Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

1 Predictive models to the COVID-19

Francisco Nauber Bernardo Gois[1], Alex Lima[1], Khennedy Santos[1], Ramses Oliveira[1], Valdir Santiago[1], Saulo Melo[1], Rafael Costa[1], Marcelo Oliveira[1], Francisco das Chagas Douglas Marques Henrique[1], José Xavier Neto[1], Carlos Roberto Martins Rodrigues Sobrinho[1], João Alexandre Lôbo Marques[2]

[1]*SECRETARIA DA SAÚDE DO ESTADO DO CEARÁ FORTALEZA, CEARÁ, BRAZIL;* [2]*UNIVERSITY OF SAINT JOSEPH, MACAO, CHINA*

1. Introduction

The world has been facing threats in the form of pandemics periodically over the centuries. The current devastating pandemic is caused by the virus strain SARS-COV-2, which is causing the coronavirus disease 2019 (COVID-19). Because of that, some economies are crashing, and also the overall strengths and morals are heavily impacted worldwide. This pandemic already affected over 170 countries, and the numbers of infected and deceased patients are rising at an alarming rate. One key aspect to understand this pandemic starts with an understanding of the disease itself, and the progression of its natural course [1].

When new pathogen and their corresponding disease make more contagions, it is essential to establish the planning to manage the outbreak and determine their force. Forecasting techniques play a significant role in yielding accurate predictions assisting the Government in creating more reliable strategies and in making productive resolutions. Currently, event forecasting applications have become usual in society because of the significant evolution of robust computational models and hardware to process large volumes of data. These techniques use historical data, thereby enabling better predictions about the situation to occur in the future. These predictions may support governments from all over the world to be prepared for eventual forthcoming situations [1,2].

Understanding epidemic growth patterns across temporal and social factors can enhance our capacity to create epidemic transmission representations, including the critical job of predicting the estimated intensity of the outbreak morbidity or mortality impact at the end. Several studies consider the epidemic growth in a large population a stochastic event; the infection increases exponentially among subjects, each of by direct contact, closeness, or ambient traces [3]. Discover the rise kinetics of an epidemic can

Data Science for COVID-19. https://doi.org/10.1016/B978-0-12-824536-1.00023-X

help create well-grounded algorithms to predict and learn the essential features of the growth dynamics of infectious diseases. The force of the outbreak is represented in mathematical functions, modeling the transmission, and this is commonly estimated using time-series analysis describing the plague spread as a function of time [4].

Forecasting models identify patterns in those time-series and allow the analysis of epidemic predictions. A forecast model denotes an abstraction that simulates a system or object in certain details to facilitate the resolution of a problem. Mathematical models permit forecast and possible control of biological systems [3].

Several studies use different kinds of linear or nonlinear models to predict the spread of the epidemic. These models handle time-series data to deliver short-term and/or long-term predictions of an epidemic disease. Each computational forecasting model has its characteristics because each model can better fit a type of problem. Every prediction technique objects obtaining high accuracy in forecasting the tomorrow, so generalization settle great precision [5]. Several models, from rule-based scoring methods to machine learning and deep learning networks, have been suggested to answer the COVID-19 outbreak, generating several studies to support public strategies and help protect lives [6].

The present chapter focuses on the survey of epidemic forecasts intended to predict COVID-19 statistics, such as several infections and deaths, spread locations, and others. It also presents the forecasting solutions proposed by the IT team of the Secretariat of Health of the State of Ceara (SESA) and CISEC against the COVID-19 epidemic in Brazil. We organize the rest of this study as follows: Section two provides an introduction to the COVID-19 forecast models and methodologies. We describe the preliminaries of SEIR, SIR, Facebook Prophet, Kalman filters (KFs), and long short-term memory (LSTM) models used by SESA to COVID-19 in Section 3. Section 4 exhibits the adopted methodology, and Section 5 presents the experimental setup and preliminary results. We conclude this study in Section 6.

2. COVID-19 epidemic forecast

Forecasting is anticipating tomorrow using old and present information. The main class of forecasting is qualitative methods, explicative techniques, and time series models. Epidemic forecasting is the utilization of mathematical and machine learning methods to foretell the spread of epidemic diseases. Epidemic forecasting predicts epidemic size, maximum periods, and spread time. Forecasting an epidemic curve includes the use of statistics, immunological, or geolocation data [5].

History of epidemiology forecasting arises from 1760. This year, Daniel Bernoulli concluded that vaccination could increase longevity in France. In 1854, John Snow studied a cholera disease in London. He connected it to a reserve of contaminated water. In the present age of social, mobility, analytics, and computing solutions, a substantial volume of information is acquire created from social communication platforms and real-time streams of outbreaks. This comprehensive data make the computation in

epidemiology increasingly complex. Big data computational epidemiology is a developing interdisciplinary area that makes use of computational models to understanding and measuring the spatiotemporal transference of infection [7].

Diverse methods and techniques have been created to examine epidemics dynamics, counting classification, dynamics, forecast, and control strategy optimization [5].

Several studies try to predict the evolution of the epidemic curve [1,4]. Classical compartmental transmission models extensively studied the increasing growth of plague spread. These models presume exponential expansion dynamics in the lack of control measures [4]. Four kinds of solution classify big data computational epidemiology [7]:

- Descriptive analytics: This includes features of the outbreak size, duration, and other properties of diseases.
- Predictive analytics: These consist of problems of determining quantities, such as identifying the people who might be infected, the number of infection cases over time, and the top of cases.
- Preventive analytics: Defined by the network, early conditions, and epidemic model.
- Prescriptive analytics: This consists of obstacles of controlling the outbreak of epidemics, e.g., by immunization or restrictive measures.

We searched Google Scholar, IEEE, Springer, and Elsevier for analysis on COVID-19 forecast distributed later than 3 January of 2020. We use search terms: COVID-19, forecast, prediction model, machine learning, artificial intelligence, algorithm, score, deep learning, regression. We recovered 1240 titles by our systematic search. Several kinds of research were based on the publicly available data of confirmed daily cases come for the Hubei and China [8–10]. Some studies use the data from the World Health Organization website and Johns Hopkins University [11,12].

COVID-19 prediction has been made based on various forecasting techniques and different data sources. To better understand the forecasting techniques, this section categorizes these techniques into multiple types for better analysis. To didactical reasons, we separated the machine learning models that do not work only with time series in a different category. We include the Math equations and additive/multiplicative models in time series categories. Fig. 1.1 presents the systematic review process. Fig. 1.2 presents the studies categorization.

2.1 Epidemic growth models

The infectious disease outbreaks prediction usually has models that adopt an exponential increase in the lacking of restriction measures. In the initial stages of a plague, healthy infected singular contacts occur probability independent. Because of this factor, the likelihood of many infected individuals encountering a unique healthy person is potentially low. It is accepted that, in the begin of an epidemic, all infected person infects R_0 people on average [3,13].

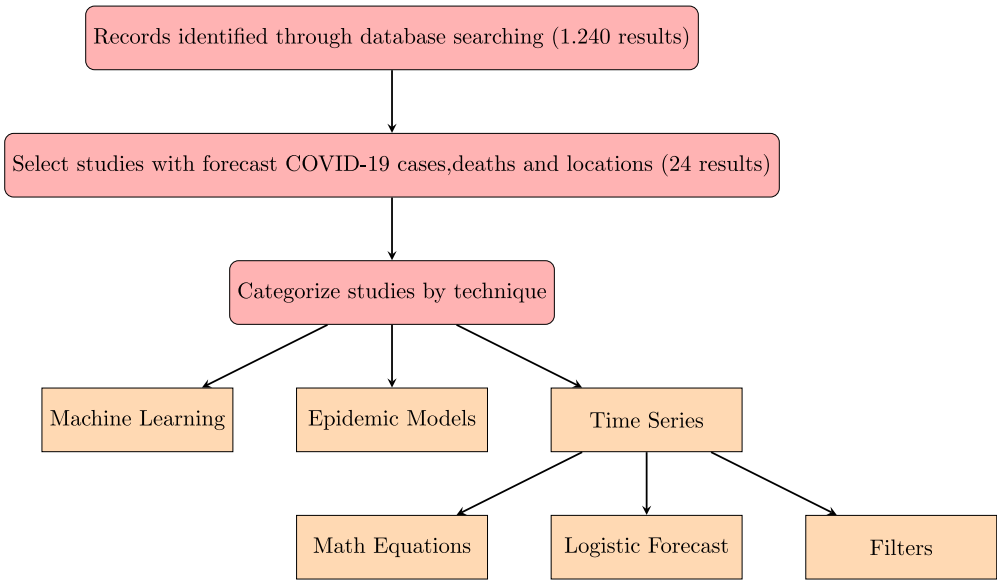


FIGURE 1.1 Systematic review process.

Henceforth, the expanding number of infections $N(t)$ increases ruled to the formula: $N(t) = N(0)e^{rt}$ where r is the infections at the beginning of the epidemic. If $R_0 > 1$, the plague increases exponentially. The basic reproduction number (R_0) is fundamental to forecast the infectious pathogen growth into a community. R_0 describes the subsequent cases that emerge from the initiation of an initial contagious case in a susceptive community through the epidemic time [3,13].

The commonly epidemiological models used are the SI, SIS, SIR, and SEIR representations. Into these models, each individual is separated into various divisions, and each division is in a state, each: Susceptible [S], Exposed or latent [E], Infectious [I] or Removed [R]. Yang et al. developed a dynamic SEIR model and AI model that can predict the COVID-19 epidemic trend within reasonable confidence. Yang et al. also use an LSTM model incorporating the results of the SEIR model, using epidemiological variables: the likelihood of contagious, incubation, and recovery rate [13].

Anastassopoulou et al. tried, with the available information, to determine the average values of the principal epidemiological variables: R_0, the case deaths ($\hat{g}$) and case healing ($\hat{b}$) ratios, with their 90% confidence intervals and customize the variables of the SIRD model to adjust the described data [8].

Jia et al. use the Bertalanffy model to explain the outbreak pattern of infectious and to describe the elements that handle and impact the outbreak of COVID-19 [14]. Teles adopted a SIR model applied in South Korea to foretelling the development of the active cases of the MERS epidemic in 2015 to predict COVID-19 cases [12]. ZHU et al. show a novel outbreak model called SEIR-HC. The study replicates the spread process of the

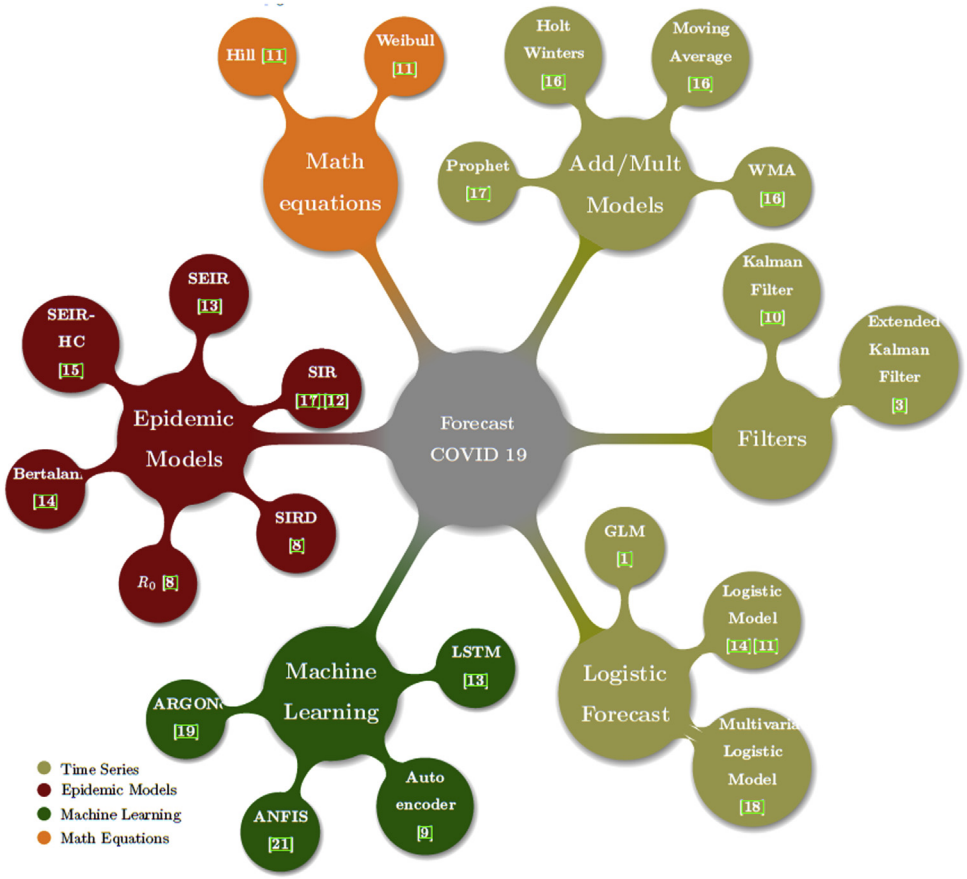


FIGURE 1.2 Research studies found with Forecast COVID 19.

COVID-19 epidemic in Wuhan city using the SEIR-HC model with an optimization algorithm, and then the propagation features and unknown data were estimated [15].

2.2 Time series

A time series is defined as a succession of features listed in time order [16]. Time series forecast models foretell the spread of diseases by analyzing one-dimensional data of infection cases, principally counting Autoregressive Integrated Moving Average (MA) model, Exponential Smoothing method Gray Model, and Markov chain method.

2.2.1 Logistic models, additive/multiplicative models, and math equations

The logistic model is a kind of time series model typically adopted in the study of epidemics. It is common to examine the threat circumstances of a particular illness and foretell the likelihood of occurrence of a particular pathology according to the risk factors [14].

Ndiaye et al. use Prophet, a solution for time-series predictions using an additive model [17]. Elmousalami and Hassanien use daily predictive models using a MA, weighted moving average (WMA), and single exponential smoothing (SES). A MA is analyzing the data points by averaging the series of data points. A MA depends on the acceptance of future observation is similar to recently previous observations. Similar to the MA, the WMA is a modification of the MA model by assigning weights to data points. SES is a smoothing time series data based on the exponential window function. Moreover, triple exponential smoothing (Holt–Winters method) is an algorithm used to forecast data points in a series [16].

Jagadish Kumar and Hembram use the Logistic equation, Weibull equation, and the Hill equation to find contagion rates in China and Italy. In this research work, data analysis is done to understand the effect of environmental factors on the spread of coronavirus disease. The cumulative infected data were examined based on several increase models. The recent data were fitted with the Gaussian distribution function [11]. Zhou et al. use univariable and multivariable logistic forecast models to investigate the threat circumstances connected with in-hospital fatality [18].

Yang et al. use three models that have been previously used in several epidemics, including SARS, Ebola, pandemic influenza, and dengue, to generate and verify short-term predictions of the cumulative number of COVID-19 reported cases in Hubei province. The study measures uncertainty based on a logistic growth model, the Richards growth model, and a sub-epidemic wave model [1]. The generalized logistic growth model increases the simple logistic growth model to adjust sub-exponential rise dynamics with a scaling of increase variables, p. Jia et al. use three varieties of numerical models: The logistic model, Bertalanffy model, and Gompertz model [14].

2.2.2 Nonlinear filter prediction models

A model is described of several numerical equations that are set to describe the interaction between various variables within specific methods. A model is not a perfect portrayal of reality. Commonly, we have no perfect understanding of the boundary conditions of the model and its uncertainty. We need to recognize the time progression of the probability density function (pdf) for the model state. With knowledge of the pdf for the model state, we can obtain knowledge about the model uncertainty. For time-based solutions, sequential data assimilation methods utilize the analysis scheme from the previous data to update the model state consecutively. Before-mentioned approaches have demonstrated helpful for several purposes, where new observations are sequentially absorbed into the model when they become ready.

Yang et al. use the ensemble KF as a short period predictor and test the success of nonpharmaceutical interventions on the epidemic spreading. The study builds an individual level–based network representation and performs stochastic reproductions to study the pestilences in Hubei Province at its initial stage and examine the plague dynamics under several situations [10]. Sameni uses an extended KF for joint parameters and variables for the estimates [3].

2.3 Machine learning prediction models

Machine learning techniques for forecasting is a part of artificial intelligence where algorithms learn from data. Machine learning models can include artificial neural networks (ANNs), deep learning, association rules, decision trees, reinforcement learning, and Bayesian networks [17].

Al-qaness et al. suggest using the adaptive neuro-fuzzy inference system (ANFIS) model that consolidates the features of both ANNs and fuzzy logic systems to anticipate COVID-19 positive cases [9]. Yang et al. use a LSTM model to predict the epidemic trend. The study used the 2003 SARS disease data, which were available for cases between April and June of 2003. The research developed a single network structure to prevent overfitting. The model was upgraded using Adam optimizer and worked for 500 iterations [13].

Liu et al. showed a methodology able to create significant and substantial short-term forecasts of COVID-19 activity, at the province level in China, by consolidating information from reports from China CDC, Internet search trends, news article trends, and information from mechanistic models. The study uses an augmented ARGONet machine learning model [19].

Rao and Vazquez use machine learning models with trip past along with the more common manifestations utilizing an online review. Before-mentioned collected data can be used in preceding screening and early identification of potential COVID-19 infected people. Thousands of data points can be received and treated by a machine learning framework that monitors people that could be contaminated and scale them into no-risk, minimal-risk, moderate-risk, and high-risk of being contaminated with the infection [20].

2.4 Discussion

The prediction representations firmly indicate that the curve of COVID-19 cases rises exponentially in nations that do not command limitations measures on travel, public gatherings, the closing of schools, universities, and workplaces. The exponential increase of cases strongly suggests that the outbreak's growth is due to an underlying biological phenomenon rather than the number of tests performed [16]. The substantial growth of the outbreak appears to be enormous even for the substantial effective Chinese logistics that make two new hospitals in a short time. Extensive capacities for this stage of health service in Hubei province or other parts in the World may prove particularly challenging [8].

But, in a limited group, the exponential rise in cases can not remain forever. Depending on the community dimension, the likelihood of infected people encountering healthy individuals drops. Therefore, the stochastic model of the outbreak spread saturates sometime [3].

Forecasting plays an essential role in every domain due to its benefits to save resources or to improve the economy. In the case of COVID-19, there are also many challenges for forecasting the death count and spread rate as the COVID-19 incubation period is very much longer, and significantly fewer datasets are available for the purpose [1].

There is a relationship in the growth kinetics of infected people, although the rate of infections is different due to various reasons. The infection curve of China and the Republic of Korea has almost reached its saturation value because of various reasons, for example, medical facilities, prevention, and public awareness. Furthermore, the distribution of daily infected people is well fitted with Gaussian function [11].

Hu et al. use a modified auto-encoder with multiple-step prediction, the model obtain an estimated average errors of 6−10 steps prediction of 1.64%, 2.27%, 2.14%, 2.08%, and 0.73%, respectively [9]. Yang et al. use an ensemble KF model to a short-term forecast of the COVID-19 curve in Wuhan City. The model can predict everyday cases and the plague hill. Identifying the daily cases from predictions 3 days ahead of the time supports proper supply provision. Yang et al. research conclusions show that decreasing the contagious time with control actions such as initial case identification and separation can decrease the plague dimension substantially [10]. Elmousalami et al. results indicate that SES is the most accurate model for forecasting confirmed, recovered, and death cases of COVID-19 [16].

Teles study used in Portugal explains that quarantine can be valuable in "flattening the curve." The study presents results that lowered the transmission rate to a fraction of the value from the initial representation used in Korea with restriction measures [12]. Yang et al. simulation results indicate that the Chinese government control epidemic using restriction measures. Unless remain and hardy control actions, the disease spread in Hubei Province would turn into continual growth, if the contagion rate is lowered by 25%, the epidemic would reach a top in the middle of February and fade out in late September. Using social distance in each city, the number of contagious cases would rise in the middle of February and decline to zero in the middle of June. With improved restrictive measures and social distancing control, the epidemic dynamics would rise at nearby mid-February and approximately the epidemic path in March. This fact can be crucial advice for nations going into the exponential increase of the outbreak in the present days [10].

Jia et al. results show that the Logistic model, Gompertz model, and Bertalanffy model has a superior prediction in the subsequent stagings of the outbreak. Between them, the Logistic model obtains good results for data in Wuhan, while Gompertz obtains more reliable results in predict the data in non-Hubei regions [14]. FPASSA-ANFIS model has a great potential to predict the number of confirmed cases within 10 days. Also, FPASSA-ANFIS surpasses other prediction models using RMSE, mean absolute error (MAE), MAPE, RMSRE, and $R\hat{A}^2 2$ validation methods [21]. Fanelli and Piazza's conclusions appear to indicate that there is a certain pattern in the growth curve of cases of COVID-19. Time-cases plots of the confirmed cases in communities of China, Italy, and France manifest the same pattern, which falls on the same pattern on average [3].

3. Material and methods

3.1 Epidemiologic predictors

When it comes to contagious diseases, it is frequent to use compartmental models, such as the SIR and SEIR models. Differential equations models SIR and SEIR, seeking the variations of the model parameters to project the spreading behavior of a given disease, are applied to the new coronavirus, where many works use these models [3,22].

3.1.1 SIR model

In 1921, Martinie created the Susceptible Infectious Removed model (SIR), which are spread in a human community by a vector; i.e., susceptive individuals acquire the infection from contagious vectors, and susceptive vectors acquire the disease from contagious people [15,23,24]. The SIR model, in principle, explains the process of a virus spread. On the other hand, this factor is not ever consonant with the contagious path. Some viruses do not confer any long-lasting immunization [15].

The SIR model is among the most fundamental compartmental representations, and several models are extended of this basic one, including the SEIR case. The SEIR model defines three partitions: S for the amount of susceptible, I for the number of infectious, and R for the number of recuperated or death (or immune) people [25].

The equations that describe the SIR model are described in 1, 2, and 3. All related to a unit of time, usually in days. Then at each instant of time t, the values of each compartment can be changed [23,25].

$$\frac{dS}{dt} = -\frac{\beta IS}{N}, \tag{1.1}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \tag{1.2}$$

$$\frac{dR}{dt} = \gamma I. \tag{1.3}$$

The modeling is simple, since $S(t) + I(t) + R(t) = N$ results in N, which represents the total population. Then in each t, individuals moved from S to I. The model removes the individuals infected with the disease from the compartment. Eq. (1.1) describes the model, where β is the average number of people comes into contact with another person multiplied by the likelihood of infection in that contact.

Eq. (1.1) shows use of the faction mentioned above removing the number of infected people, in the I compartment the new ones infected by the rate are added, with the removal of those who were recovered or died, introducing the term μ, which represents the recovery and mortality rate.

The last Eq. (1.7) explains the variation of the recovered patients and the number of deaths compartment, which is described by μ on those infected patients.

FIGURE 1.3 SIR model and the transitions between the compartments.

Fig. 1.3 illustrates all compartment transitions, showing the transition rate for each time in the arrows.

This model requires as input the amount of the susceptible, infected, and cured or dead population, all referring to time 0. And the necessary rates, it is transmission probability, recovery rate, and mortality (Fig. 1.4).

3.1.2 SEIR model

Because the SIS and SIR model exclusively supports the cases without an incubation period, which is not the case for several classes of contagious infections, Cooke proposed a spread model for the case that after a specific period, the susceptibles person can get infectious. This model is named as the SEIR model [26] (Fig. 1.5).

The SEIR model differs from the SIR in one compartment, the E representing Exposure, which refers to diseases that are not manifested at the exact moment of infection, having an incubation period. Like COVID-19, which has an ordinary incubation period of 14 days.

The model is defined with four differential equations, described in Eqs. (1.4–1.7). Some small changes are made, starting with the addition of the new Eq. (1.5), which represents the calculation of individuals exposed to the virus.

The model added a new rate, the incubation rate, σ, which is the rate of latent individuals becoming infectious (typical period of incubation is $1/\sigma$) [26].

$$\frac{dS}{dt} = -\frac{\beta IS}{N}, \tag{1.4}$$

$$\frac{dE}{dt} = \frac{\beta IS}{N} - \sigma E, \tag{1.5}$$

$$\frac{dI}{dt} = \sigma E - \gamma I, \tag{1.6}$$

$$\frac{dR}{dt} = \gamma I. \tag{1.7}$$

Analogous to the SIR representation, the sum of the compartments, which are now $S(t) + E(t) + I(t) + R(t) = N$, results in the total population.

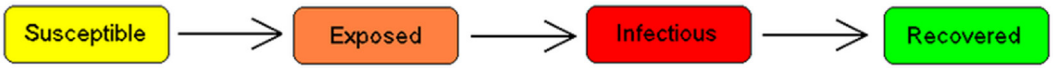


FIGURE 1.4 SEIR model with the transitions between the compartments [26].

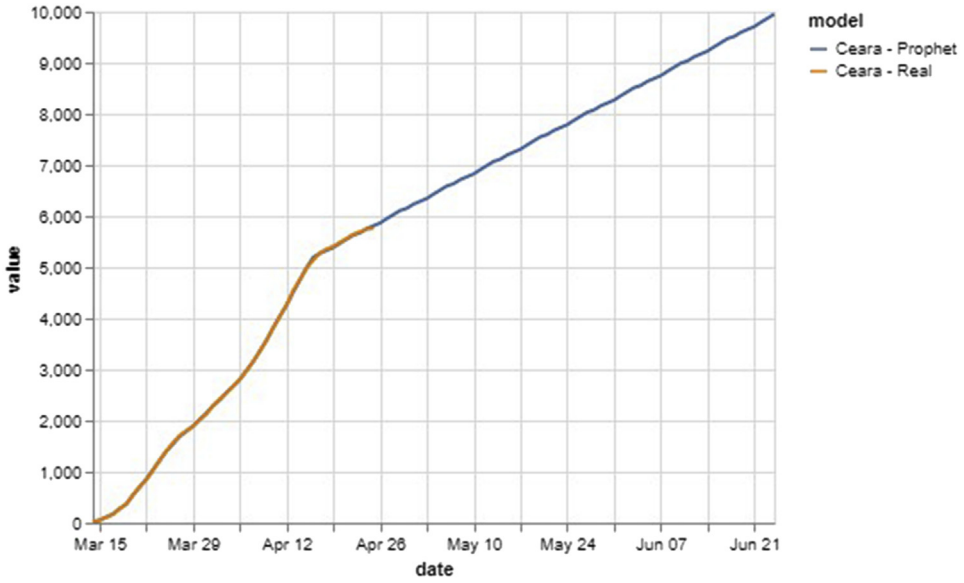


FIGURE 1.5 Prophet short term results to Ceará dataset.

3.2 Nonlinear additive and multiplicative methods

3.2.1 Prophet

Prophet is an approach for prediction of time series data based on an additive model. Prophet uses seasonality and day-off effects to calculate nonlinear tendencies. It operates appropriately with historical series that have regular periodical patterns and diverse seasons of past data. Prophet is resilient to missing data and variations in the bias and generally works well with outliers [27].

This method is a helpful method for time series with many distortions, lack of data, and drastic changes. What led us to use it since the lack of data on COVID-19 is excellent because it is a new disease.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{1.8}$$

The Prophet Eq. (1.8) shows the following features, decomposing the time series into three elements: trend $g(t)$, seasonality $s(t)$, and holidays $h(t)$.

- $g(t)$: piecewise linear or logistic increase curve for modeling nonseasonal changes in time series.
- $s(t)$: seasonal changes.
- $h(t)$: effects of day-off.
- ε_t: error term accounts for any not common changes not accommodated by the model.

3.3 Holt Winters

Exponential smoothing is an ordinary procedure used to predict a time series left out the requirement of applying a parametric model [28]. The Holt–Winters also named to as double exponential smoothing, is an addition of exponential smoothing created for trended and periodic time series.

The Holt–Winters model [29] is an expansion of the Holt method [30], developed by Winters and divided into two groups, multiplicative and additive Holt–Winters. The multiplier model was selected for the analysis in this chapter because it trends forecast values by seasonality, being the best for data with trends and increasing seasonality as a function of time.

The exponential and Holt–Winters procedures are susceptible to regular events or anomalies. Outliers influence prediction methods in two forms. First, the smoothed values are affected. Smoothed values depend on the present and historical values of the series, plus the outliers. The other influence concerns the choice of the parameters used in the recursive updating design [28].

The use of the multiplicative method is explained by the characteristics of the data, using the numbers of infections and deaths of COVID-19; the curve presents an exponential shape. The trend and seasonality data have an increase according to the number of days; thereby, the multiplicative model is ideal.

In the Holt–Winters multiplicative method, the periodic partition is formulated in relative terms and used to fit the time series periodically. Eqs. (1.9–1.11) describe the multiplicative method.

$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}). \tag{1.9}$$

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} \tag{1.10}$$

$$I_t = \beta \frac{y_t}{S_{t-1} + b_{t-1}} + (1-\beta)I_{t-L} \tag{1.11}$$

where S_t is the overall smoothing, b_t is the inclination smoothing, and I_t is the periodically smoothing. y_t refers to the real data at a period of t. L is the time. The α, γ, and β are constants between 0 and 1. The model minimizes the mean square error equation using α, γ, and β.

3.4 Kalman filter

The KF is a method that utilizes a set of measures observed over a period, including noise and gives estimations according to the used set, by considering a joint probability distribution across the variables for each time frame. The KF, also named as linear quadratic estimation, is an optimal estimator which suggests parameters of interest from indirect, inexact, and dubious observations.

The KF aims to find the "most reliable estimate" from noisy input. It is recursive, KF treats the new measures as they appear. The filter presents a recursive resolution to the linear optimal filtering problem to stationary as well as nonstationary situations. It is also recursive and measures the new state from the previous estimates and the new data. Unique the previous estimate needs storage, reducing the need for saving the whole past noted data [31]. Filtering methods allow the recursive evaluation of model parameters. These techniques have found application in various disciplines, and across the last two decades, have been used to contagious infection epidemiology [32].

The KF dynamics rise from the constant periods of forecast and filtering. The change aspects of these periods are determined and translated in Gaussian probability density functions. Following new constraints on the system changes, the KF dynamics converge to a steady-state filter, and the steady-state gain is inferred. The learning method connected with the filter, which describes the new data conveyed to the state measure by the latter system measure, is presented.

The KF gives a linear minimum error variance estimate of the state characterized by a state-space model. The KF has the support of leading with noise in the couple, model, and the data. The main goal of the KF is to diminish the mean squared error within the real and measured data. Consequently, it gives the accurate as a possible measure of the data in the mean squared error function. Thought from this fact, it should be plausible to determine that the KF has much in common with the chi-square. The chi-square merit function is typically applied as a model to fit a collection of model variables to a method named least-squares fitting. The KF is usually named as recursive least squares [33].

3.5 State space derivation

The differential equations of the KF can be incorporated into a state-space component. Let $Y_t, Y_{t-1}, \ldots, Y_1$ denoted the observed values of a feature in time t, $t-1$, …, 1. We assume that Y depends on an unobservable quantity θ, known as system state variables. The goal of KF is make inferences of θ. The relation between Y_t and θ is given by the equation [33,34]:

$$Y_t = F_t\,\theta_t + \upsilon_t \tag{1.12}$$

where F_t is a known quantity. F_t is the noiseless connection between the t state vector and the measurement vector, and is assumed stationary over time. The observation error υ_t is the associated with measurement error [34–36]. The main difference between KF and conventional linear models is that KF regression coefficients are not constant change over time as the system equation:

$$\theta_t = G_t\theta_{t-1} + w_t \tag{1.13}$$

where θ is the state vector at time t; G_t is the state transition matrix of the progress from the position at $t-1$ to the state at t, and is presumed stationary over time; w_t is the associated white noise with recognize covariance; υ_t and the system equation error w_t are presumed to be mutually independent random variables, spectrally white, and with

normal probability distributions. w_t and v_t are sequences of white, gaussian noise with zero mean:

$$E[w_t] = E[v_t] = 0, \tag{1.14}$$

The KF is the filter that gets the least mean-square state error estimation. When Y_0 is a Gaussian vector, the state and perceptions noises w_t and v_t are white and Gaussian, and the state and observation dynamics are linear. For the minimization of the MSE to support the optimal filter, it must be plausible to evaluate model errors using Gaussian distributions. The covariances of the noise models are considered stationary in period and are given by;

$$Q = E\left[w_t w_t^T\right] \tag{1.15}$$

$$R = E\left[v_t v_t^T\right] \tag{1.16}$$

The mean squared error is given by:

$$P_k = E\left[e_t e_t^T\right] = E\left[\left(Y_t - \widehat{Y}_t\right)\left(Y_t - \widehat{Y}_t\right)^T\right] \tag{1.17}$$

where P is the error covariance matrix at time t. Considering the previous estimation of $\widehat{Y}$ is named $\widehat{Y}'$, and was obtained by observation of the system. It is welcome to estimate using a write an update equation, mixing the old estimation with new measurement data.

4. Methodology

The proposed analysis considers public data available of new confirmed cases and deaths reported daily for the state of Ceará, in the northeast region of Brazil, from the 15 of March until the 24 of April. The data were obtained from an open API available on https://github.com/integrasus/api-covid-ce, validated according to the Ceara Integrasus Platform (available at https://indicadores.integrasus.saude.ce.gov.br/indicadores/indicadores-coronavirus/coronavirus-ceara). The database has the following attributes:

- Categorical result of COVID-19 exam
- City of patience provided by Brazilian Geographic Institute
- Asthma indicator
- Indicator of cardiovascular problems
- Date of death
- Date of exam result
- Date of begin of the symptoms
- Date of exam notification
- Exam final result

We planned three experiments. The first experiment aims to find the best model for short-term prediction. The model should only use the state of Ceará and find out which

are the best models. The second experiment aims to validate the selected models for the long-term prediction of the total number of active cases. Thus, we use China confirmed cases dataset from January to the April 27, 2020. The third experiment involves performing the long-term prediction of confirmed cases of COVID-19. In this path, data on cases of COVID-19 infection from China, Italy, Korea, and Brazil were used. The dataset has features of data and the number of infected in cumulative form.

4.1 Performance metrics

The accuracy of the suggested approach is evaluated by applying a set of performance metrics as follows:

4.1.1 Root mean square error

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(y'_i - y_i\right)^2} \tag{1.18}$$

where y' and y are the foretold and real values, respectively.

4.1.2 Mean absolute error

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left|y'_i - y_i\right| \tag{1.19}$$

where y' and y are the foretold and real values, sequentially.

4.1.3 Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y'_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y'_i - \overline{y}_i\right)^2} \tag{1.20}$$

where y' and y are the predicted and original values, respectively. $\overline{y}$ is the average of original values. The lowest value of RMSE and MAE indicates the most suitable approach. The greater rate of R^2 shows a better correlation for the method.

5. Results

The results are the most critical factors for the analysis of the pandemic, since it shows the possible epidemic evolution according to the proposed models. In this section, the results are presented for each model and a comparison between them is performed. The comparisons are based on standard metrics for regression models analysis, such as root mean square error, mean absolute error (MAE), and R squared (R^2).

To obtain more detailed and reliable results, three different environments for the results projection are proposed. The first one characterizes short-term forecasts, which projects values for the following weeks. For a long-term analysis, two environments are used, one using a data set from China, where the epidemic process has already passed

through all the steps, from its beginning to a presumed end. The last environment uses a data set that merges the data from Ceará so far with that one from China, using data from China to complete the data from Ceará until a possible end of the epidemic.

The results presented are based on data from the 15th of March of 2020, when there was a significant increase in the number of cases of COVID-19 at the State of Ceará, Brazil.

5.1 Short-term analysis

For the short-term analysis, we used only the data set from Ceará individually and Brazil as a whole. Table 1.1 presents the error results by RMSE, MAE, and R^2. Prophet and Holt Winter obtain better results than the other models used in the experiment; this fact can be explained due to the exponential nature of the epidemic curve. Epidemiological models had worse results than the time series and machine learning models. This fact does not necessarily indicate that the other models had worse results, but because of the nature of the COVID-19 disease, the models' parameters could be better adjusted.

The Prophet, a nonlinear method that adapts to seasonality, trend, and holidays of the time series, has been shown to correctly predict the number of COVID-19 cases in Ceará, making it the best method for short-term forecasting, the result shown in Fig. 1.6. Holt–Winters is a method applied to time series. We use the multiplicative method due to the growth in the curve in the data, generally an exponential shape. The method has excellent efficacy in series with high seasonality, which is not much presented in data from the epidemic in Ceará. But this does not rule out the method. We use it to predict the number of cases in Ceará in a short term; it is noteworthy to see that the method adjustment the data trend, with small variations at the beginning (Figs. 1.6 and 1.7).

Fig. 1.8 shows the prediction using the SEIR model in comparison with Ceara's real results. The SEIR model, as mentioned, is a mathematical method that makes predictions using differential equations. The values returned by SEIR project a continuous curve without more substantial distortions, and in most cases, have an exponential shape. To predict the number of cases with the SEIR model, we use the values within the I compartment, which refers to people who were infected. As the compartment is varied in t, the input data used were those obtained at the beginning of the epidemic in Ceará (Fig. 1.9).

Table 1.1 Method errors to short-term experiments.

Method	MAE	RMSE	R2
Prophet	11.49	16.06	0.999
Holt–Winters	133.90	149.08	0.994
KF + SEIR + CE	216.65	245.89	0.983
Kalman filter	342.83	388.52	0.959
SEIR	564.79	723.29	0.858
KF + SEIR	517.85	758.68	0.844

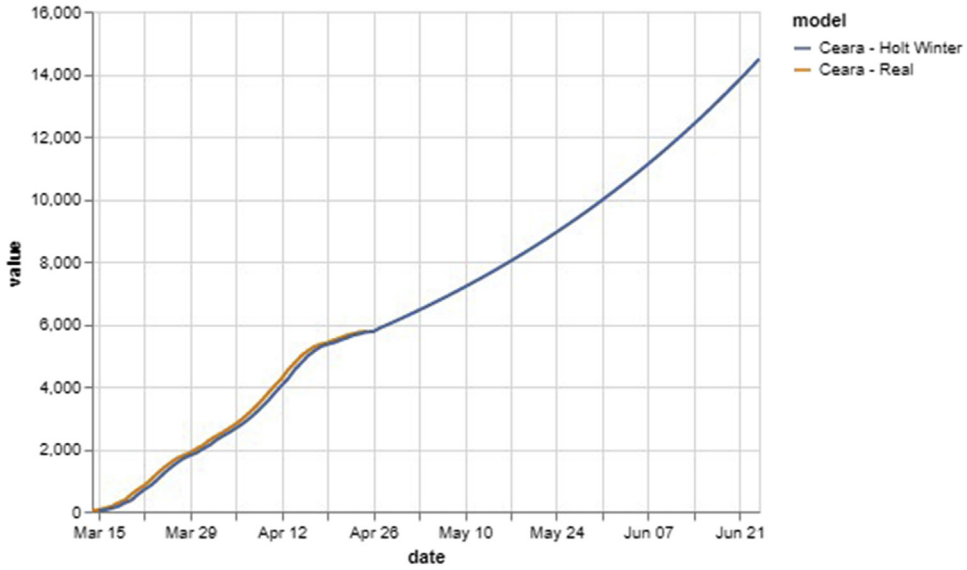


FIGURE 1.6 Holt Winters short term results to Ceará dataset.

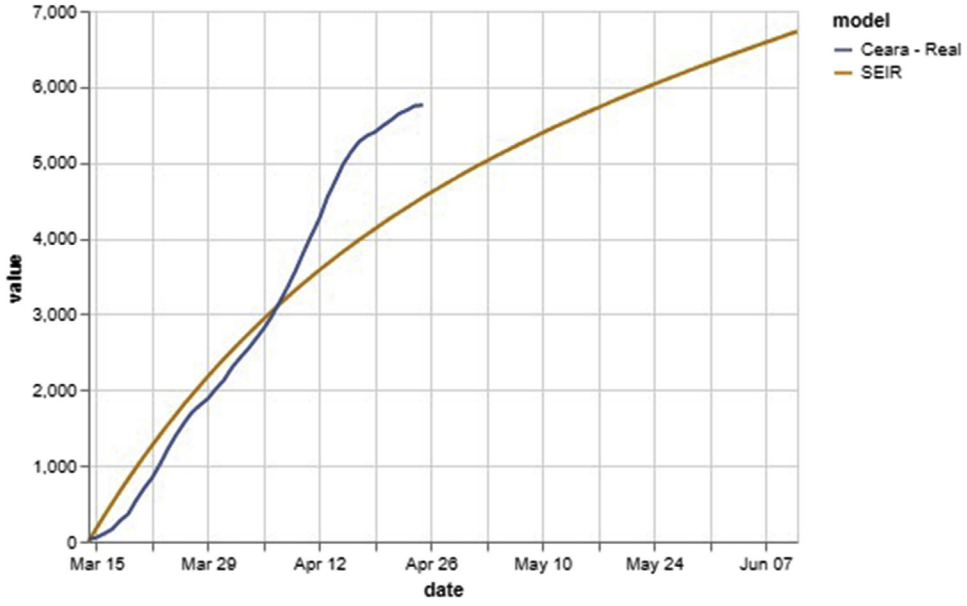


FIGURE 1.7 SEIR result short term Ceará.

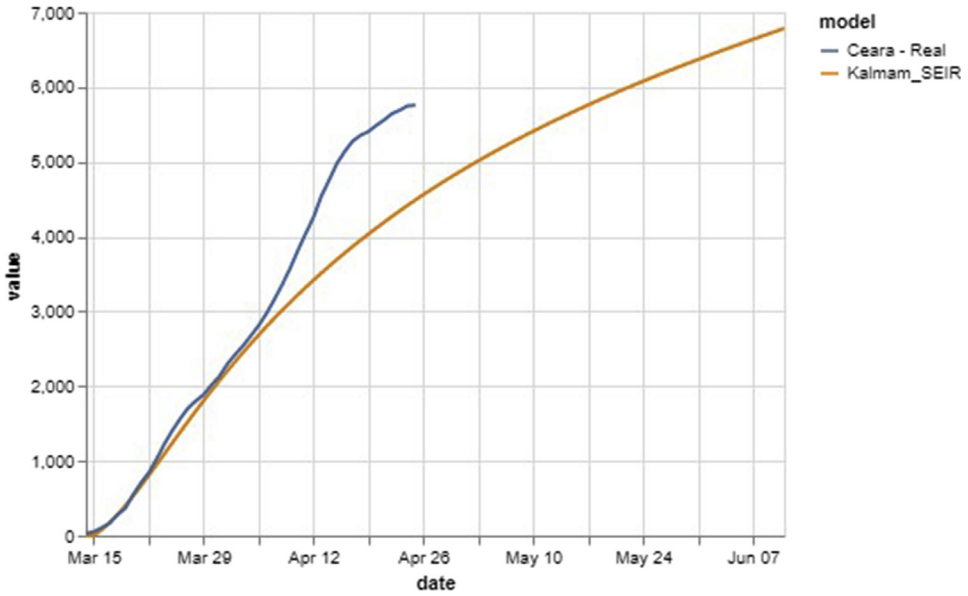


FIGURE 1.8 Kalman Filter and SEIR short term results to Ceará dataset.

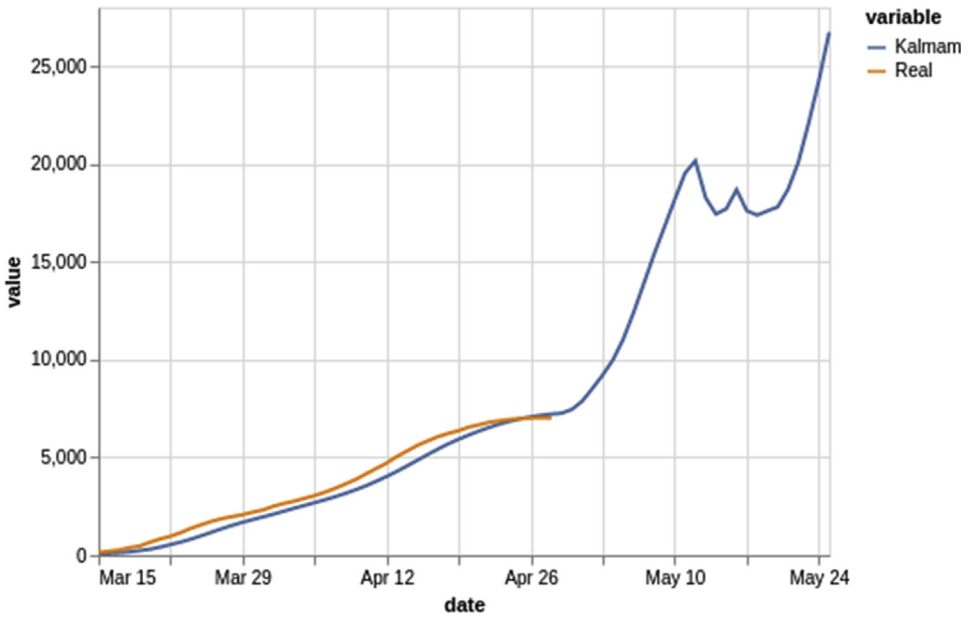


FIGURE 1.9 Kalman filter result short term Ceará.

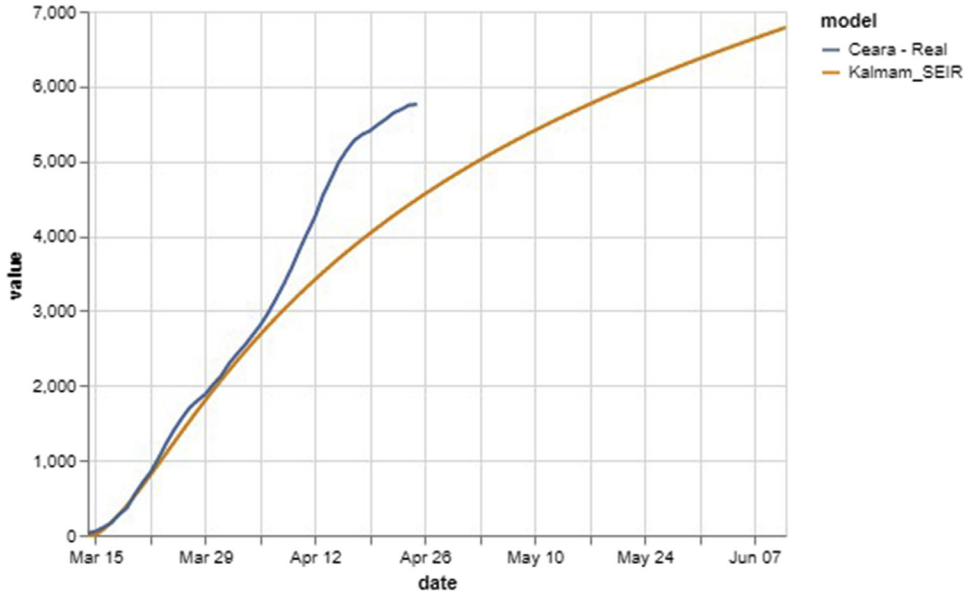


FIGURE 1.10 Kalman filter and SEIR result short term Ceará.

For the KF, we use three approaches, the first shown in Fig. 1.10, which uses only the KF, as it is an adaptive method, it is necessary other data, the forecast is based on data from Brazil. Adapting the filter to the data proved to be effective, and making it a suitable method for short-term forecasts. The second approach using the KF is to use the SEIR method, generating the result shown in Fig. 1.10, in this case, the data generated from the SEIR model were used in the filter. The third and last is the use of the hybrid data set, which consists of joining the data from Ceará and data generated from the SEIR model, before applying the data in the KF (Fig. 1.11).

5.2 Long-term results for China data set

To validate the approaches for long-term methods, we use data from the epidemic in China, since in Ceará and in everyone, who was infected with COVID-19 still in the process of evolution or decay. And the data from China become better for this validation, which has the entire evolution of the epidemic, from beginning to end. The errors for each method are shown in Fig. 1.12. For a better comparison, we allocated all methods in just one graph. The KF obtained the best score among the other methods because KF adapts to the data.

Epidemiological models have returned to a worse scenario that happened. Because the model predicts that the epidemic would last more months, but with the measures adopted by the country, the scenario was changed.

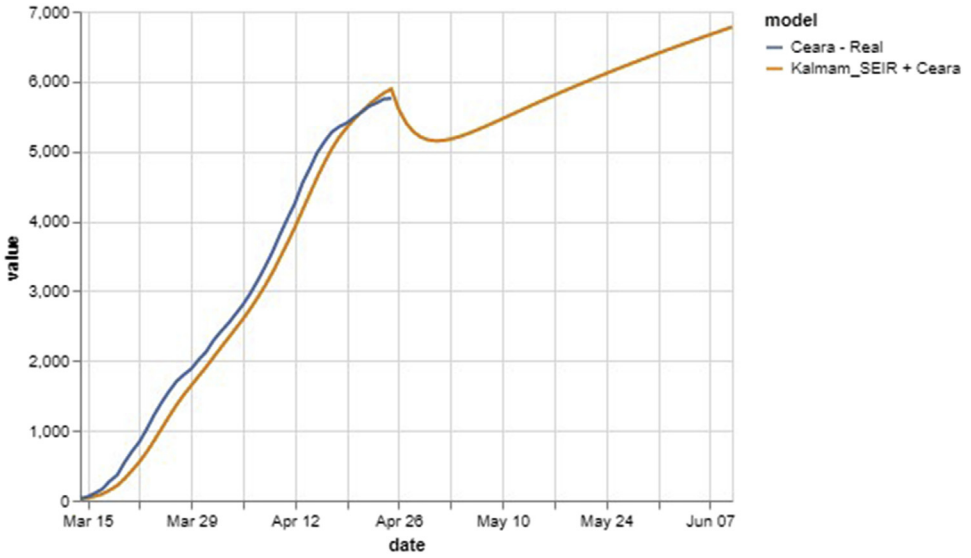
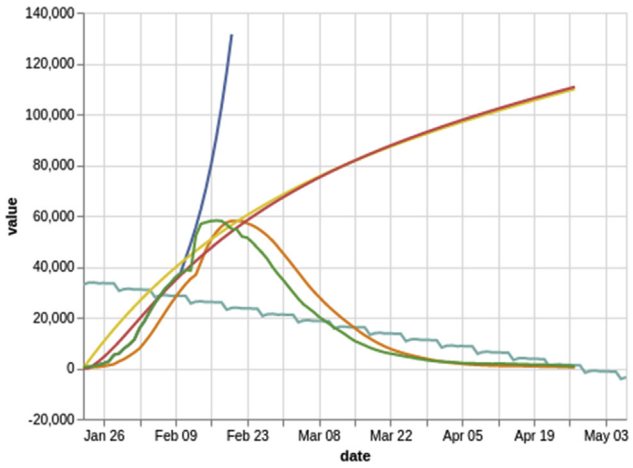


FIGURE 1.11 Kalman filter, SEIR + Ceará result short term Ceará.



Method	MAE	RMSE	R2
Kalman Filter	3038.40	4099.15	0.952
Prophet	11825.02	16070.89	0.275
Holt Winters	9158.26	21149.54	0.007
KF + SEIR	53263.11	67200.52	-11.614
SEIR	53957.37	66963.28	-11.553

FIGURE 1.12 Long-term results and method errors to China data set.

The Holt–Winters method predicts a high-order exponential growth, as it deals with an exponential smoothing method, when faced with the considerable growth of cases it tends to continue growing according to the curve, tending to large values.

Prophet model the one considered best in the short term, not obtain a reasonable efficiency for the long term, due in no small amount of data and the change of up and down, the method tries to look for seasonality and trend. However, the data do not have such evident characteristics, thus resulting in small periods of fall and rise.

5.3 Long-term results for hybrid data set

Long-term results with the hybrid data set are those using data from Ceará, Brazil, and data from China to complement the rest of the epidemic. In this approach, it is possible to verify the behavior of the models in long-term forecasts, with the help of data from China.

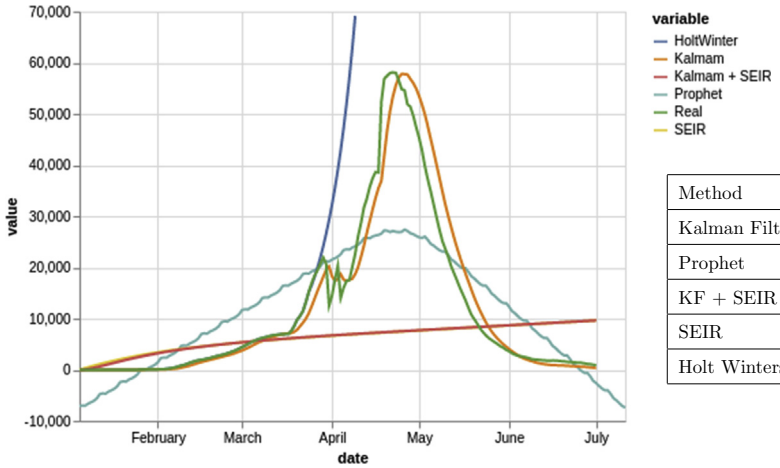
Fig. 1.13 shows the results based on the hybrid data set. There is a strong resemblance to the long-term result using data from China; the best model is still the KF. Like the forecast that uses data from China, Holt–Winters obtained an exponential increase and tended to high values. The errors for each method using the hybrid data set are shown in Table.

The Prophet method has a large error for long-term predictions, but now its prediction has taken a different form compared to the result using data from China. It is noticeable that he was able to model the shape of the growth, peak, and decay of the curve, but the forecast values for the number of cases were different, resulting in a big error.

In this case, the epidemiological models again proved to be almost identical, with a small difference at the beginning of the forecast. The forecast also appears shows that the growth of the virus would continue for a long term.

6. Final considerations

Forecasting plays an essential role in several areas of study due to its benefits on saving resources or improving the decision-making process to benefit the economy. In the case of the COVID-19 outbreak, there are many challenges for forecasting as the COVID-19 incubation period is much more extended than other epidemic processes, and a small number of datasets are available for this purpose.



Method	MAE	RMSE	R2
Kalman Filter	2314.96	3947.14	0.937
Prophet	6946.42	9622.75	0.631
KF + SEIR	9785.30	16207.05	-0.046
SEIR	9861.01	16235.39	-0.050
Holt Winters	16601.70	43231.31	-7.560

FIGURE 1.13 Long-term results and method errors to long-term experiments using China and Ceara data set.

The prediction models infer that the amount of COVID-19 cases expands exponentially in its increasing phase. The exponential increase of cases strongly suggests that the epidemic growth is due to an underlying biological phenomenon rather than due to the number of tests performed. Some studies imply that there is a certain generality in the temporal evolution of COVID-19. Some time-frame plots of the confirmed infected individuals of China, Italy, and France demonstrate a generalization on the curve of cases and a pattern on the collapse of the health system of each country. Although these facts, in a limited community, the exponential growth of cases can not remain forever. Hence, the stochastic model of infection spread saturates sometime.

This chapter covered forecasting techniques to predict the number of positive instances of COVID-19 based on real data obtained from different epidemic locations, such as China, Brasil, in general, and, more specifically, the State of Ceará, in the northeast region of Brazil. By using classic epidemiological methods and innovative machine learning techniques, together with historical data of the pandemic in the State, the proposed techniques obtained results close to the real number of cases, and in some scenarios, the exact number of cases. In short-term predictions, Prophet and Holt–Winter obtained better results; this fact can be explained due to the exponential nature of the epidemic curve. In long-term predictions, the KF obtained the best score among the other methods.

Finally, it is essential to highlight that determining the indicator of the actual number of infected people depends on a wide variety of circumstances such as massive testing, social isolation, under-reporting of cases, among others. By not considering these factors, the purpose of this paper is limited to predicting the number of diagnosed cases to assist public policy decision-making in combating the new coronavirus pandemic.

References

[1] G.R. Shinde, A.B. Kalamkar, P.N. Mahalle, Forecasting models for coronavirus (COVID-19): a survey of the state-of-the-art, SN Comput. Sci. 1 (4) (April 2020).

[2] S.B. Bastos, D.O. Cajueiro, Modeling and Forecasting the Early Evolution of the Covid-19 Pandemic in Brazil, 2020 arXiv:2003.14288.

[3] D. Fanelli, F. Piazza, Analysis and forecast of COVID-19 spreading in China, Italy and France, Chaos, Solitons Fractals 134 (2020) 109761, https://doi.org/10.1016/j.chaos.2020.109761, arXiv:2003.06031.

[4] C. Viboud, L. Simonsen, G. Chowell, A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks, Epidemics 15 (2016) 27–37, https://doi.org/10.1016/j.epidem.2016.01.002.

[5] P. Manliura Datilo, Z. Ismail, J. Dare, A review of epidemic forecasting using artificial neural networks, Int. J. Epidemiol. Res. 6 (3) (2019) 132–143, https://doi.org/10.15171/ijer.2019.24.

[6] L. Wynants, B. Van Calster, M.M.J. Bonten, G.S. Collins, T.P.A. Debray, M. De Vos, M.C. Haller, G. Heinze, K.G.M. Moons, R.D. Riley, E. Schuit, L.J.M. Smits, K.I.E. Snell, E.W. Steyerberg, C. Wallisch, M. van Smeden, Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal, BMJ (Clin. Res. ed.) 369 (2020) m1328, https://doi.org/10.1136/bmj.m1328.

[7] M.K. Pandey, K. Subbiah, Performance analysis of time series forecasting using machine learning algorithms for prediction of ebola casualties, Commun. Comput. Inf. Sci. 899 (January) (2019) 320−334, https://doi.org/10.1007/978-981-13-2035-4_28.

[8] C. Anastassopoulou, L. Russo, A. Tsakris, C. Siettos, Data-based analysis, modelling and forecasting of the COVID-19 outbreak, PLoS One 15 (3) (2020) e0230405, https://doi.org/10.1371/journal.pone.0230405.

[9] Z. Hu, Q. Ge, S. Li, L. Jin, M. Xiong, Artificial Intelligence Forecasting of Covid-19 in China, 2020, pp. 1−20. URL, http://arxiv.org/abs/2002.07112.

[10] Q. Yang, C. Yi, A. Vajdi, L.W. Cohnstaedt, H. Wu, X. Guo, C.M. Scoglio, Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China, Infect. Dis Model (2020). https://doi.org/10.1101/2020.03.27.20045625.

[11] J. Kumar, K. P. S. S. Hembram, Epidemiological study of novel coronavirus (COVID-19). arXiv:2003.11376. URL http://arxiv.org/abs/2003.11376.

[12] P. Teles, Predicting the evolution of SARS-Covid-2 in Portugal using an adapted SIR Model previously used in South Korea for the MERS outbreak (April). arXiv:2003.10047. URL http://arxiv.org/abs/2003.10047.

[13] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, J. He, Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, J. Thorac. Dis. 12 (3) (2020) 165−174. https://doi.org/10.21037/jtd.2020.02.64.

[14] L. Jia, K. Li, Y. Jiang, X. Guo, T. Zhao, Prediction and analysis of Coronavirus Disease 2019 (December). arXiv:2003.05447. URL http://arxiv.org/abs/2003.05447.

[15] H. Zhu, Transmission dynamics and control methodology of COVID-19: a modeling study, Appl. Math model. (2020). https://doi.org/10.1101/2020.03.29.20047118.

[16] H.H. Elmousalami, A.E. Hassanien, Day Level Forecasting for Coronavirus Disease (COVID-19) Spread: Analysis, Modeling and RecommendationsarXiv:2003.07778. URL http://arxiv.org/abs/2003.07778.

[17] B.M. Ndiaye, L. Tendeng, D. Seck, Analysis of the COVID-19 pandemic by SIR model and machine learning technics for forecastingar. Xiv:2004.01574. URL http://arxiv.org/abs/2004.01574.

[18] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen, B. Cao, Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study, Lancet 395 (10229) (2020) 1054−1062. https://doi.org/10.1016/S0140-6736(20)30566-3.

[19] D. Liu, L. Clemente, C. Poirier, X. Ding, M. Chinazzi, J.T. Davis, A. Vespignani, M. Santillana, A Machine Learning Methodology for Real-Time Forecasting of the 2019-2020 COVID-19 Outbreak Using Internet Searches , News Alerts , and Estimates from Mechanistic Models (D). arXiv:3122774.

[20] A.S. Rao, J.A. Vazquez, Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey in the populations when cities/towns are under quarantine, Infect. Control Hosp. Epidemiol. 1400 (2020). https://doi.org/10.1017/ice.2020.61.

[21] M.A.A. Al-qaness, A.A. Ewees, H. Fan, M. Abd El Aziz, Optimization method for forecasting confirmed cases of COVID-19 in China, J. Clin. Med. 9 (3) (2020) 674. https://doi.org/10.3390/jcm9030674.

[22] X. Zhou, X. Ma, N. Hong, L. Su, Y. Ma, J. He, H. Jiang, C. Liu, G. Shan, W. Zhu, S. Zhang, Y. Long, Forecasting the worldwide spread of covid-19 based on logistic model and SEIR model, medRxivarXiv: https://www.medrxiv.org/content/early/2020/04/08/2020.03.26.20044289.

[23] E. Beretta, Y. Takeuchi, Global stability of an SIR epidemic model with time delays, J. Math. Biol. 33 (3) (1995) 250−260. https://doi.org/10.1007/BF00169563.

[24] D. Schenzle, An age-structured model of pre-and post-vaccination measles transmission, Math. Med. Biol. J. IMA 1 (2) (1984) 169–191.

[25] L. Stone, B. Shulgin, Z. Agur, Theoretical examination of the pulse vaccination policy in the SIR epidemic model, Math. Comput. Model. 31 (4–5) (2000) 207–215. https://doi.org/10.1016/S0895-7177(00)00040-6.

[26] K.L. Cooke, Stability analysis for a vector disease model, Rocky Mt. J. Math. 9 (1) (1979) 31–42. https://doi.org/10.1216/RMJ-1979-9-1-31.

[27] S.J. Taylor, B. Letham, Forecasting at scale, Am. Statistician 72 (1) (2018) 37–45.

[28] S. Gelper, R. Fried, C. Croux, Robust forecasting with exponential and holt-winters smoothing, J. Forecast. 29 (3) (2010) 285–300. https://doi.org/10.1002/for.1125.

[29] P.R. Winters, Forecasting sales by exponentially weighted moving averages, Manag. Sci. 6 (3) (1960) 324–342. URL https://EconPapers.repec.org/RePEc:inm:ormnsc:v:6:y:1960:i:3:p:324-342.

[30] C.C. Holt, Forecasting seasonals and trends by exponentially weighted moving averages, Int. J. Forecast. 20 (1) (2004) 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015. URL http://www.sciencedirect.com/science/article/pii/S0169207003001134.

[31] S. Haykin, Kalman Filtering and Neural Networks, vol. 47, John Wiley & Sons, 2004.

[32] W. Yang, A. Karspeck, J. Shaman, Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics, PLoS Comput. Biol. 10 (4) (2014). https://doi.org/10.1371/journal.pcbi.1003583.

[33] B. Cazelles, N.P. Chau, Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic, Math. Biosci. 140 (2) (1997) 131–154. https://doi.org/10.1016/S0025-5564(96)00155-1.

[34] R.J.M.N.D. Singpurwalla, Understanding the Kalman Filter, 2005.

[35] J.K. Uhlmann, S.J. Julier, A new extension of the Kalman filter to nonlinear systems, Signal Proces. Sensor Fusion Target Recognit. VI (3068) (1997) 182–194.

[36] J. Mandel, J.D. Beezley, L. Cobb, A. Krishnamurthy, Data driven computing by the morphing fast Fourier transform ensemble Kalman filter in epidemic spread simulations, Proced. Comput. Sci. 1 (1) (2010) 1221–1229. https://doi.org/10.1016/j.procs.2010.04.136.