Journal of Data Science 19 (2), 269–292
April 2021

DOI: 10.6339/21-JDS991
Data Science in Action

Time Series Regression Models for COVID-19 Deaths

Marinho G. Andrade[1,*], Jorge A. Achcar[2], Katiane S. Conceição[1], and Nalini Ravishanker[3]

[1] *Department of Applied Mathematics and Statistics, ICMC, University of São Paulo, São Carlos/SP, Brazil*
[2] *Faculty of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto/SP, Brazil*
[3] *Department of Statistics, University of Connecticut, Storrs, CT, USA*

Abstract

This article develops nonlinear functional forms for modeling count time series of daily deaths due to the COVID-19 virus. Our models explain the mean levels of the time series while accounting for the time-varying variances. A Bayesian approach using Markov chain Monte Carlo (MCMC) is adopted for analysis, inference and forecasting of the time series under the proposed models. Applications are shown for time series of death counts from several countries affected by the pandemic.

Keywords *Bayesian approach; nonlinear model; pandemic cycle*

1 Introduction

Coronavirus (COVID-19) is a new pandemic viral infection that has been spreading worldwide in the year 2020, and is caused by a newly discovered coronavirus. The virus emerged in the city of Wuhan, China, at the end of 2019, and as of August 20, 2020, it has accounted for 22,263,347 globally confirmed cases and 782,471 deaths according to the World Health Organization (World Health Organization, 2020a). The disease incidence rate grows on an exponential scale with global geographical expansion to almost all countries of the world. As the epidemic started on December 31, 2019, WHO (World Health Organization) was informed of a cluster of cases of pneumonia of unknown cause detected in Wuhan City, Hubei Province of China. The COVID-19 was identified as the causative virus by Chinese authorities on January 7, 2020 (World Health Organization, 2020b). On January 30, 2020, following the recommendations of the Emergency Committee, WHO declared that the outbreak constitutes a public health emergency of international concern.

The COVID-19 virus affects different people in different ways. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness, and will recover without requiring special treatment (World Health Organization, 2020d). Common symptoms include fever, tiredness and dry cough. Other symptoms include: shortness of breath, aches, and pains, sore throat, very few people will report diarrhea, nausea, or a runny nose (World Health Organization, 2020b). Older people and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop severe symptoms. It is believed that the COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes. While there are many ongoing clinical trials aimed at possible treatments, there are no specific vaccines or treatments for COVID-19 at this time.

[*] Corresponding author. Email: marinho@icmc.usp.br.

Received August 1, 2020; Accepted November 1, 2020

In the Situation Report World Health Organization (2020c) of WHO (January 21, 2020), there were only 282 confirmed cases in the region covering China, Japan, Republic of Korea, and Thailand. Since December 2019, the pandemic has shown dynamic patterns of varying extents of increase, leveling off, or decrease in different areas. Figure 1 in Section 2 presents time series plots of the daily notified cases of COVID-19 in China, USA, Spain, Italy, Brazil and India in the period ranging from January 1, 2020 to May 30, 2020. It can be seen that during this period, USA has the majority of daily notifications compared to other countries. Smaller numbers of COVID-19 notification cases are observed in China, possibly indicating that the pandemic is under control due to containment strategies taken by the authorities and/or the use of diagnostic testing on the entire population. The situations in other countries vary and are discussed in Section 2. In order to track the virus, the World Health Organization has updated the Laboratory Testing Strategy on March 21, 2020 (World Health Organization, 2020d) according to different transmission scenarios: countries with no cases; countries with one or more occurrences (sporadic cases); countries experiencing clusters of cases across periods, geographic locations, or common exposure (clustered cases); or countries experiencing more significant outbreaks or sustained and pervasive local transmission (community transmission).

There is considerable recent literature related to COVID-19. The research emerges from many different fields of study including viral origin and structure (e.g., Lan et al., 2020; Shang et al., 2020; Lam et al., 2020), epidemiology (e.g., Ferretti et al., 2020), preclinical research (e.g., Kim et al., 2020), diagnostic and serology (e.g., Ju et al., 2020; Wölfel et al., 2020) and therapy and clinical trials (e.g., Shen et al., 2020). In these studies, the main focus has been identification of the symptoms of COVID-19 in order to develop an antigen (clinical trials and diagnostics), to track the presence of the virus using an app that builds a memory of proximity contacts (epidemiology) and the identification of the viral origin and the COVID-19 structure. According to Lan et al. (2020) and Andersen et al. (2020), the virus appears to be optimized for binding to the human ACE2 receptor; the spike protein of SARS-CoV-2 has a functional polybasic cleavage site at the S1-S2 boundary through the insertion of 12 nucleotides. Moreover, the receptor-binding domain (RBD) in the spike protein is the most variable part of the coronavirus genome, that is, six RBD amino acids have been shown to be critical for binding to ACE2 receptors and for determining the host range of SARS-CoV-like viruses (Andersen et al., 2020; Wan et al., 2020). A great research effort is seen worldwide in the last three months related to different ways of understanding the pandemic with hundreds of papers being published in a very short time period, (e.g., Kandel et al., 2020; Pung et al., 2020; Chan et al., 2020; Huang et al., 2020; Wu et al., 2020; Lu et al., 2020; Chen et al., 2020b; Li et al., 2020; Lai et al., 2020; Lupia et al., 2020; Shereen et al., 2020; Chen et al., 2020a; Sohrabi et al., 2020; Han et al., 2020; Chen, 2020; Wu et al., 2020; Zhao et al., 2020). Most papers are related to the transmission mechanisms of coronavirus among different persons, genomic characterization and epidemiology of COVID-19, new treatments to COVID-19, clinical features of patients infected with the virus, developments of vaccines, effects of the confinement of entire populations to minimize the spread of the disease and not overburden the health systems and to decrease the mortality rate, especially for the elderly and people with comorbidities. Another point of great interest is related to predictions of the quarantine time adopted by almost all countries in the world following WHO directions to decrease and delay new infections, but that could be catastrophic in economic terms for many countries in the world.

Mathematical models of epidemic dynamics have been used to understand patterns and predict the evolution of the pandemic. The input for these models is the time series of the number of susceptible, infected, and recovered (SIR) persons (Kermack and McKendrick, 1927; Lili Wang et al., 2020). Notification by health centers of the number of infected people depends on the efficiency with which tests and diagnostics are performed on people with suspected infection. One well-known problem is that when underreporting of the number of infected people is prevalent, the reported data can provide very optimistic predictions for the pandemic state. Another problem is caused by delays in reporting the death counts on some days, resulting in delays and discrepancies in reporting. A model that tries to solve the problem with inconsistent data and underreporting in the number of infected people was proposed by Lili Wang et al. (2020). However, detecting discrepant data and eliminating the effect of these data on model fit is easier than estimating unreported data on the number of people infected because it is unobserved data.

An analysis that can contribute to assessing and predicting the future state of a pandemic can be done by considering statistical models for the mean number of deaths in a regions. It is also crucial to understand that some factors not considered in such a model curve can lead to errors in forecasting of the number of deaths for a given day. Among these, one of the main factors is data consolidation. Consolidation is done by medical reports based on clinical and laboratory tests that prove the cause of death. Due to an excess of cases, some laboratory tests to certify the confirmation of the cause of death by COVID-19 delay the confirmation, causing an accumulation of late notifications that are reported on certain days together with the consolidated cases on the same day. This issue can cause significant peaks in the number of deaths on certain days, especially when the pandemic reaches its peak, together with other factors, such as the capacity of the health system to be exhausted, difficulty in medical care, etc., lead to the heteroscedasticity observed in the daily death counts data.

The use of nonlinear regression models has been well known in several areas such as, modeling of so called dose-response relation, tumor growth, pharmacology and in the analysis of epidemic data. Some papers have proposed nonlinear regression models to model the number of cases or number of deaths by COVID-19. Luo (2020) uses the logistic model to predict the end of the pandemic. Zhang et al. (2020) uses Gamma models to predict turning points, durations and attack rates of COVID-19. Girardi et al. (2020) presents a robust Bayesian approach on five parameter log-logistic curve models to model the number of deaths curve in some cities in Italy. Tsallis and Tirnakli (2020) present a unified nonlinear functional form for predicting COVID-19 peaks for several countries. The pattern of the COVID-19 death curve is highly variable depending on many factors, restrictive measures, and the capacity of the health services in different countries. For this reason, it is not possible to use a single function to model the epidemic in different locations in the world. In this paper, a nonlinear regression statistical model is proposed for the daily counts of COVID-19 deaths in a few different regions of the world, using a few nonlinear functions. It is essential that the model considered for the mean of the time series reproduces the asymmetric pattern of these series. Another critical aspect to be included in the model is the conditional heteroscedasticity of the time series. A good model for the variance is fundamental to calculate probabilities and more accurate credible intervals for the forecasts.

In this paper, we have described seven nonlinear functions for modeling the time series of the number of daily deaths caused by the COVID-19 pandemic. Applications of the proposed models are shown using time series of daily deaths from six countries with different characteristics and at different stages of the pandemic. The countries we selected for this analysis fall into three groups. Group G1 consists of a country (China) where the epidemic originated but is currently reported to be well under control. Group G2 consists of European countries (Italy and Spain) which were the first countries in Europe to detect cases of contamination. But at the moment, they have already passed the peak of the pandemic. Group G3 consists of large countries (USA and Brazil) of continental dimensions located in the American continent.

Table 1: Pandemic numbers by country on August 20, 2020.

Country	Population	Date (No.) of the 1st. confirmed case	Total No. cases (Aug 20, 2020)	Total No. deaths (Aug 20,2020)
China	1,411,124,099	Dec 31, 2019 (27)	89,455	3,391
USA	333,761,149	Jan 21, 2020 (1)	5,482,416	171,821
Spain	45,680,787	Feb 1, 2020 (1)	380,421	29,554
Italy	60,020,970	Jan 31, 2020 (3)	255,054	35,462
Brazil	217,256,782	Feb 26, 2020 (1)	3,466,609	112,116
India	1,394,618,848	Jan 1, 2020 (1)	2,767,273	51,091

Although their population sizes are nearly the same, they have vastly different development conditions. In these two countries, the peak was only recently reached. Group G4 consists of an extremely densely populated country (India) where the epidemic seems to have spread slowly and where the peak still to be reached in the coming months. A result of a July serological survey of 6,936 people across three suburbs of Mumbai found that more than 50% of people across parts of India's financial hub of Mumbai have coronavirus antibodies. It is indicating that the population may have inadvertently achieved the controversial "herd immunity" protection from the coronavirus. It explains why a steep drop in infections is being seen among the closely-packed population, despite new cases increase overall in the hard-hit country. (Read more at: https://www.bloombergquint.com/coronavirus-outbreak/herd-immunity-seems-to-be-developing -in- mumbai-s-poorest-areas).

The purpose of our analysis is to match the best nonlinear functional form for the pattern of the expected number of daily deaths curve for each of the four groups, G1-G4, particularly to identify the timing of the peak and the rate and nature of leveling off in the curve. Such an understanding is crucial to understanding how COVID-19 deaths differentially progressed in these countries, bringing different approaches for its management and control. Further, such an analysis is useful to anticipate the peak timing and the life-cycle in a future second wave of this epidemic, or similar epidemics in the future - although we all hope this will not be the case. We also note that although we have shown our analysis for time series collected at the national level, a similar analysis is possible at different smaller zonal levels, for example, for many US, Brazilian, or Indian states.

The paper is organized as follows. In Section 2, we present a description of the data sets used and discuss the main objective of the paper. Section 3 develops the functions for modeling the mean and variance of the time series and the Bayesian approach used to do model fitting and make inference about the model parameters. Section 4 shows results from time series from the different countries, while Section 3.4 presents a Bayesian forecasting procedure for future daily death counts. Section 5 presents a summary and discussion.

2 COVID-19 Data Description

We have considered data from six countries, i.e., China, USA, Spain, Italy, Brazil and India. A summary of pandemic numbers by country is presented in Table 1.

For each country, we model and predict the count time series of daily deaths by COVID-19, starting from the day when the first infected case was confirmed in the country. An exception is the time series of the number of deaths in China, where the first record was made on December 31, 2019, with 26 infected cases and 0 deaths. Therefore, for different countries, the time series start on different days. For all series, we consider observations until July 31, 2020 for model fitting. We analyze the series of daily deaths from these six countries individually. We restrict ourselves to modeling time series of deaths and not time series of new infected cases because the series of infected cases depends on the number of tests that each state has performed and the data may not accurately reflect the size of the problem. The time series used in this paper were taken from the website `https://countrymeters.info/pt` and the European Centre for Disease Prevention and Control (ECDC) website `https://www.ecdc.europa.eu/en/publications-data`. Figures 1a and 1b show the number of new daily infected cases and the logarithm of the accumulated daily number of infected cases. Figures 1c and 1d show the number of daily deaths and the logarithms of the accumulated daily number of deaths. As we mentioned earlier, we only model the time series of daily deaths in this paper.

In Section 3, we describe a Bayesian approach for fitting a nonlinear statistical model to represent the mean pattern for the daily series of new deaths in each country (Ratkowsky, 1983; Bates and Watts, 1988). The detected pattern is then used to forecast future values of the time series.

3 The Model

Let $\{Y(t),\ t = 1, 2, \ldots, T\}$ be a time series defined over nonnegative integers, representing the number of daily deaths by COVID-19 in a single location. We consider the following model for $Y(t)$:

$$Y(t) = \eta(t) + Z(t), \tag{1}$$

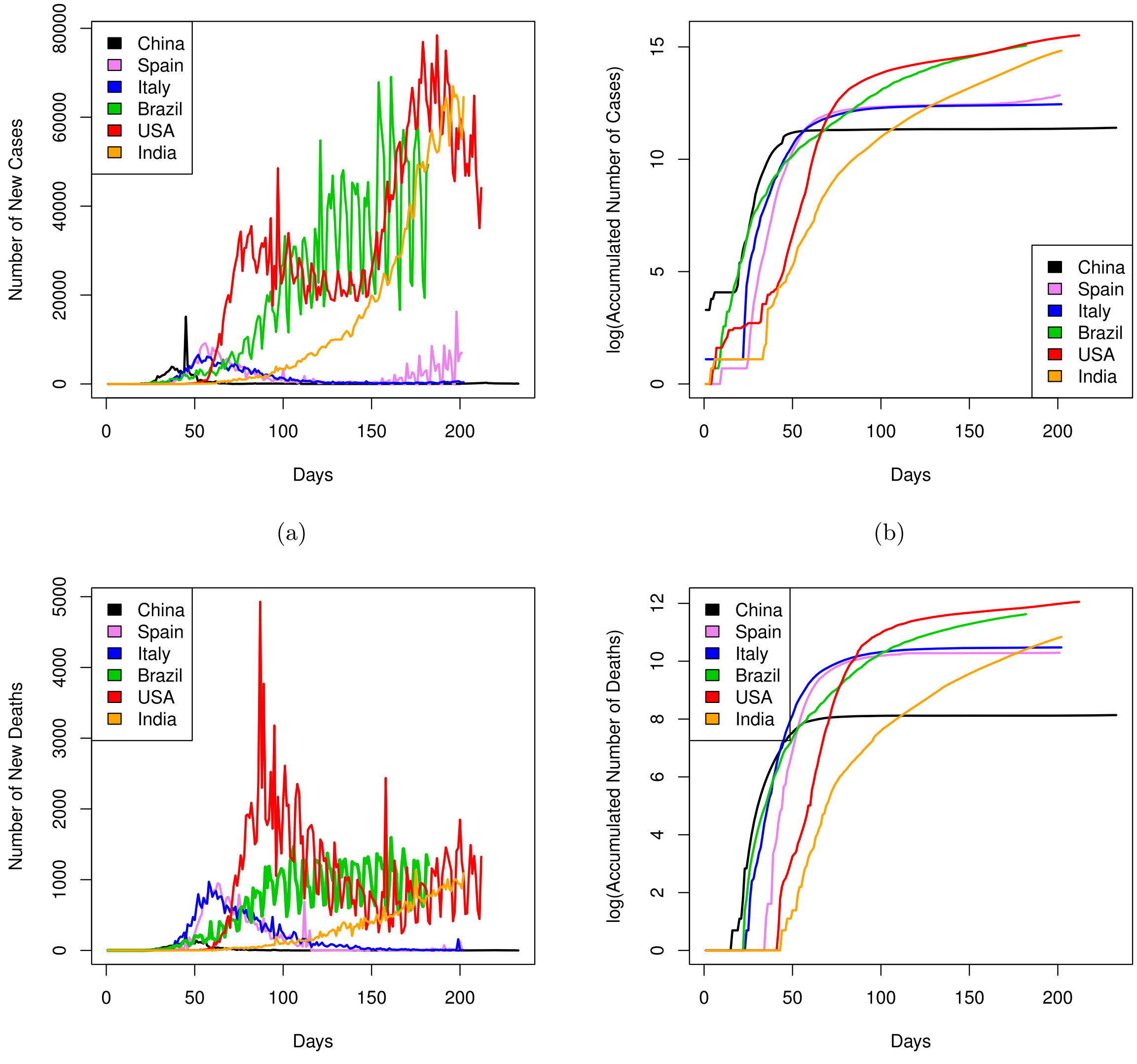
where $\eta(t)$ is a deterministic nonlinear function of time whose forms are defined in Table 2, and $Z(t)$ is an autoregressive conditionally heteroscedastic stochastic process given by

$$\begin{aligned} Z(t) &= \sum_{i=1}^{p} \phi_i Z(t-i) + W(t), && (2)\\ W(t) &= \sigma(t)\varepsilon(t), && (3)\\ \sigma^2(t) &= \alpha_0 + \alpha_1 W^2(t-1), && (4) \end{aligned}$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the pth order autoregressive polynomial with all roots outside the unit circle, B is the backshift operator, $Var(W(t)) = \sigma^2(t)$ which follows an autoregressive conditionally heteroscedasticity (ARCH) model of order 1 (Engle, 1982, 1983) with $\alpha_0 > 0$ and $0 \leqslant \alpha_1 < 1$ and $\varepsilon(t)$ is an i.i.d. normally distributed process with $E(\varepsilon(t)) = 0$, $Var(\varepsilon(t)) = 1$.

This model enables us to model the temporal dependence in the first two moments of the residuals from the nonlinear regression fit. Although a variance stabilizing transformation such as the logarithmic transformation of the $Y(t)$ may be considered, we instead use a conditionally heteroscedastic model.

Plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the regression residuals and the squared residuals, together with the Lagrange Multiplier (LM) test (Engle, 1982) enables us to verify the presence of autoregressive or ARCH effects in the residuals.



(a) (b)

(c) (d)

Figure 1: (a) Number of new daily infected cases and (b) the logarithm of the accumulated daily number of infected cases. (c) Number of new daily deaths and (d) the logarithm of the accumulated daily number of deaths.

We use as candidates for the $\eta(t)$ model the functions provided in Table 2. Although these functions are widely known, we make some observations about them here.

(i) The generalized exponential function (Gupta and Kundu, 2001) has excellent shape flexibility depending on the parameter θ_2. If $\theta_2 > 1$, this function has a unimodal asymmetric form, whose mode is given by $\theta_1^{-1} \log(\theta_2)$.

Table 2: Alternative functions for $\eta(t)$ to represent expected values of $Y(t)$.

Function	$\eta(t)$
Gaussian	$\theta_0 \exp\left(-\frac{(t-\theta_1)^2}{\theta_2}\right)$
Log-Normal	$\frac{\theta_0}{t} \exp\left(-\frac{(\log(t)-\theta_1)^2}{\theta_2}\right)$
Alpha	$\frac{\theta_0}{t^2} \exp\left(-\frac{1}{2}\left(\frac{\theta_1}{t}-\theta_2\right)^2\right)$
Generalized Exp.	$\frac{\theta_0(1-\exp(-\theta_1 t))^{\theta_2-1}}{\exp(\theta_1 t)}$
Rational function	$\frac{\theta_0+\theta_1 t}{1+\theta_2 t+\theta_3 t^2}$
Logistic	$\frac{\theta_0 \exp(-\theta_2 t)}{1+\theta_1 \exp(-\theta_2 t)}$
Gamma	$\theta_0 t^{\theta_1} \exp(-\theta_2 t)$

(ii) The Rational function has been widely used in the context of statistical modeling (especially process modeling), as an empirical technique to fit curves (Kulikov, 2001; Press et al., 2007). This function also has an asymmetric shape, which is an important characteristic of our data. A brief description of the class of Rational functions is presented in in the supplementary material.

(iii) The alpha function has been used in the context of Product Life-Cycles (Petrescu, 2009) that are widely used in business applications. This curve has two inflection points, and it allows us to model four phases of the pandemic cycle. The first phase when growth accelerates, the second when the speed of the growth diminishes before reaching the peak (in general this can be a reflection of containment measures), the third phase occurring after the maximum when the curve starts to decline, and the fourth, after starting the relaxation of containment.

Given $\hat{\eta}(t)$, we construct the residual process $\widehat{Z}(t)$ given as

$$\widehat{Z}(t) = Y(t) - \hat{\eta}(t).$$

To simplify the notation we denote $\mathcal{F}_t = \{Y(t-1), ..., Y(1), \eta(t-1), ..., \eta(1)\}$ all information available until time t, we define the conditional mean $E(Y(t)|\mathcal{F}_t) = \mu(t)$ given by

$$\mu(t) = \eta(t) + \sum_{i=1}^{p} \phi_i Z(t-i), \tag{5}$$

One of the biggest challenges in modeling the pandemic is to find the time, denoted by τ_{peak}, at which the highest number of deaths (or cases) occurs. Finding τ_{peak} allows us to assess the expected value of the stochastic process $E(Y(\tau_{peak})|\mathcal{F}_{\tau_{peak}}) = \mu(\tau_{peak})$. To find the value of τ_{peak}, consider a positive integer t_{peak} such that $\mu(t_{peak}) \geqslant \mu(t)$, $\forall\ t \in \mathbb{N}$. Then, τ_{peak} is estimated by t_{peak}.

Another challenge is to evaluate the probabilities $P(Y(t) \leqslant U)$ for fixed values of U that allows us to estimate the end of the pandemic. However, evaluating these probabilities is only possible by making assumptions about the probability distribution of the white noise $W(t)$ and finding the appropriate models for the variances $\sigma^2(t)$. As we see above, we have assumed that $\{W(t), t \geqslant 0\}$ are i.i.d. $N(0, \sigma^2(t))$ variables and $\sigma^2(t)$ follows a stationary ARCH(1) process.

The probability $P(Y(t) < U)$ can be easily calculated as

$$P(Y(t) \leqslant U) = \Phi\left(\frac{U - \mu(t)}{\sigma^2(t)}\right), \forall\ t > 0,$$

where $\Phi(z)$ is the standard normal N(0,1) cumulative distribution function (cdf). The estimate of t_{peak}, as well the estimates of the probabilities $P(Y(t) < U)$ ($\forall\ t > 0$), can provide extremely useful information for decision-making and resource management to cope with a pandemic.

When there is interest in analyzing the *accumulated* number of deaths (or cases), we can consider the accumulated processes given by

$$S(t) = \sum_{k=0}^{t} Y(k),$$

in which case, we have

$$E(S(t)) = \sum_{k=0}^{t} \mu(k) \qquad \text{and} \qquad Var(S(t)) = \sum_{k=0}^{t} \sigma^2(k).$$

For accumulated processes, the main objective is to look for sigmoidal functions that best fit these curves. A function $S(t)$ in sigmoidal form ideally has its origin in $S(0)$, inflection point (peak) occurring at the beginning of the first third of the stage before approaching the maximum value, with an asymptote to be reached when the process enters a stage of decay. In a pandemic, this change in curvature occurs closer to the peak point. Some sigmoidal functions, used to model growth curves, are presented in Desta et al. (1999) and Sonnino (2020).

3.1 Model Fitting

The Bayesian approach provides an attractive method for analyzing the time series of daily death counts and generally is more reliable in small-sample inference for nonlinear regression than the least squares approach (Katz et al., 1981; Dorndorf et al., 2019). Let $\{Y(t), t \geqslant 0\}$ be the stochastic process and $\{y_t, t = 1, 2, \ldots T\}$ denote the realization, where $Y(t) = y_t$ for all $t \in \mathbb{N}$.

Let us denote the parameter vector characterizing (1)-(4) by $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots \theta_m,\ \phi_1, \ldots, \phi_p, \alpha_0, \alpha_1)$. For instance, in modeling $\mu(t)$, m will depend on the chosen function from Table 2; $m = 2$ for the Gaussian function and $m = 3$ for the Rational function and p is the order of the autoregressive model considered for $Z(t)$. Denote the data set by $D_T = (y_1, \ldots, y_T)$.

The likelihood function of the parameters given the data is given by

$$\mathcal{L}(\boldsymbol{\theta}|D_T) = \prod_{t=1}^{T} f(y_t|y_{t-1}, \ldots, y_1) f(y_1),$$

where the conditional probability density function (pdf) of y_t given its history is given by (Tsay (2010))

$$f(y_t|y_{t-1}, \ldots, y_1) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} \exp\left\{-\frac{(y_t - \mu(t))^2}{2\sigma^2(t)}\right\}.$$

The prior specification for the parameter vector $\boldsymbol{\theta}$ is as follows. We assume for the parameter vector $\boldsymbol{\theta}^* = (\theta_0, \theta_1, \dots \theta_m,\ \phi_1, \dots, \phi_p)$ a normal prior distribution $N(\boldsymbol{a}, b^2\boldsymbol{I})$, whose prior density can be written as

$$\pi(\boldsymbol{\theta}^*) = (2\pi b)^{-(m+p)/2} \exp\left\{-\frac{1}{2b^2}(\boldsymbol{\theta}^* - \boldsymbol{a})'(\boldsymbol{\theta}^* - \boldsymbol{a})\right\},$$

where $\boldsymbol{I}$ is an identity matrix of dimension $(m+p)\times(m+p)$. Values for the $\boldsymbol{a}$ and b hyperparameters may be assigned, as we do, to have approximately non-informative prior (we considered, $a = 0$ and $b = 1$). For parameter $\alpha_0 > 0$ we assume a Log-Normal prior $LN(a_0, b_0)$, and for parameter $\alpha_1 \in (0, 1)$ we assume a Normal(0, 1) prior for parameter $\xi = logit(\alpha_1)$. Assuming priori independence for the parameters, we denote by $\pi(\boldsymbol{\theta})$ the joint prior density function for the parameter vector $\boldsymbol{\theta}$. From Bayes' rule, the posterior density function for the vector of parameters $\boldsymbol{\theta}$ is then given by

$$\pi(\boldsymbol{\theta}|D_T) \propto \mathcal{L}(\boldsymbol{\theta}|D_T)\pi(\boldsymbol{\theta}).$$

The nonlinearity of the mean and the conditional heteroscedasticity of the process precludes a closed form for the posterior density of $\boldsymbol{\theta}$. Bayesian estimation of $\boldsymbol{\theta}$ must be implemented numerically using Markov chain Monte Carlo (MCMC) methods. Although MCMC-based algorithms are time-consuming and slow to converge in complex modeling situations, the modeling considered in this paper were computationally feasible and convergence as monitored by the Geweke criterion (Geweke, 1992) was reached in less than 1 minute for each model using a core $i7$. However, the computational effort grows rapidly when the order p of the autoregressive model increases.

We ran the analysis by modeling $\mu(t)$ by each function in Table 2. We used the Gibbs sampler with the Metropolis-Hastings algorithm (Chib and Greenberg, 1995). Specifically, we generated a chain of size 50,000, deleting a initial burn-in sample of 30% of the generated chain and resampled a value once every 10 generated values (thinning). This resulted in a final sample of size $G = 3500$, which is considered a random sample from the posterior distribution of $\boldsymbol{\theta}$. We obtained such results corresponding to each function for $\mu(t)$. To select the best fitting AR(p) model, we used the BIC criterion (Schwarz, 1978), and to select the best fitting function $\mu(t)$, we used the conditional predictive ordinate (CPO) (Gelfand et al., 1992), which is briefly described below.

3.2 The Conditional Predictive Ordinate

Use of the CPO provides a useful cross-validation approach and is a computationally efficient measure of model fit. This measure is calculated using the predictive probability density function of observing y_t in the future after having observed D_{t-1} given by

$$f(y_t|D_{t-1}) = \int_{\boldsymbol{\Theta}} f(y_t|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D_{t-1})d\boldsymbol{\theta}, \tag{6}$$

where $\boldsymbol{\Theta}$ is the parametric space.

Considering the equation (6) and the generated chain from the MCMC procedure, we can estimate the CPO$_t$ by

$$\widehat{\text{CPO}}_t = \left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{f(y_t|\boldsymbol{\theta}^{(g)})}\right]^{-1}. \tag{7}$$

In many situations, it is more convenient for numerical reasons, to calculate $\log(\mathrm{CPO}_t)$, rather than CPO_t. In this case, $\log(\widehat{\mathrm{CPO}})$ is given by the sum of the estimates $\log(\widehat{\mathrm{CPO}}_t)$. The fitted model with the smallest $-\log(\widehat{\mathrm{CPO}})$ is selected as the model that best fits the data.

3.3 Influential Points

Let $\boldsymbol{y_{(-t)}} = (y_1, y_2, \ldots, y_{t-1}, y_{t+1}, \ldots, y_T)$ the observation vector after removal of the $t-th$ observation of D_T. We denote the joint posterior densities of the parameter vector $\boldsymbol{\theta}$, obtained from the original data set by $\pi(\boldsymbol{\theta}|D_T)$ and from the data set after removal of the $t-th$ observation by $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{(-t)})$. The influence of the observation y_t can be evaluated by calculating the Kullback-Leibler (K-L) divergence between these two posterior densities. Specifically,

$$\mathrm{KL}(\pi, \pi_{(-t)}) = \int \pi(\boldsymbol{\theta}|D_T) \log\left(\frac{\pi(\boldsymbol{\theta}|D_T)}{\pi(\boldsymbol{\theta}|\boldsymbol{y}_{(-t)})}\right) d\boldsymbol{\theta}. \tag{8}$$

It can be shown that (8) can be expressed as a posterior expectation (Conceição et al., 2013)

$$\mathrm{KL}(\pi, \pi_{(-t)}) = -\log(\mathrm{CPO}_t) + E_{\pi(\boldsymbol{\theta}|D_T)}\{\log\left(f(y_t|\boldsymbol{\theta})\right)\}, \tag{9}$$

where CPO_t is the conditional predictive ordinate (CPO) density of the observation y_t.

Given a sample $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(G)}\}$, generated from posterior density $\pi(\boldsymbol{\theta}|D_T)$, we can estimate the effect of observation y_t by

$$\widehat{\mathrm{KL}}(\pi, \pi_{(-t)}) = -\log(\widehat{\mathrm{CPO}}_t) + \frac{1}{G}\sum_{g=1}^{G} \log\left(f(y_t; \boldsymbol{\theta}^{(g)})\right),$$

where $\widehat{\mathrm{CPO}}_t$ is evaluated by Equation (7).

A measure of calibration for the $\mathrm{KL}(\pi, \pi_{(-t)})$ divergence proposed by McCulloch (1989), denoted by ρ_t, is derived from the solution of the equation $\mathrm{KL}(\pi, \pi_{(-t)}) = \mathrm{KL}(\mathrm{B}(0.5), \mathrm{B}(\rho_t)) = -\log 4\rho_t(1-\rho_t)/2$, where $\mathrm{B}(\rho_t)$ denotes the Bernoulli distribution with a success probability ρ_t. This implies describing results using the full posterior density, $\pi(\boldsymbol{\theta}|D_T)$, instead of the posterior density by removing the *t-th* observation, $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{(-t)})$, is equivalent to describing an event not seen as having probability ρ_t, when the correct probability is 0.5. Solving the equation for ρ_t, we have that $\rho_t = \frac{1}{2}\left\{1 + \sqrt{1 - \exp[-2\mathrm{KL}(\pi, \pi_{(-t)})]}\right\}$. This implies that $0.5 \leqslant \rho_t \leqslant 1$. For $\rho_t \gg 0.5$ it can be considered that the *t-th* observation is an influential point.

3.4 Forecasting

Forecasts of future values of the process $Y(T+h)$ for some lead time $h > 0$ made from origin T are obtained as the conditional expectations $E(Y(T+h)|D_T)$ of the predictive density $f(y_{T+h}|D_T)$, given by

$$f(y_{T+h}|D_T) = \int_{\boldsymbol{\Theta}} f(y_{T+h}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D_T)d\boldsymbol{\theta}, \tag{10}$$

where $\boldsymbol{\Theta}$ is the parametric space. Thus, the forecast for $Y(T+h)$, denoted here by $\widehat{Y}(T+h) = E(Y(T+h)|D_T)$, is given by

$$\widehat{Y}(T+h) = \int_0^{\infty} y_{T+h} f(y_{t+h}|D_T) dy_{T+h}. \tag{11}$$

Replacing the equation (6) in equation (11) and changing the order of integrals, we find an appropriate equation to calculate the predictions given by

$$\widehat{Y}(T+h) = \int_{\boldsymbol{\Theta}} E(Y(t+h)|\boldsymbol{\theta} D_T)\pi(\boldsymbol{\theta}|D_T)d\boldsymbol{\theta}.$$

Monte Carlo estimates for $\widehat{Y}(T+h)$ can be obtained considering the generated samples from the posterior density of the parameter vector $\boldsymbol{\theta}$. Let us start by considering the chains provided by the MCMC algorithm for the parameter vector $\boldsymbol{\theta}^{(g)},\ g = 1, \ldots, G$. Thus, we have the values $\mu^{(g)}(T+h), g = 1, \ldots, G$ evaluated when replacing each element of the vector $\boldsymbol{\theta}^{(g)}$ in the function $\mu(T+h)$. A Monte Carlo estimate for $\widehat{Y}(T+h)$ can be calculated by

$$\widehat{y}_{T+h} = \frac{1}{G}\sum_{g=1}^{G}\mu^{(g)}(T+h).$$

3.4.1 Forecasts in the Presence of Heteroscedasticity

The presence of conditional heteroscedasticity does not affect the calculation of the forecasts, since $E(\sigma(t)W(t)|D_t) = 0$ for all $t \geqslant 0$. However, when interest lies in calculating probabilities associated with $Y(T+h)$, the presence of heteroscedasticity requires that $\sigma^2(T+h)$ be predicted as well. Considering the proposed model given in (4) for $\sigma^2(t)$, we have that $\widehat{\sigma}^2(T+h) = E(\sigma^2(T+h)|D_T)$ is given by

$$\widehat{\sigma}^2(T+h) = \alpha_0\left(\sum_{j=0}^{h-1}\alpha_1^j\right) + \alpha_1^h\widehat{Z}^2(T), \tag{12}$$

where $\widehat{Z}(T) = Y(T) - \widehat{\mu}(T)$, $\alpha_0 > 0$ and $0 \leqslant \alpha_1 < 1$. However, since $0 \leqslant \alpha_1 < 1$, when $h \to \infty$, the equation (12) results in $\widehat{\sigma}^2(T+h) \to \alpha_0/(1-\alpha_1)$.

Monte Carlo estimates for $\widehat{\sigma}^2(T+h)$ can be calculated by considering the values $\widehat{\sigma}^{2(g)}(T+h), g = 1, \ldots, G$, evaluated by replacing each element of the vector $\boldsymbol{\alpha}^{(g)} = (\alpha_0^{(g)}, \alpha_1^{(g)})$ in equation (12). Thus, the Monte Carlo estimate $\widehat{\sigma}^2(T+h)$ is given by

$$\widehat{\sigma}^2_{T+h} = \frac{1}{G}\sum_{g=1}^{G}\widehat{\sigma}^{2(g)}(T+h).$$

The credible intervals (CI) for the forecasts $\widehat{y}_{T+h}$ with probability $1-\alpha$ can be calculated by $\widehat{y}_{T+h} \pm z_{\alpha/2}\widehat{\sigma}_{T+h}$, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of a normal $N(0,1)$ distribution. When the process is homoscedastic, the variance is constant for all $T+h$. However, when there is heteroscedasticity, it is necessary to express $\sigma^2(T+h)$ in terms of the parameters α_0 and α_1.

The probabilities associated with $Y(T+h)$, in the presence of heteroscedasticity, can be calculated by

$$P(Y(T+h) < U) = \Phi\left(\frac{U - \widehat{y}_{T+h}}{\widehat{\sigma}^2_{T+h}}\right).$$

In the next section, we present the results of the best model fitted and the forecast for 100 days ahead given the observed data. We mention that 100 days is a long forecasting horizon for these inherently short memory time series models, and forecasts closer to the forecast origin will be more reliable. All results were obtained using the Software (R Core Team, 2020).

Table 3: Natural logarithm of the conditional predictive ordinate (CPO(p)) for the functions $\mu(t)$ for the daily number of new COVID-19 deaths in six countries

Fuction	$-\log(\text{CPO}(p))$					
	China	USA	Spain	Italy	Brazil	India
Gaussian	254(1)	309(2)	337(3)	231(3)	197(3)	483(0)
Log-Normal	253(1)	307(2)	331(3)	232(1)	164(6)	479(0)
Alpha	263(1)	305(2)	330(0)	**214(0)**	**161(6)**	490(3)
Generalized Exp.	261(1)	307(2)	330(1)	224(1)	165(6)	479(0)
Rational function	260(1)	**304(1)**	**328(3)**	230(1)	188(3)	483(0)
Logistic	**247(1)**	310(2)	335(3)	230(3)	205(2)	486(0)
Gamma	252(1)	309(2)	333(3)	225(3)	167(6)	479(0)

4 COVID-19 Results for the Six Countries

In Section 2, we described count time series of daily deaths by COVID-19 for each of the six countries summarized in Table 1. The observations available until July 31 (sample size) were used for model fitting. We scale the population of each country in order to better handle the data numerically and graphically. For China and India, we denote the data in multiples of 100 million. Corresponding to the total population of each of these countries (of about 1.411 billion and 1.395 billion respectively), we get a scale factor of 14.11 for China and 13.95 for India. For the remaining four countries, we denote the data as multiples of one million, resulting in scale factors of 333.76 for USA, 45.68 for Spain, 60.02 for Italy, and 217.26 for Brazil, based on each of their total populations. We fit each of the models presented in Table 2 under the Bayesian approach described in Section 3.1. Table 3 shows the CPO(p) for each model, where p refers to the order of the autoregressive model. In this table, it is possible to see that the Alpha model was selected for two of the six data sets, the Rational function model was selected for two of the data sets, and the Logistic Model was selected for data set from China. For India, we have three models as the same criterion value, and we can use any of these models to analyze the data set from India. This is due to the fact that India is still in the first part of the pandemic and has not yet reached its peak.

As we discussed in the introduction, the main objectives in fitting these models to the data are (a) to understand patterns in the mean curves, and (b) to predict future values of the time series. Therefore, our aim is to use the fitted models for the average number of daily deaths to identify the state of the pandemic and to use the fitted curve to evaluate whether the peak days of the epidemic are approaching, are occurring, or have passed. Before analyzing each country's data, it is essential to understand that some factors that can increase the forecast error are not considered in the forecasting model for the number of deaths for any given day. One of the main factors in this list is data consolidation. As mentioned in Section 1, these factors can lead to outliers and heteroscedasticity, which causes high prediction errors. However, the trajectory of the average curve and the identification of peak days is more critical in our analysis than accurate forecasting of the number of deaths on any given day. Aiming to achieve this objective with proposed models, we used a K-L divergence calibration measure to identify outliers, which we then smoothed suitably.

The fitted models were used to forecast the pattern of the mean number of deaths curve together with its 95% credible interval, for the next 100 days beyond the observed data. We present these results for the six countries. The last 20 days with observations from August 1 to August 20 were held out from the model fitting and were used to evaluate the accuracy of forecasts for these days. We present these results in Subsection 4.7.

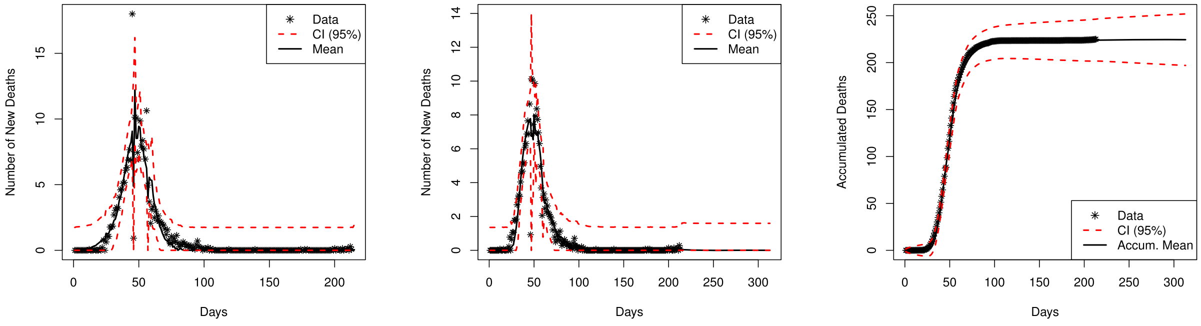
Tables in the supplementary material present the posterior summaries and the corresponding 95% credible intervals for the parameters of the selected models fitted after applying a smoothing to account for outliers identified by the K-L divergence calibration measure. The residual analysis, including the Lagrange Multiplier test for the ARCH model for $Z(t)$, is also presented in the supplementary material.

4.1 Results for China

We considered the time series of new daily deaths notified between December 31, 2019 and July 31, 2020 in China. According to the CPO criterion shown in Table 3, the Logistic with AR(1) model was selected as providing the best fit to the data. It is shown in Figure 2a that on February 12 ($t = 45$) the daily added number of deaths had clear jumps with significantly large sizes. Such sizable jumps cannot happen within one day, rather they represent an accumulation of deaths that have not been reported on previous dates prior to February 12 (Lili Wang et al., 2020). In our analysis, we considered the K-L divergence (see Subsection 3.3) to identify outliers. We chose to do two analyses, i.e., (a) to keep these outliers in the data set, or (b) to replace these outliers by the central moving average of the seven values observed with the outlier values at the center.

Proceeding with the analysis (a), Figure 2a presents the time series of daily death counts, the fits under the Logistic model for $\mu(t)$, and the 95% credible interval. Considering the K-L divergence as a measure of calibration, we accept as an influential point the values whose $\rho_i > 0.8$. With this criterion, values that occurred on February 12 and February 23 ($t = 45$ and $t = 56$) respectively, were considered to be outliers. Therefore, for these days, it is not expected that the model will be able to predict the observations $y(45) = 254$ and $y(56) = 150$ well. Proceeding with the analysis (b), we considered the time series of new daily deaths, replacing the outliers by the central moving average. According to the CPO criterion, the Alpha with AR(3) model was selected as providing the best fit to the data ($-\text{LogCPO}(3) = 183$). Figure 2b presents the time series of daily deaths, the fits under the Alpha with AR(3) model for $\mu(t)$, and the 95% credible interval. In Figure 2c, we present the accumulated number of deaths and the 95% credible interval. We observe from the plots in Figure 2 that the Alpha with AR(3) model gives a good fit for this time series. In addition, Figure 2 also shows the forecast for the next 100 days (from July 31) and the credible intervals around the forecasts.

Using the fitted Alpha with AR(3) model, we can estimate τ_{peak}, the day with the maximum number of deaths as $t_{peak} = 47$, and the expected number of deaths $\widehat{\mu}(47) = 11.4$ (or 161 people, by multiplying by the scale factor 14.11). The 95% credible interval for this day was (7.9, 14.9) (or, (111, 210) people). Specifically for this data set, these estimated values cannot be compared with the values that were observed, i.e., $t_{peak} = 45$ and $y(45) = 17.9$ (or 253 people) because February 12 was an outlier. Proceeding with the analysis (b), the observed values were $t_{peak} = 47$ and $y(47) = 10.1$ (or 143 people); the predicted values using this fitted model were $\widehat{\mu}(47) = 6.7$ (or 95 people) and the 95% credible interval for this day was (6.3, 13.9) (or, (89,196) people). We note that the number of deaths observed on this peak day is included in the estimated credible interval.



(a) (b) (c)

Figure 2: China model: (a) mean of the Logistic with AR(1) model; (b) mean of the Alpha with AR(3) model; (c) accumulated number of deaths by COVID-19 in China from December 31, 2019 to July 31, 2020, and forecasts for the next 100 days.

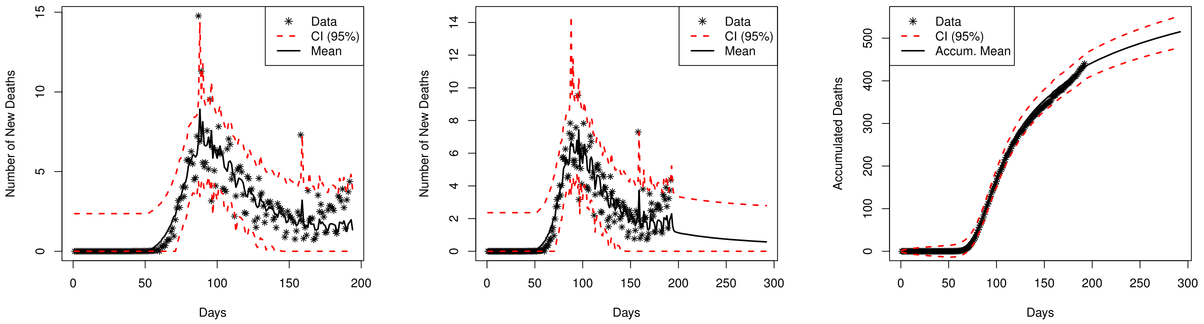
4.2 Results for USA

We considered the time series of daily deaths notified by COVID-19 between January 21, 2020 until July 31, 2020 in USA. Table 3 shows the CPO for each fitted model, and the Rational function with AR(1) model was selected as the best fitting model. Figure 3a presents the time series of new daily deaths, the fitted mean $\widehat{\mu}(t)$ with the Rational function model, and the 95% credible interval. We can observe from Figure 3a that there are some outliers. Based on a K-L measure of $\rho_i > 0.8$, we identified observations at $t = 87$ and $t = 89$ as outliers, with $y(87) = 14.7$ (4906 people) and $y(89) = 11.3$ (3771 people). Replacing the outliers by the central moving average and refitting all the models, the newly fitted Rational function with AR(1) model was selected as providing the best fit to the data according to the CPO criterion ($-\text{LogCPO}(1) = 261$). Figure 3b presents the time series of daily deaths, the fits under the newly fitted model for $\mu(t)$, and the 95% credible interval. In Figure 3c, we present the accumulated number of deaths and the 95% credible interval. Figure 3 shows the goodness of fit of the Rational function model for this data set. In addition, Figure 3 shows the forecasts for the next 100 days (from July 31) for the number of new daily deaths and the cumulative number of deaths.

For the US, the observed peak day was $t_{peak} = 95$ and $y(95) = 9.5$ (or 3171 people, multiplying by 333.76). For this peak day, the forecast provided by the fitted model was $\widehat{\mu}(95) = 5.9$ (or 1969 people) with 95% credible interval (4.5, 10.6) (or (1502, 3538) people). We note that the number of deaths observed on the peak day is included in the estimated credible interval.

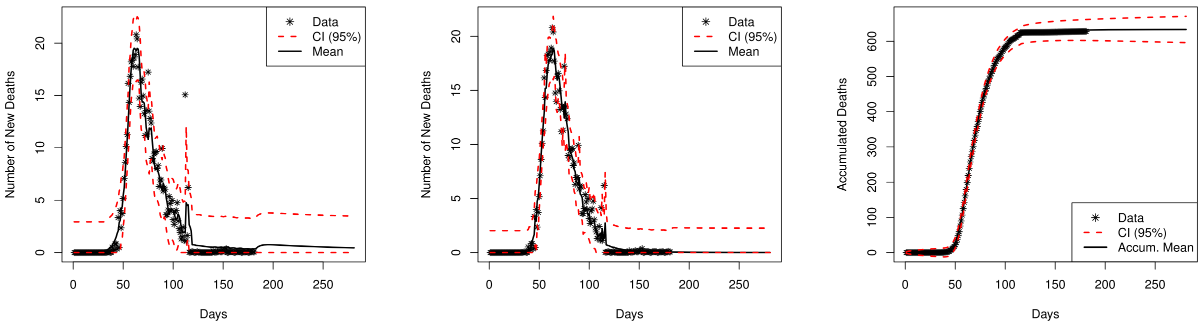
4.3 Results for Spain

For Spain, we consider the time series of new daily deaths notified between February 1, 2020 until July 31, 2020. Based on the CPO criterion shown in Table 3, the Rational function with AR(3) model was selected as the best model for this data. Figure 4a presents the time series of new daily deaths, the fitted mean by the Rational function with AR(3) model $\widehat{\mu}(t)$, and the 95% credible interval. The measure of calibration for the K-L divergence identified as outliers (with $\rho_i > 0.8$) the observations at $t = 112$, and $y(112) = 15.1$ (690 people). Replacing the outliers by the central moving average and fitting all the models again, according to the CPO criterion, the



(a) (b) (c)

Figure 3: USA model: (a) mean of the Rational function with AR(1) model; (b) mean Rational function with AR(1) model without outliers; (c) accumulated number of deaths by COVID-19 in USA from January 21, 2020 to July 31, 2020 and forecasts for the next 100 days.

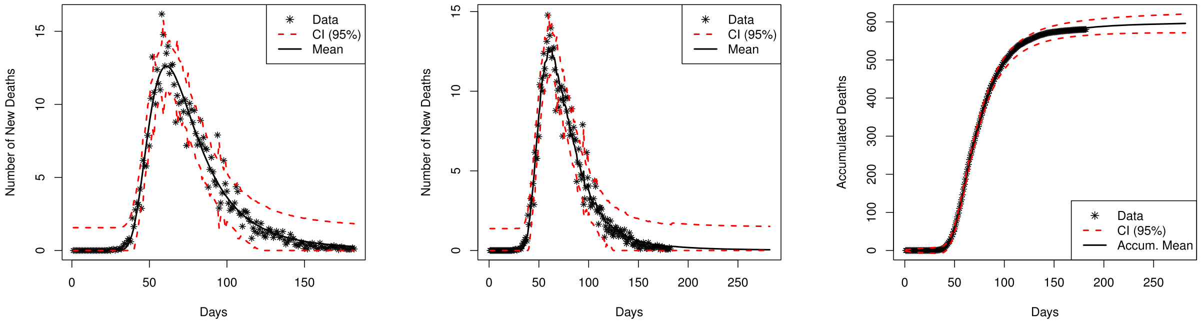


(a) (b) (c)

Figure 4: Spain model: (a) mean of Rational function with AR(3) model; (b) mean of the Alpha with AR(1) model; (c) accumulated number of deaths by COVID-19 in Spain from February 1, 2020 to July 31, 2020 and forecasts for the next 100 days.

Alpha with AR(1) model was selected as providing the best fit to the data ($-\text{LogCPO}(1) = 254$). Figure 4b presents the time series of daily death counts, the fits under the new fitted model for $\mu(t)$, and the 95% credible interval. In Figure 4c, we present the accumulated number of deaths and the 95% credible interval. We can observe from Figure 4 the goodness of fit of the Alpha with AR(1) model for the asymmetric data set. These plots also show the forecasts for the next 100 days (from July 31 as origin) for the number of new daily deaths and the cumulative number of deaths.

The day with the maximum number of deaths observed was $t_{peak} = 63$ and $y(63) = 20.8$ (or 950 people, multiplying by the scale factor of 45.68). The estimate of the number of deaths given by the fitted model with their credible interval on this day is given by $\widehat{\mu}(63) = 18.3$ (or 836 people), with 95% credible interval (15.9, 21.8) (or (726, 996) people). We note that the number of deaths observed on that peak day is included in the estimated credible interval.



(a) (b) (c)

Figure 5: Italy model: (a) mean of the Alpha with AR(0) model; (b) mean of the Alpha with AR(1) model; (c) accumulated number of deaths by COVID-19 in Italy from January 31, 2020 until July 31, 2020 and forecasts for the next 100 days.

4.4 Results for Italy

For Italy, we model the time series of new daily number of notified deaths of COVID-19, beginning on January 31 and ending on July 31, 2020. Table 3 presents the LogCPO for each assumed model, from which we see that the Alpha without autoregression model ($p = 0$) was initially selected as the best fit to the data. Figure 5a presents the time series of new daily death counts, the fitted mean $\widehat{\mu}(t)$, and the 95% credible interval. The discrepant point analysis indicated points $y(52) = 13.2$ (or 792 people) and $y(58) = 16.2$ (or 972 people) as outliers. By replacing these points with their respective seven-point centered moving averages, the best fitting new model was the Alpha with AR(1) model ($-\text{LogCPO}(1) = 193$). Figure 5b presents the time series of new daily number of deaths, the fitted mean by the Alpha with AR(1) model $\widehat{\mu}(t)$, and the 95% credible interval. In Figure 5c, we show the accumulated number of deaths and the 95% credible interval. We can observe from Figure 5 that the Alpha with AR(1) model provides a good fit to these data. Figure 5 also shows the forecasts for the next 100 days (from July 31 as origin) for the number of new daily deaths and the cumulative number of deaths.

For Italy, the day with the maximum number of deaths was $t_{peak} = 58$ and $y(58) = 16.2$ (or 972 people, multiplying by the scale factor of 60.02), and this point was indicated as an outlier and replaced by $y(58) = 688$. The new peak day was $t_{peak} = 59$ with $y(59) = 14.8$ (or 888 people). The estimated of the number of deaths from the Alpha with AR(1) model was $\widehat{\mu}(59) = 12.4$ (or 744 people), with a 95% credible interval of (10.6, 14.8) (or (636, 888) people).

4.5 Results for Brazil

In Brazil, the pandemic only started after the other countries that we analyzed, so we consider the series of mortality beginning on February 26, when the first case was notified until July 31, 2020. According to the CPO criterion, shown in Table 3, the Alpha with AR(6) model was selected as providing the best fit to the data and the K-L divergence did not not identify any outliers. A peak cannot be identified for Brazil because, as of June 10, the curve in the number of new deaths per day oscillates at a level with an average of 1000 deaths per day, showing a flat line instead of a negative slope. This pattern of the curve is possibly due to political

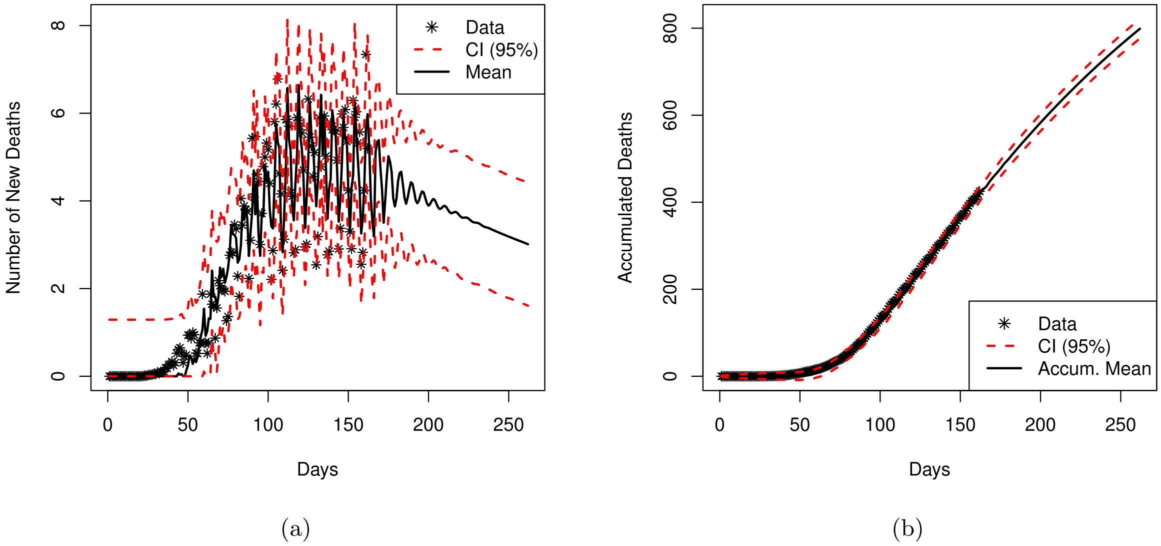


Figure 6: Brazil model: (a) mean of the Alpha with AR(6) model; (b) accumulated number of deaths by COVID-19 in Brazil from February 26, 2020 until July 31, 2020 and forecasts for the next 100 days.

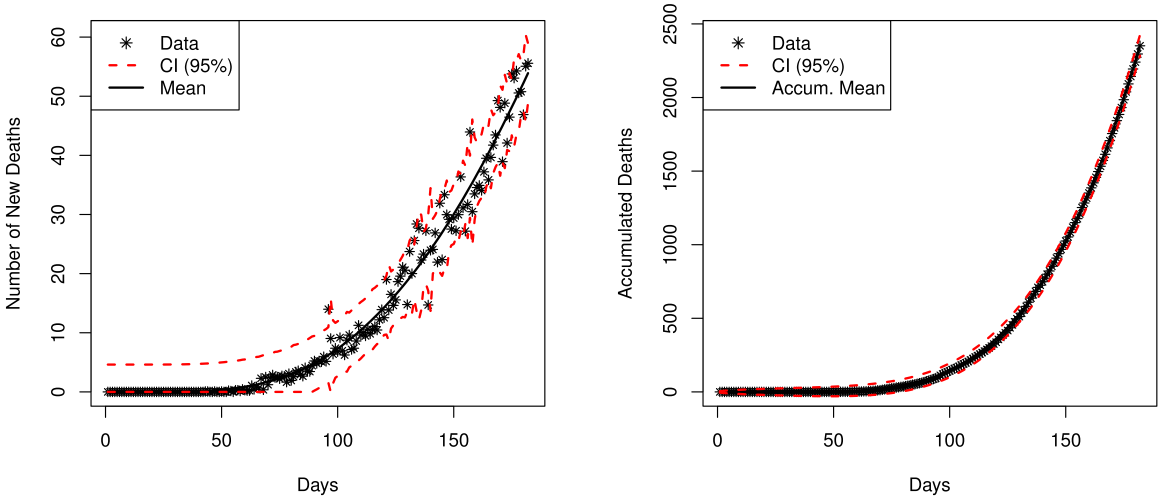
conflict in decision making regarding restrictive measures to contain the spread of the virus (see additional details in the supplementary material). The maximum number of deaths observed in the Brazilian data set was $y(161) = 7.3$ (or 1586 people, multiplying by the scale factor of 217.26); for that day the Alpha with AR(6) model provided a prediction of $\widehat{\mu}(161) = 5.7$ (or 1238 people) with a 95% credible interval of (5.2, 8.1) (or (1130, 1760) people).

Figure 6a presents the time series of new daily deaths, the fitted mean $\widehat{\mu}(t)$ and the 95% credible interval. Figure 6b shows the accumulated number of deaths and the 95% credible interval. We can observe in Figure 6a that the fitted Alpha unimodal model shows a smooth decay. Figure 6 also shows the forecasts for the next 100 days (from July 31) for the number of deaths and the cumulative number of deaths in Brazil.

4.6 Results for India

For India, we consider the time series of new daily deaths notified between January 3, 2020, and July 3, 2020. However, although the first notification occurred on January 31, it was on March 5 that India's pandemic started, with 22 cases reported. The first death was reported on March 13.

According to the CPO criterion, shown in Table 3, three models can be chosen. However, the discrepant point analysis indicated points $y(175) = 80.9$ (or 1128 people, multiplying by the scale factor of 13.95) as outlier. After replacing this outlier value by the value calculated using its moving average centered, and fit all the models again, the Log-Normal without autoregressive model ($p = 0$) presented the highest value of the criterion ($-\text{LogCPO}(0) = 403$) and it was selected as providing the best fit to the data. The fact that the peak has not yet occurred implies that a possible asymmetry is not yet evident. Hence while the data support the Log-



(a)

(b)

Figure 7: India model: (a) mean of the Log-Normal model; (b) accumulated number of deaths by COVID-19 in India from January 31, 2020 until July 31, 2020.

Normal model now, it may not be suitable for the entire series. However, the Log-Normal model currently meets the objectives of the analysis, which is to provide a forecast of the short-term trend for the pandemic data.

Figure 7a presents the time series of daily new deaths, the fitted mean with Log-Normal model $\widehat{\mu}(t)$, and the 95% credible interval. Figure 7b shows the accumulated number of deaths and the 95% credible interval. We can observe at the fitted mean model shown in Figure 7 the curve with an increasing curvature indicating that observation are before the peak. These plots also show the cumulative number of deaths.

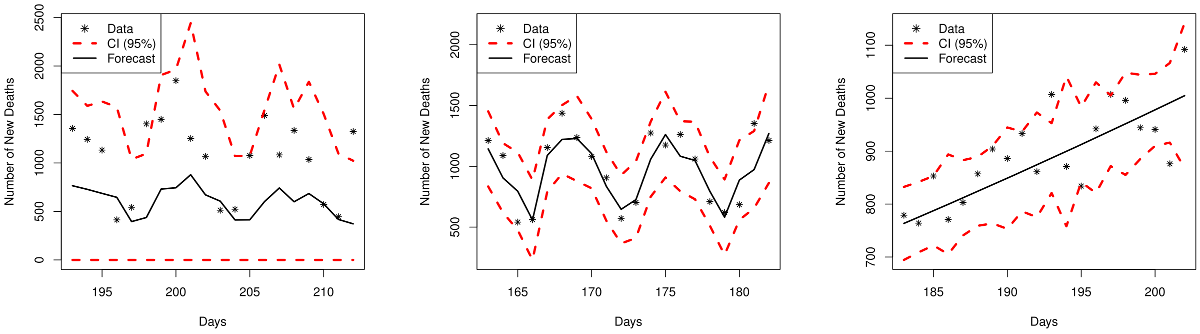
4.7 Forecast results

Any pandemic trend assessment, reduction, leveling off, or growth, must be persistent for at least 14 days (since 14 days is the virus incubation time). Therefore, if our models are used to forecast this trend for 20 days, they can shed light on whether the country is leaving the peak or whether it has not even reached the peak yet. In this subsection, we present the use of fitted models to obtain 20-day ahead forecasts.

The accuracy of the predictions made with the fitted models can be measured considering the 20 days observed from August 1 to August 20, held out from model fitting in order to do forecast evaluation. The forecast errors for these days were computed as the difference between the observed counts and the corresponding forecasts. The forecast evaluation criteria used were the square root of the Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). The Mean Absolute Percent Error (MAPE) and the mean arctangent absolute percentage error (MAAPE) (Kim and Kim, 2016) were calculated for countries where there is no zero observation between August 1 and August 20 (USA, Brazil, India) due to MAPE's deficiency in evaluating errors when the observation is small. For these 20 days of forecast, we also present the minimum,

Table 4: Forecasting evaluation criteria for six countries

Country	Forecast Error				Observations			Forecast		
	RMSE	MAE	MAPE(%)	MAAPE(%)	Min.	Mean	Max.	Min.	Mean	Max.
China	1.99	1.57	–	–	0	2	6	0	1	20
USA	582.86	483.20	40.68	37.59	413	1055	1848	0	606	2442
Spain	11.30	7.63	–	–	0	8	29	0	5	24
Italy	6.70	6.04	–	–	2	6	13	0	12	15
Brazil	149.61	112.21	11.89	11.62	541	992	1437	230	960	1271
India	58.71	48.45	5.28	5.28	764	896	1092	694	881	1142



(a)

(b)

(c)

Figure 8: Number of new daily deaths observed and predicted from August 1 to August 20: (a) USA; (b) Brazil; (c) India.

average and maximum values of the number of observed deaths predicted by the model selected for each country. These values and evaluation criteria are shown in Table 4.

Figure 8 presents the observed daily new deaths for 20 days along with the forecasts and 95% credible intervals for USA, Brazil and India. We can see in these graphs the trends in the number of daily deaths in each country.

Two other points that we want to draw attention to are: On May 22, the Spanish Ministry of Health reported an additional 56 deaths in the last 24 hours. However, the Catalan authorities reported an additional 632 deaths that had previously occurred and were not assigned to dates, meaning that the cumulative total increased by 688 deaths. In this work, we considered on May 22, just 56 deaths. On May 25, the government decreased the number of total cases by -372 and the number of deaths to -1918. The discrepancy is the result of the validation of the same data by autonomous communities and the transition to a new surveillance strategy. We calculated our forecasting disregarding these negative numbers.

Another interesting forecasting result that we can obtain from the fitted models is the "end date" of the pandemic. Estimating the "end date" is not trivial and warrants a few different considerations. Theoretically, we can define the "end date" as the day with a high percentage of the total deaths predicted in the pandemic life cycle curve. Our forecasts consider two alternatives for estimating "end dates":

i. The date when 99% of expected of the total deaths have already occurred;

Table 5: Forecast the “end dates” of the pandemic

Countries	Start date	End date (97%)	End date (99%)
China	Dec 31, 2019	89 (Mar 29)	114 (Apr 23)
USA	Jan 21, 2020	360 (Dec 16)	564 (Aug 10)*
Spain	Feb 1, 2020	181 (Jul 31)	207 (Aug 27)
Italy	Jan 31, 2020	182 (Jul 31)	271 (Oct 29)
Brazil	Feb 26, 2020	392 (Mar 18)*	483 (Jun 19)*
India	Jan 31, 2020	–	–

(∗) date for year 2021; −− Not evaluated by model

ii. The date when 97% of expected of the total deaths have already occurred.

Table 5 reports the two alternative estimations of COVID-19 “end dates” for countries studied in this paper.

It is worth mentioning that China announced a relaxation of the quarantine and a partial reopening on March 30. For Brazil, these “end dates” are inaccurate because the country may be at a peak on June 10, or may not yet have reached its peak. However, Brazil released a reopening plan for June 1, at the height of the pandemic.

5 Final Comments

The COVID-19 pandemic is a huge public health problem. Three elements have been shown to be fundamental to differentiate the response to the crisis caused by the coronavirus: i) public attitudes of the authorities to the uncertainties brought about by the new virus; ii) the ability of the government to implement preventive measures; and iii) the structure and competence of public health systems to serve patients. In the paper in the supplementary material, we briefly report information on these attitudes for each of the countries we consider.

In this paper, we have described some nonlinear functions for modeling the time series of the number of daily deaths caused by the COVID-19 pandemic. Applications of the proposed models are shown using time series of daily deaths from six countries.

We note that it is essential that the model considered for the mean of the time series reproduces the asymmetric pattern of these series. Another critical aspect to be included in the model is the conditional heteroscedasticity of the time series. A good model for the variance is fundamental to calculate probabilities and more accurate credible intervals for the forecasts, and the ARCH(1) model that we used seems to do well.

A Bayesian approach has several advantages for estimation and forecasting in such situations, and can be an excellent tool. We use Markov Chain Monte Carlo procedures for carrying out Bayesian inference. The need for estimating model parameters with high precision led to some difficulties in choosing a specific nucleus to generate candidates for each parameter. A random walk with a Gaussian kernel whose variance was controlled for each model based on the rejection rate of the generated samples, which is always between 35% and 60%, seemed adequate.

To speak of a single peak for very large countries such as the USA, India, and Brazil is similar to speaking of a single peak for Europe. The states or regions in these countries have different peaks. But a global peak can be assessed and used to predict trends. An essential and useful take-away from our analysis is that the fitted models allow the prediction of the expected number of deaths on the day of the peak of the pandemic. However, we noticed that the variability of the series grows very close to this peak, and, usually, the model predicts a credible interval which includes the observed number of deaths.

These models can be used to forecast just 20 days ahead in order to inform what the trend of the pandemic pattern is likely to be. Further, as expected, forecasts will also change with the addition of new data. Also, the selected models will need to be fitted again in the future (perhaps once every 10 days). Nevertheless, our models come quite close to the actual counts in most cases and give reasonable estimates that can be of assistance to health authorities in decision making regarding understanding the progress of the pandemic.

Supplementary Material

Supplementary material online include: rational functions and nonlinear rational polynomial model; tables with fitted model parameters and residual analysis; a brief Report for each of the countries that we considered; data and R code needed to reproduce the results.

Acknowledgments

We are indebted to the editor and referees for their valuable comments, criticisms, and suggestions which have substantially improved the manuscript. Marinho G. Andrade was supported by the Brazilian organizations FAPESP (2019/21766-8); Jorge A. Achcar was supported by the Brazilian organization CNPq (301923/2019-1); Katiane S. Conceição was supported by the Brazilian organization FAPESP (2019/22412-5).

References

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, 26(4): 450–452.

Bates DM, Watts DG (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *The Lancet*, 395(10223): 514–523.

Chen D, Xu W, Lei Z, Huang Z, Liu J, Gao Z, et al. (2020a). Recurrence of positive SARS-CoV-2 RNA in COVID-19: A case report. *International Journal of Infectious Diseases*, 93: 297–299.

Chen J (2020). Pathogenicity and transmissibility of 2019-nCoV: A quick overview and comparison with other emerging viruses. *Microbes and Infection*, 22: 69–71.

Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. (2020b). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *The Lancet*, 395(10223): 507–513.

Chib S, Greenberg E (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4): 327–335.

Conceição KS, Andrade MG, Louzada F (2013). Zero-modified Poisson model: Bayesian approach, influence diagnostics, and an application to a Brazilian leptospirosis notification data. *Biometrical Journal*, 55(5): 661–678.

Desta F, Mac Siurtain MP, Colbert JJ (1999). Parameter estimation of nonlinear growth models in forestry. *Silva Fennica*, 33(4): 327–336.

Dorndorf A, Kargoll B, Paffenholz JA, Alkhatib H (2019). A bayesian nonlinear regression model based on t-distributed errors. In: *IX Hotine0-Marussi Symposium on Mathematical Geodesy* (P Novák, M Crespi, N Sneeuw, F Sansò, eds.), 127–135. Springer, Berlin.

Engle RF (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4): 987–1007.

Engle RF (1983). Estimates of the variance of US inflation based upon the ARCH model. *Journal of Money, Credit and Banking*, 15(3): 286–301.

Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491): 1–7.

Gelfand AE, Dey DK, Chang H (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). *Technical Report 462*, Department of Statistics, Stanford University, Stanford, California.

Geweke J (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*, 4: 641–649.

Girardi P, Greco L, Mameli V, Musio M, Racugno W, Ruli E, et al. (2020). Robust inference for nonlinear regression models from the Tsallis score: Application to COVID-19 contagion in Italy. ArXiv preprint: https://doi.org/10.1002/sta4.309.

Gupta RD, Kundu D (2001). Generalized exponential distribution: Different method of estimations. *Journal of Statistical Computation and Simulation*, 69(4): 315–337.

Han Q, Lin Q, Jin S, You L (2020). Coronavirus 2019-nCoV: A brief perspective from the front line. *Journal of Infection*, 80(4): 373–377.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223): 497–506.

Ju B, Zhang Q, Ge X, Wang R, Yu J, Shan S, et al. (2020). Potent human neutralizing antibodies elicited by SARS-CoV-2 infection. BioRxiv preprint: https://doi.org/10.1101/2020.03.21.990770.

Kandel N, Chungong S, Omaar A, Xing J (2020). Health security capacities in the context of COVID-19 outbreak: An analysis of International Health Regulations annual report data from 182 countries. *The Lancet*, 395: 1047–1053.

Katz D, Azen S, Schumitzky A (1981). Bayesian approach to the analysis of nonlinear models: Implementation and evaluation. *Biometrics*, 37: 137–142.

Kermack WO, McKendrick AG (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772): 700–721.

Kim S, Kim H (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3): 669–679.

Kim YI, Kim SG, Kim SM, Kim EH, Park SJ, Yu KM, et al. (2020). Infection and rapid transmission of SARS-CoV-2 in ferrets. *Cell Host & Microbe*, 27(5): 704–709.

Kulikov VS (2001). Rational function. In: *Encyclopedia of Mathematics* (M Hazewinkel, ed.). http://encyclopediaofmath.org/index.php?title=Rational_function&oldid=48438.

Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): The epidemic and the challenges. *International Journal of Antimicrobial Agents*, 55(3): 1–9.

Lam TTY, Jia N, Zhang YW, Shum MHH, Jiang JF, Zhu HC, et al. (2020). Identifying SARS-

CoV-2-related coronaviruses in Malayan pangolins. *Nature*, 583: 282–285.

Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581: 215–220.

Li JY, You Z, Wang Q, Zhou ZJ, Qiu Y, Luo R, et al. (2020). The epidemic of 2019-novel-coronavirus (2019-nCoV) pneumonia and insights for emerging infectious diseases in the future. *Microbes and Infection*, 22(2): 80–85.

Lili Wang YZ, Jie He BZ, Wang F, Lu Tang MK, Barker D, Eisenberg MC, et al. (2020). An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China (with discussion). *Journal of Data Science*, 18(3): 409–432.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet*, 395(10224): 565–574.

Luo J (2020). Predictive monitoring of COVID-19. *White paper*, Singapore University of Technology.

Lupia T, Scabini S, Pinna SM, Di Perri G, De Rosa FG, Corcione S (2020). 2019-novel coronavirus outbreak: A new challenge. *Journal of Global Antimicrobial Resistance*, 21: 22–27.

McCulloch RE (1989). Local model influence. *Journal of the American Statistical Association*, 84(406): 473–478.

Petrescu E (2009). A statistical distribution useful in product life cycle modeling. *Management and Marketing*, 4(2): 165–170.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007). Cambridge University Press, New York.

Pung R, Chiew CJ, Young BE, Chin S, Chen MI, Clapham HE, et al. (2020). Investigation of three clusters of COVID-19 in Singapore: Implications for surveillance and response measures. *The Lancet*, 395(10229): 1039–1046.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ratkowsky DA (1983). *Nonlinear Regression Modelling: A Unified Practical Approach*. Marcel Dekker, New York.

Schwarz G (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464.

Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. (2020). Structural basis of receptor recognition by SARS-CoV-2. *Nature*, 581: 221–224.

Shen C, Wang Z, Zhao F, Yang Y, Li J, Yuan J, et al. (2020). Treatment of 5 critically ill patients with COVID-19 with convalescent plasma. *JAMA*, 323(16): 1582–1589.

Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R (2020). Covid-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*, 24: 91–98.

Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. (2020). World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). *International Journal of Surgery*, 76: 71–76.

Sonnino G (2020). Dynamics of the COVID-19 comparison between the theoretical predictions and the real data. ArXiv preprint: https://arxiv.org/abs/2003.13540.

Tsallis C, Tirnakli U (2020). Predicting COVID-19 peaks around the world. *Frontiers in Physics*, 8: 217.

Tsay RS (2010). *Analysis of Financial Time Series*. John Wiley & Sons, New Jersey, 3 edition.

Wan Y, Shang J, Graham R, Baric RS, Li F (2020). Receptor recognition by the novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS coronavirus. *Journal of virology*, 94(7): 1–9.

Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature*, 581(7809): 465–469.

World Health Organization (2020a). Coronavirus disease (COVID-19) pandemic. https://www.who.int/health-topics/coronavirus#tab=tab_1.

World Health Organization (2020b). Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

World Health Organization (2020c). Coronavirus disease (COVID-19): Situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

World Health Organization (2020d). Laboratory testing strategy recommendations for COVID-19: Interim guidance. https://www.who.int/publications/i/item/laboratory-testing-strategy-recommendations-for-covid-19-interim-guidance.

Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host & Microbe*, 27(3): 325–328.

Zhang X, Ma R, Wang L (2020). Predicting turning point, duration and attack rate of COVID-19 outbreaks in major western countries. *Chaos, Solitons & Fractals*, 92: 214–217.

Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*, 92: 214–217.