar-UFMG 010. Available at https://ideas.repec.org/p/cdp/tecnot/tn010.html, Cedeplar, Universidade Federal de Minas Gerais, Minas Gerais, Brasil.

RITCHIE H, MATHIEU E, RODÉS-GUIRAO L, APPEL C, GIATTINO C, ORTIZ-OSPINA E ET AL. 2020. Coronavirus Pandemic (COVID-19) (accessed 10 january 2020). Available at https://ourworldindata.org/coronavirus. Our World in Data, Oxfordshire, United Kingdom.

SAK H, SENIOR A & BEAUFAYS F. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: XV Annual Conference of the International Speech Communication Association (INTERSPEECH), p. 338-342. doi:10.21437/Interspeech.2014-80.

SHARMA K, KOIRALA A, NICOLOPOULOS K, CHIU C, WOOD N & BRITTON PN. 2021. Vaccines for COVID-19: where do we stand in 2021? Paediatr Respir Rev 39: 22-31. doi:10.1016/j.prrv.2021.07.001

SORJAMAA A. 2010. Methodologies for time series prediction and missing value imputation. Dissertation for the degree of doctor of science in technology. Faculty of Information and Natural Sciences, the Aalto University School of Science and Technology. Espoo, Finland. http://lib.tkk.fi/Diss/2010/isbn9789526034539/isbn9789526034539.pdf.

TAYLOR SJ & LETHAM B. 2017. Forecasting at Scale. PeerJ Preprints 5:e3190v2 35(8): 48-90. doi:10.7287/peerj.preprints.3190v2.

THE LANCET. 2020. COVID-19 in Brazil: “So what?”. The Lancet 395(10235): 1461. doi:10.1016/S0140-6736(20)31095-3.

TIAN S ET AL. 2020. Characteristics of COVID-19 infection in Beijing. J Infect 80(4): 401-406. doi:10.1016/J.JINF.2020.02.018.

TOMAR A & GUPTA N. 2020. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ 1(728). doi:10.1016/j.scitotenv.2020.138762.

TOMAZ JV. 2020. COVID-19 nas Favelas (accessed 01 May 2020). Available at https://painel.vozdascomunidades.com.br. Voz das Comunidades, Rio de Janeiro, Brasil.

TSAY RS. 1988. Outliers, Level Shifts, and Variance Changes in Time Series. J Forecast 7(1): 1-20. doi:10.1002/for.3980070102.

VASILEV I ET AL. 2019. Python Deep Learning Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow. Packt Publishing Ltd.

VERONESE T, ROSA R, BOLZAN M, ROCHA FERNANDES F, SAWANT H & KARLICKY‘ M. 2011. Fluctuation analysis of solar radio bursts associated with geoeffective X-class flares. J Atmos Sol-Terr Phys 73(11): 1311-1316. doi:10.1016/j.jastp.2010.09.030.

WANG P, ZHENG X, LI J & ZHU B. 2020. Prediction of Epidemic Trends in COVID-19 with Logistic Model and Machine Learning Technics. Chaos Soliton Fract 139: 110058. doi:10.1016/j.chaos.2020.110058.

WILLMOTT CJ. 1982. Some comments on the evaluation of model performance. Bull Am Meteorol Soc 63(11): 1309-1313. doi:10.1175/1520-0477(1982)063(1309:SCOTEO)2.0.CO;2.

WORLDOMETER. 2020. COVID-19 Coronavirus Pandemic (accessed 10 January 2020). Available at https://www.worldometers.info/coronavirus. Worldometers.info, Dover, Delaware, USA.

YAN B, TANG X, LIU B, WANG J, ZHOU Y, ZHENG G, ZOU Q, LU Y & TU W. 2020. An Improved Method for the Fitting and Prediction of the Number of COVID-19 Confirmed Cases Based on LSTM. Comput Mater Cont 64(3): 1473-1490. doi:10.32604/cmc.2020.011317.

YUDISTIRA N. 2020. COVID-19 growth prediction using multivariate long short term memory. arXiv preprint arXiv:200504809.

ZHOU B, SHE J, WANG Y & MA X. 2020. Duration of Viral Shedding of Discharged Patients With Severe COVID-19. Clin Infect Dis 71(16): 2240-2242. doi:10.1093/cid/ciaa451.

ZIVOT E & WANG J. 2007. Modeling financial time series with S-Plus®. Vol. 191. Springer Science & Business Media.

How to cite

ARANTES FILHO LR, RODRIGUES ML, ROSA RR & GUIMARÃES LNF. 2022. Predicting COVID-19 cases in various scenarios using RNN-LSTM models aided by adaptive linear regression to identify data anomalies. An Acad Bras Cienc 94: e20210921. DOI 10.1590/0001-3765202220210921.

Manuscript received on June 23, 2021;
accepted for publication on January 12, 2022

LUIS RICARDO ARANTES FILHO[1]
http://orcid.org/0000-0002-8470-8377

MARCOS L. RODRIGUES[1]
https://orcid.org/0000-0002-9199-6928

REINALDO R. ROSA[1,2]
http://orcid.org/0000-0002-2962-4322

LAMARTINE N.F. GUIMARÃES[1,3]
http://orcid.org/0000-0002-0302-9162

[1]Instituto Nacional de Pesquisas Espaciais (INPE), Programa de Pós-Graduação em Computação Aplicada (PG-CAP), Av. dos Astronautas, 1758, Jd. da Granja, 12227-010 São José dos Campos, SP, Brazil
[2]Instituto Nacional de Pesquisas Espaciais (INPE), Laboratório Associado de Computação e Matemática Aplicada (LABAC), Av. dos Astronautas, 1758, Jd. da Granja, 12227-010 São José dos Campos, SP, Brazil
[3]Instituto Tecnológico de Aeronáutica (ITA), Programa de Pós-Graduação em Ciências e Tecnologias Espaciais (PG-CTE), Departamento de Ciência e Tecnologia Aeroespacial (DCTA), Praça Marechal Eduardo Gomes, 50, Vila das Acacias, 12228-900 São José dos Campos, SP, Brazil

Correspondence to: **Luis Ricardo Arantes Filho**

E-mail: luisricardoengcomp@gmail.com

Author contributions

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

APPENDIX - LSTM BASIC CONCEPTS

Recurrent Neural Networks (RNN) are Artificial Neural Networks (ANNs) applicable to classification and pattern recognition in sequential data and time series, such as voice, language, and textual data. Their data classification, clustering, and predictive capabilities come from the feature of allowing or not allowing the learning process in memory cells that are represented by connections among the neurons in the network (Gulli & Pal 2017, Manaswi et al. 2018, Vasilev et al. 2019).

RNN has a similar structure to a Multi-Layer Perceptron (MLP) (Goodfellow et al. 2016) network with the addition of loops and interconnections in the neurons of the intermediary layer. In this type of neural network, the learning of a given element depends on the previous elements in the data series. The figure A1 shows a simple RNN described with an architecture similar to the classic MLP network, however it is possible to check the interconnections and loops in the intermediate neurons. The recurrent layer has interconnections between neurons established in the same layers (Buduma & Locascio 2017).

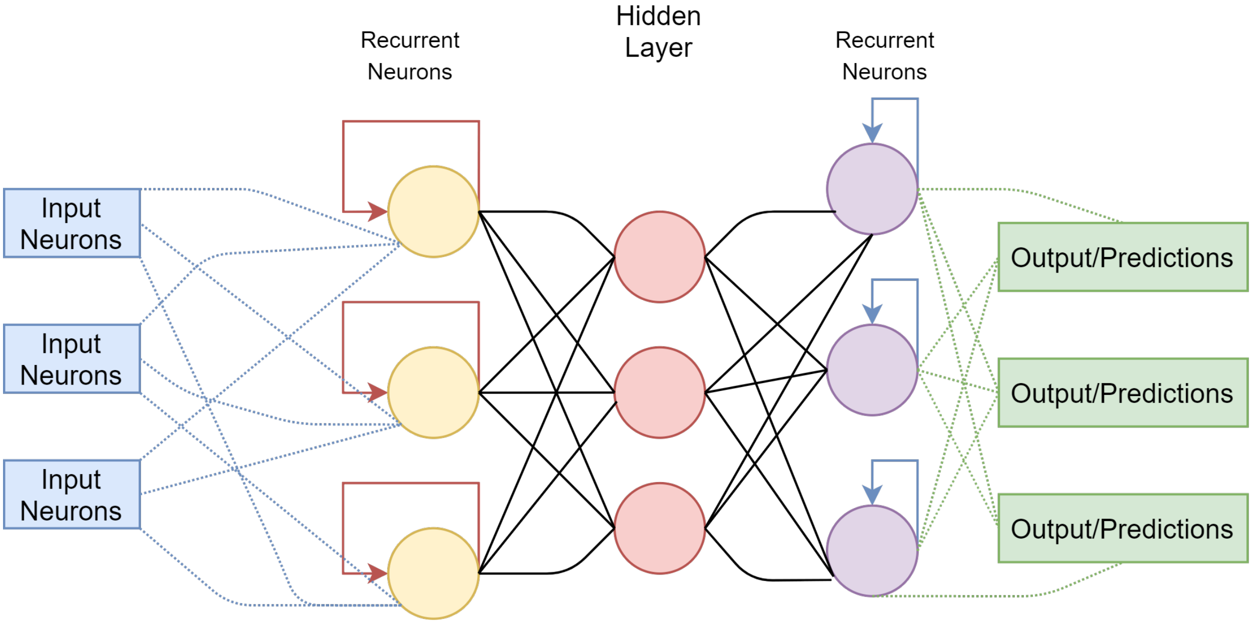


Figure A1. Basic Structure of a RNN.

This architecture ensures that output at time $\{t = n\}$ is dependent on inputs at time $\{t = n-1, t = n-2, ..., t = 1\}$. The training process is done by the backpropagation algorithm Goodfellow et al. (2016), if the data sequence is long there is the possibility of the function converging to local minimums or maximums, making learning slow and inefficient due to the simple architecture of the neural network, as pointed out by Manaswi et al. (2018).

Time series data applications require a more robust and improved type of topology that can take into account the input history. An RNN includes the ability to maintain internal memory with feedback and, therefore, support temporal behavior. The proposal of Hochreiter & Schmidhuber (1997b) called Long Short-Term Memory (LSTM) networks, improves the classic RNN networks. LSTM has been developed to deal with the problem of training convergence in large sequences of input data.

The LSTM networks try to establish relationships in data that have long-term temporal dependencies. This network includes memory cells in the intermediate layer (memory neurons) that hold information for long periods during training epochs. These cells are activated according to gates (*g*) (gate units) that control the information flux by activating each memory cell at a time, further details about LSTM Cell structure can be seen in Sak et al. (2014), Gulli & Pal (2017), Manaswi et al. (2018).

The recurrent connections add memory to the network and allow it to learn and take advantage of the ordered nature of the time series. The internal memory indicates that the output (prediction) of the network is dependent on the recent context in the input sequence and not the data that has just been selected as an input to the network (Manaswi et al. 2018).

There are several models and representations of LSTM networks to produce a sequence of predictions by a sequence of inputs, such of them are described bellow:

1. Model one to one: One input point generates only one output (prediction);
2. Model one to many: One input point generates N output points(predictions);
3. Model many to one: N input points generates only one output point (prediction);
4. Model many to many: N input points generates a sequence of N output points (predictions).

Figure A2 shows some LSTM models with different input/output relationships.

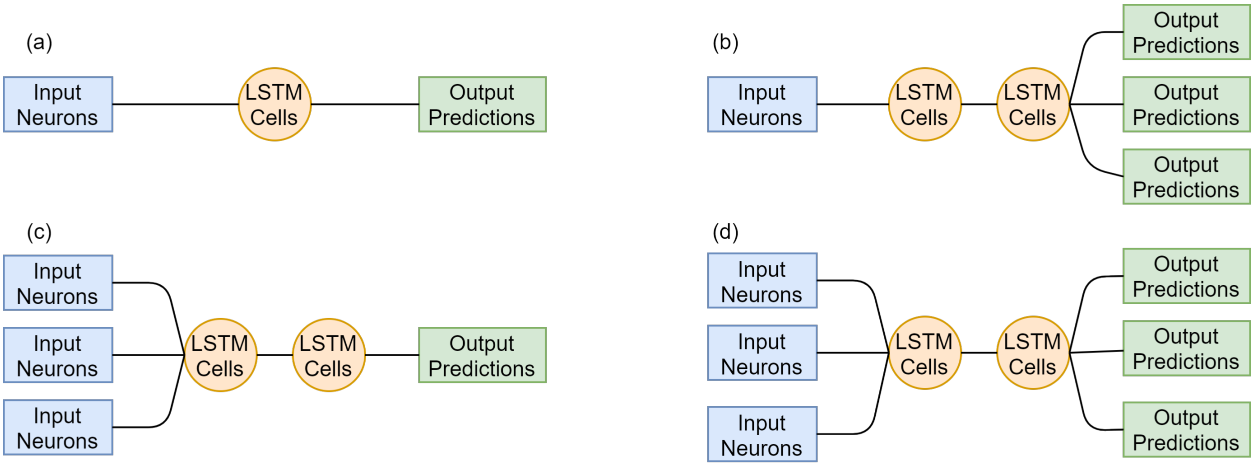


Figure A2. LSTM networks with sequences of inputs generating the outputs. In (a) we show a 1 to 1 model, in (b) we show the 1 to N model, in (c) we show the N to 1 model, and in (d) we show the N to N model. In this work, all the LSTM prediction models work as a N to N model.

In this paper, we try to develop an LSTM neural networks approach focusing mainly on the construction and parameterization of the models, checking how the models behave when time series data with several complex characteristics are applied. The models described at first subsections of Materials and Methods, featuring an N:N modeling (Figure A2), thus we seek to understand how the choice of inputs and outputs impact the predictive ability of the models. In this article, we also point out that, for the validation of the predictions, the data went through a deep statistical analysis, verifying in each series possible anomalies that affect the prediction capacity of the neural networks.

We point out that all models, as well as access to data and anomaly analysis models, are available online for access by the scientific community and society in general (https://github.com/marcosmlr/docs_lstm_covid/blob/master/DatasetsCOVID.ipynb, https://github.com/marcosmlr/lstm_covid). Each model can be trained and validated via the Google Colab platform without the need to purchase more sophisticated hardware or tools, this kind of feature allows these models to be applied in places that do not have powerful and sophisticated computing structures.