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.

J. Math. Anal. Appl. 514 (2022) 125202

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Correcting notification delay and forecasting of COVID-19 data

Alessandro J.Q. Sarnaglia, Bartolomeu Zamprogno, Fabio A. Fajardo Molinares, Luciana G. de Godoi *, Nátaly A. Jiménez Monroy

Laboratory of Statistics and Natural Computing - LECON, Statistics Department, UFES, Vitória, Brazil

ARTICLE INFO

Article history:
Received 31 August 2020
Available online 30 March 2021
Submitted by S.G. Krantz

Keywords:
COVID-19
Notification delay
Prediction
Overdispersion

ABSTRACT

Since the first official case of COVID-19 was reported, many researchers around the world have spent their time trying to understand the dynamics of the virus by modeling and predicting the number of infected and deaths. The rapid spread and highly contagiousness motivate the necessity of monitoring cases in real-time, aiming to keep control of the epidemic. As pointed out by [3], some pitfalls like limited infrastructure, laboratory confirmation and logistical problems may cause reporting delay, leading to distortions of the real dynamics of the confirmed cases and deaths. The aim of this study is to propose a suitable statistical methodology for modeling and forecasting daily deaths and reported cases of COVID-19, considering key features as overdispersion of data and correction of notification delay. Both, reporting delays and forecasting consider a Bayesian approach in which the daily deaths and the confirmed cases are modelled using the negative binomial (NB) distribution in order to accommodate the population heterogeneity. For the correction of notification delay, the mean number of occurrences regarding time t notified at time $t+j$ (mean delayed notifications) is associated to the temporal and the delay lag evolution of the notification process through a log link. With regard to daily forecasting, the functional form adopted for the number of deaths and reported cases of COVID-19 is related to the sigmoid growth equation. A variable regarding week days or days off was considered in order to account for possible reduction of the records due to the lower offer of tests on days off. To illustrate the methodology, we analyze data of deaths and infected cases of COVID-19 in Espírito Santo, Brazil. We also obtain long-term predictions.

1. Introduction

The new coronavirus (SARS-CoV-2) is contagious among humans and causes the COVID-19 disease. COVID-19 was firstly reported in December 2019 after the appearence of an unidentified pneumonia. In Brazil, the first confirmed case of infection by SARS-CoV-2 was reported by the Health Ministry on February 26, 2020. Subsequently, in March 2020, the World Health Organization (WHO) announced the COVID-19

* Corresponding author.
E-mail address: luciana.godoi@ufes.br (L.G. de Godoi).

https://doi.org/10.1016/j.jmaa.2021.125202
0022-247X/

as a pandemic because of the growing number of infected cases outside China, where the outbreak started. According to the reports of the panel at WHO (https://covid19.who.int/), in July 2020 there were more than 17 million cases of COVID-19, including near 700.000 deaths around the world, affecting more than 200 countries and territories. At the same month, official data in Brazil indicated approximately 2.5 million confirmed cases of COVID-19 and more than 90.000 deaths.

The severity of the COVID-19 range from mild to severe respiratory symptoms. Some researchers call attention to the existence of severe neurological complications [21]. Older people and people of any age with comorbidities (obesity, type 2 diabetes mellitus, serious heart conditions, etc.) present higher risks for severe illness, requiring hospitalization, intensive care and/or mechanical ventilation. The most serious cases of COVID-19 may lead to death.

As pointed out by [27], the virus has the potential to spread rapidly and infect a large fraction of the population, overwhelming health care systems. Given the rapid rate of spread, [27] suggest that a combination of control measures, including early and active surveillance, quarantine and especially strong social distancing efforts, is needed to slow down or stop the spread of the virus.

Unfortunately, the COVID-19 pandemic is evolving rapidly and is not only a medical emergency and public health tragedy, but it is also affecting economic activities. With no urgent actions, the socioeconomic effects could have wide implications for trade, travel, provision of aid, economic markets, supply chains and the daily lives of people living around the world [30]. As pointed out by [19], COVID-19 is a medical problem with immense societal consequences. The world's scientists need to come together to find the proper solution for controlling this pandemic event, manage its consequences, and prevent future recurrences of similar pandemics.

To prepare the health care system for COVID-19 patients, it is necessary to quickly identify cases and keep control of the epidemic. [3] discuss the difficulties in monitoring epidemics in real-time and indicate the reporting delay as a crucial issue because it distorts the relationship between the reported disease incidence and the true disease incidence. According to them, reporting delays may be due to laboratory confirmation, logistical problems, infrastructure difficulties, and so on.

Many institutions and research groups around the world are dedicated to modeling and prediction of the number of confirmed cases and deaths associated to COVID-19. Different methodologies have been considered for these purposes, as can be seen in [29], [20] and [22]. In [13], attention is drawn to the problem of collective dynamics in human populations in different scenarios, such as crowd disasters, crime, terrorism, war and disease spreading. The authors discuss the complexity to propose analytic and predictive models. Regarding global pandemics, [13] also present a history of the development of mathematical models in this context until nowadays, showing that, despite of challenges, complex science has produced major advances in modeling the dynamics of global epidemics and it includes quantitative, realistic, and even predictive models, bringing together statistical data analysis, modeling efforts, analytical approaches, and laboratory experiments. One of the most popular modeling strategies in this scenario is the use of compartmental models [6,28,15,26, e.g.], including the well-known SIR model and its extensions, such as the SEIR model [4,10] and the SIDARTHE model [11], among others. Basically, SIR-type models partition the population in "compartments" and define a system of nonlinear ordinary differential equations describing the transitions among these groups, which must be solved numerically. An improvement of the SIR model, including more realistic assumptions such as the effect of births and deaths due to other causes is suggested by [1]. The research of [12] shows that a SEIR model underestimates peak infection rates and substantially overestimates epidemic persistence after the peak has passed. The mathematical structure of SIR model and a discussion about the limitation of the method in the literature is described by [8].

Regarding stochastic models, [31] provide an estimate of the size of the epidemic in Wuhan on the basis of the number of cases exported from Wuhan to cities outside mainland China and forecast the extent of the domestic and global public health risks of epidemics, accounting for social and non-pharmaceutical prevention interventions. For this, they consider a stochastic modelling in terms of the SEIR model with the

basic reproductive number (R_0) being estimated using the Gibbs sampling and non-informative flat prior. The R_0 is defined by [7] as the average number of infectious contacts that an infected individual has before recovering and becoming immune (or dying). It is one of the most crucial quantities in infectious diseases and, as pointed out in [16], R_0 measures how contagious a disease is. For $R_0 < 1$, the disease is expected to stop spreading, but for $R_0 = 1$ an infected individual can infect on an average 1 person, that is, the spread of the disease is stable. The disease can spread and become epidemic if $R_0 > 1$. The nowcasting considered in [31] is related to the impact of the social distancing measures, use of face masks and improved personal hygiene and other in the transmissibility of the virus and not with the reporting delays as proposed by [3]. An extensive simulation of the epidemic forecasts for Wuhan and five other Chinese cities assuming that the transmissibility of SARS-CoV-2 was reduced by 0%, 25%, and 50% after Wuhan was quarantined on Jan 23, 2020 and with 0% and 50% mobility reduction inter-city was performed by [31].

In Brazil, [9] provides a web page and an app with daily updates of the number of infected people and deaths and also presents the short (1 to 2 weeks) and long term prediction for COVID-19. The statistical methodology considered by them is a hierarchical Bayesian model where the number of infected or deaths is modelled by a Poisson distribution with a time invariant non-linear predictor for the mean. A well known limitation of the Poisson distribution is the equidispersion, which intrinsically assumes that the mean and the variance of the response variable are equal. For many observed count data, it is common to identify overdispersion, which occurs when the sample variance is greater than the sample mean [14]. The simplest strategy to deal with overdispersion is to use the negative binomial regression and it is recommended when the extra variations presented on the data are caused by the heterogeneity of the population [5].

[7] show that the population heterogeneity can significantly impact the disease-induced immunity due to SARS-CoV-2 and argue that many SIR-type models assume a homogeneously mixing population in which all individuals are equally susceptible, and equally infectious if they become infected. The authors propose to accommodate this heterogeneity by categorizing the community into different age cohorts, with heterogeneous mixing between the different age cohorts, and their social active level.

From the previous discussion, the heterogeneity mentioned by [7] may induce overdispersion on COVID-19 data. In order to accommodate this phenomenon, we propose an extension of the model developed by [9], considering a negative binomial distribution instead of the Poisson. We have performed a reparameterization of the model in terms of more meaningfully quantities, allowing an easier prior elicitation. We have also incorporated other important features in the model. Specifically, we include an explanatory variable regarding week days and days off, in order to account for possible reduction of the records due to the lower offer of tests on days off, which, to the best of our knowledge, has not been considered in any mathematical or statistical analysis. We have also allowed time variation of the model parameters in order to account for the unstable nature of the pandemic.

We use data from Espírito Santo State in Brazil (ES/BR) to illustrate the proposed methodology. The purpose here is to predict the daily number of confirmed infections and deaths caused by COVID-19 for short and long term. Two main reasons have motivated us to analyze these data: (1) since 16/04/2020, the technical report of the non-governmental organization Open Knowledge Brasil (OKBR) identifies the state of ES/BR as one of the most transparent states in the dissemination of data regarding the COVID-19 in Brazil; (2) unlike most of the states in Brazil, the daily number of confirmed cases and deaths are aggregated at the date of occurrence (day of realization of the test or day of the death), not the date of notification, which is much more advisable to better reproduce the pandemic dynamics.

Despite the benefit of reason (2) aforementioned, it is worth to point out that even with a good transparency, the lack of reagents of molecular biology tests have caused delay in the laboratory confirmation of the COVID-19 in ES/BR (see [24,25,2]). This naturally causes updates of the numbers of previous days. Therefore, prior fitting the proposed model, in order to correct for delayed notifications, we extent the method in [3] by considering week days and days off and dropping the assumption of a delay window.

The remainder of this article is organized as follows: in Section 2, we present the methodology for correcting the notification delay and to predict the daily deaths and daily reported cases. In Section 3, we apply the methodology developed in Section 2 on COVID-19 dataset from Espírito Santo/Brazil. Finally, we make some concluding remarks in Section 4. The method proposed in this paper is implemented in R [23]. All codes are available with the authors upon request.

2. Methodology

2.1. Correcting notification delay

We are interested on the counts of some event at the time t, denoted by Y_t. In particular, we will apply the method in this section to the daily number of deaths and daily reported cases of COVID-19. These data naturally present a notification delay, so that Y_t is not truly known at time t and notifications occurred at t may be reported at instants $s \geq t$. In this paper, inspired by the study in [3], we will describe this behavior as follows. Let $Y_{t,s}$ be the total of occurrences at t notified until s, $s \geq t$. We assume

$$Y_{t,T+K} = \begin{cases} \sum_{k=1}^{K} Z_{t,T+k-t} + Y_{t,T}, & 1 \leq t \leq T; \\ \sum_{k=t-T+1}^{K} Z_{t,T+k-t} + Y_{t,t}, & T < t \leq T+K-1, \end{cases} \tag{1}$$

where $Z_{t,j}$ represents the number of occurrences regarding time t notified at time $t+j$ (delayed notifications). In this context, j will be referred to as the delay lag. The model in Equation (1) is particularly appealing in this case, since we do not have the whole evolution of the data, in particular, the data was provided only from T to $T+K$, such that, when $t \leq T$, it is only possible to obtain $Z_{t,j}$, for $j = T+1-t, \ldots, T+K-t$. For simplicity, we may write $T+K = N$.

Here, aiming to account for possible overdispersion, we assume that the delayed notifications $Z_{t,j}$ follow a Negative Binomial distribution with $\mathbb{E}(Z_{t,j}) = \lambda_{t,j}$ and $\mathbb{V}(Z_{t,j}) = \lambda_{t,j} + \frac{\lambda_{t,j}^2}{\phi}$, which will be denoted by $Z_{t,j} \sim \text{NB}(\lambda_{t,j}, \phi)$. The mean $\lambda_{t,j}$ satisfy

$$\log \lambda_{t,j} = \lambda + \alpha_t + \beta_j + \gamma_{t,j}, \tag{2}$$

where λ denotes the overall mean, α_t and β_j accommodate respectively the temporal and the delay lag evolution of the notification process and $\gamma_{t,j}$ allows for temporal changes in the delay lag effect. Equation (2) could be easily generalized to incorporate covariate effects. For simplicity, the parameter vector and the collection of observed notifications are represented by

$$\Phi = (\lambda, \alpha_1, \ldots, \alpha_{T+K}, \beta_0, \ldots, \beta_{T+K-1}, \gamma_{1,0}, \ldots, \gamma_{T+K,T+K-1}, \phi)$$

and

$$\mathcal{Z}_O = \{Z_{t,j}, \quad j = \max\{1, T+1-t\}, \ldots, T+K-t, \quad t = 1, \ldots, T+K\},$$

respectively. Fig. 1 shows an illustration of the data. Note that the set $j = \max\{1, T+1-t\}, \ldots, T+K-t$, $t = 1, \ldots, T+K$, may be rewritten as $t = \max\{1, T+1-j\}, \ldots, T+K-j$, $j = 1, \ldots, N-1$.

In order to correct the notification delay, we resort to the following Bayesian approach. Aiming to accommodate the unstable nature of α_t, β_j and $\gamma_{t,j}$, we assume the following evolution structure:

$$\begin{aligned} &\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, W_\alpha), \quad t = 2, \ldots, N, \\ &\beta_j | \beta_{j-1} \sim \mathcal{N}(\beta_{j-1}, W_\beta), \quad j = 1, \ldots, N-1, \end{aligned}$$

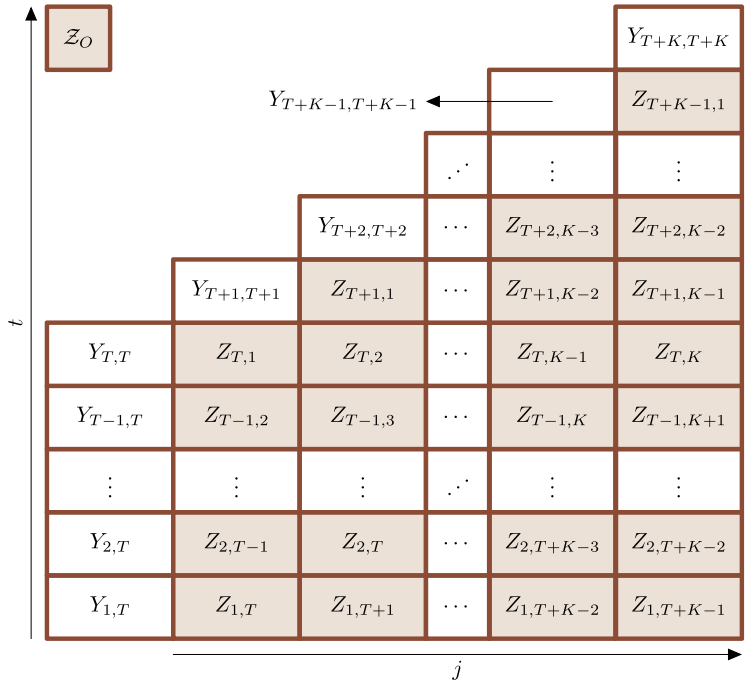


Fig. 1. Organization of delayed notification data. The $Z_{t,j}$ (colored rectangles) denote the number of occurrences regarding time t notified at time $t+j$. Note that the number of available delayed notifications reduces as t increases. The $Y_{t,s}$ is the total of occurrences at t reported until s, $s \geq t$ (see Equation (1)). For example, the total of occurrences at t updated in $T+K$ ($Y_{t,T+K}$) is given by the total of occurrences at t reported until T ($Y_{t,T}$) plus the delayed notifications $Z_{t,j}$, $j = \max\{1, T+1-t\}, \ldots, T+K-t$.

$$\gamma_{t,j}|\gamma_{t-1,j} \sim \mathcal{N}(\gamma_{t-1,j}, W_\gamma), \quad t = t_1^{(j)}, \ldots, N-j, \quad j = 1, \ldots, N-1,$$

where $t_1^{(j)} = \max\{1, T+1-j\}$ and we fix the variances as $W_\alpha = W_\beta = W_\gamma = W = 1/1600$ to ensure the parameters do not change more than 5% with probability 0.95. Fixing a correction window $L \geq 1$, for each $l = 1, \ldots, L$, from Equation (1), we observe that the (future) unobserved total $Y_{N-L+l,N+l}$ may be written as function of the observed total $Y_{N-L+l,N}$ and the unobserved delayed notifications $Z_{N-L+l,L-k}$, $k = 0, \ldots, l-1$. More precisely, we have

$$Y_{N-L+l,N+l} = Y_{N-L+l,N} + \sum_{k=0}^{l-1} Z_{N-L+l,L-k}, \quad l = 1, \ldots, L.$$

For simplicity, we define the collection

$$\mathcal{Z}_U = \{Z_{N-L+l,L-k}, \quad k = 0, \ldots, l-1, \quad l = 1, \ldots, L\}$$

of unobserved variables. One illustration of the $\mathcal{Z}_O$ and $\mathcal{Z}_U$ collections is provided in Fig. 2.

The delay correction is implemented by drawing samples from the posterior distribution of $\mathcal{Z}_U, \Phi|\mathcal{Z}_O$. Assuming independence of $\mathcal{Z}_U$ and $\mathcal{Z}_O$ (conditional on Φ), this sampling can be performed using Markov Chain Monte Carlo (MCMC) methods. Here, we take as prior distributions $\lambda \sim \mathcal{N}(0, 100)$, $\alpha_1 \sim \mathcal{N}(0, W)$, $\beta_1 \sim \mathcal{N}(0, W)$, $\gamma_{t_1^{(j)},j} \sim \mathcal{N}(0, W)$, $j = 1, \ldots, N-1$ and $\phi \sim \mathcal{G}(10, 1)$, where $\mathcal{G}(\alpha, \beta)$ denotes the gamma distribution with expectation $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$. This means that $\mathbb{E}(\phi) = 10$ and $\mathbb{V}(\phi) = 10$ *a priori*. These values of mean and variance for ϕ express our prior belief of overdispersion for the delayed notifications in the considered data.

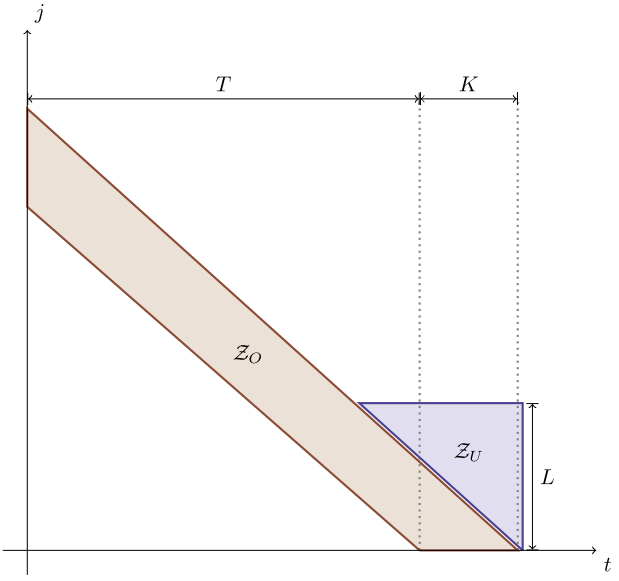


Fig. 2. Observed and unobserved delayed notifications data sets represented by $\mathcal{Z}_O$ (red) and $\mathcal{Z}_U$ (blue), respectively. The unobserved set $\mathcal{Z}_U$ consists of the first L unobserved delayed notifications. The idea is to generate plausible observations of the unobserved set $\mathcal{Z}_U$ based on information of the observed set $\mathcal{Z}_O$. Note that, for fixed t, the smaller the number of delayed notifications in $\mathcal{Z}_O$, the greater the number of delayed notifications in $\mathcal{Z}_U$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

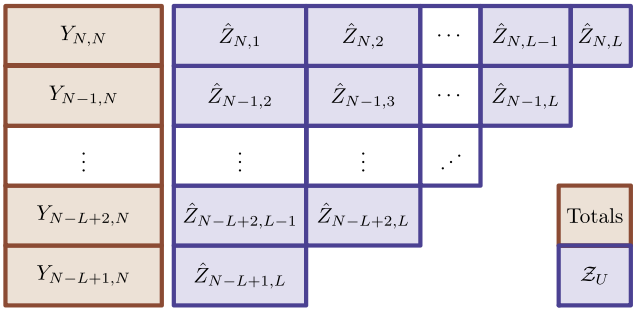


Fig. 3. Schematic for notification delay correction. The $N = T + K$ denotes the more recent time. The strategy consists of generating plausible upcoming delayed notifications (up to a total of L) and synthetically update the total of occurrences ($Y_{t,N}$). Note that it is necessary to generate more synthetic unobserved delayed notifications to update total of occurrences for more recent days. Specifically, we aggregate the generated $\hat{Z}_{N-L+l,L-l+1}, \hat{Z}_{N-L+l,L-l+2}, \ldots, \hat{Z}_{N-L+l,L}$ to the observed total of occurrences $Y_{N-L+l,N}$, forming the delayed corrected total $\hat{Y}_{N-L+l,N+l}$, $l = 1, \ldots, L$.

The delay corrected observations are

$$\hat{Y}_{N-L+l,N+l} = Y_{N-L+l,N} + \sum_{k=0}^{l-1} \hat{Z}_{N-L+l,L-k}, \quad l = 1, \ldots, L,$$

where, in this case, $\hat{Z}_{N-L+l,L-k}$ denote the sample mean calculated on the correspondent draws from $\mathcal{Z}_U, \Phi|\mathcal{Z}_O$. An illustration of the delay correction is presented in Fig. 3.

2.2. Forecasting

We now discuss the methodology for daily deaths and daily reported cases prediction. For simplicity, since we will apply the method here discussed to data corrected for delayed notifications, we drop the previous notation and denote the count variable by Y_t. Once again, in order to account for possible overdispersion,

we assume $Y_t \sim \text{BN}(\mu_t, \theta)$. We will consider a Bayesian approach to fit the model. The main step here is to choose a suitable functional form for μ_t. In this paper, inspired by [9], the starting point is to consider a generalized logistic curve to describe the expected cumulative growth denoted by $\mathscr{U}_t$. In particular, we assume

$$\mathscr{U}_t = \frac{a}{(1+\exp\{-c(t-b)\})^f}, \quad a, b, c, f > 0,$$

where a denotes the maximum value of $\mathscr{U}_t$, f is a skewness parameter and, when $f = 1$, b and c denote the inflection point and the logistic growth rate (or steepness) of $\mathscr{U}_t$, respectively. In this context, the associated functional form for μ_t is given by

$$\mu_t = \frac{\partial}{\partial t}\mathscr{U}_t = \frac{acf\exp\{c(t-b)\}}{(1+\exp\{c(t-b)\})^{f+1}}, \quad t = 1, 2, \ldots. \tag{3}$$

From the pandemic point of view, the day of maximum and the maximum number of occurrences are key features. We denote these quantities by $\mathcal{T}$ and $\mathscr{M}$, respectively. Rewriting (3) in terms of $\mathcal{T}$ and $\mathscr{M}$ will make inference and prior elicitation simpler. From $\mathcal{T} = \text{argmin}_t(\mu_t)$ (which may be obtained from $\mu'_{\mathcal{T}} = 0$) and $\mathscr{M} = \mu_{\mathcal{T}}$, we obtain these values in terms of the original parameters as

$$\mathcal{T} = b + \frac{\log f}{c} \quad \text{and} \quad \mathscr{M} = ac\left(\frac{f}{f+1}\right)^{f+1}.$$

Note that, as mentioned above, when $f = 1$, the inflection point is $\mathcal{T} = b$. Therefore, Equation (3) may be rewritten as

$$\mu_t = \mathscr{M}\frac{(f+1)^{f+1}\exp\{-c(t-\mathcal{T})\}}{(f+\exp\{-c(t-\mathcal{T})\})^{f+1}}. \tag{4}$$

Let $\mathscr{C}$ denote the cumulative total of occurrences, such is $\mathscr{C} = \lim_{t\to\infty}\mathscr{U}_t = a$. Thus, we may rewrite

$$c = \frac{\mathscr{M}}{\mathscr{C}}\left(\frac{f+1}{f}\right)^{f+1}. \tag{5}$$

For the daily reported cases data, a preliminary exploratory study has shown that it might be necessary to include a factor regarding week days and days off. This covariate will be denoted by $X_t = 1 - \mathbb{1}_{(t \text{ is week day})}$. The preliminary investigation also indicates that X_t only affects the height of μ_t, such that Equation (4) is extended to

$$\mu_t = \mathscr{M}\exp\{\zeta X_t\}\frac{(f+1)^{f+1}\exp\{-c(t-\mathcal{T})\}}{(f+\exp\{-c(t-\mathcal{T})\})^{f+1}}, \tag{6}$$

where $\exp\{\zeta\}$ is the multiplicative effect when t refers to a day off.

Similarly to delay correction, due to the unstable nature of the phenomenon, we will assume the following dynamic evolution to parameters $\mathcal{T}$, $\mathscr{M}$ and c:

$$\begin{aligned}
\mathcal{T}_t|\mathcal{T}_{t-1} &\sim \mathcal{LN}(\log\mathcal{T}_{t-1}, W_{\mathcal{T}}),\\
\mathscr{M}_t|\mathscr{M}_{t-1} &\sim \mathcal{LN}(\log\mathscr{M}_{t-1}, W_{\mathscr{M}}),\\
c_t|c_{t-1} &\sim \mathcal{LN}(\log c_{t-1}, W_c),
\end{aligned}$$

where $V \sim \mathcal{LN}(\mu, \sigma^2)$ means that V follows the lognormal distribution with $\mathbb{E}(\log V) = \mu$ and $\mathbb{V}(\log V) = \sigma^2$ and we fix the variances as $W_{\mathcal{T}} = W_{\mathscr{M}} = W_c = W = 1/6400$ to ensure the parameters do not change more

Table 1
Means and 0.95 HPD credibility intervals of the constant parameters.

Measure	μ	ϕ
Lower	-2.0482	4.2033
Mean	-1.6053	7.3644
Upper	-1.1565	11.7102

than 2.5% with probability 0.95. For simplicity, the parameter vector and the collection of observed daily occurrence numbers are represented by

$$\Theta = (\zeta, \mathcal{T}_1, \ldots, \mathcal{T}_N, \mathscr{M}_1, \ldots, \mathscr{M}_N, c_1, \ldots, c_N, f, \theta)$$

and

$$\mathcal{Y} = \{Y_t, t = 1, \ldots, N\},$$

respectively. The inference was carried out by using MCMC for drawing samples from the posterior distribution of $\Theta|\mathcal{Y}$. The prior distributions for starting the evolutionary parameters were set as $\mathcal{T}_1 \sim \mathcal{LN}(\log \mathcal{T}_0, W)$, $\mathscr{M}_1 \sim \mathcal{LN}(\log \mathscr{M}_0, W)$ and $c_1 \sim \mathcal{LN}(\log c_0, W)$, which means that, *a priori*, at the beginning of the observation period we expect that, at the log scale, the day of the maximum and the daily maximum will be $\mathcal{T}_0$ and $\mathscr{M}_0$, respectively. We take a bad scenario with $\mathcal{T}_0 = N + 50$, that is, *a priori*, we believe that it will take 50 more days to arise the maximum. The choice of $\mathscr{M}_0$ and c_0 will be explained in Section 3. For constant parameters, we elicited the following prior distributions: $\zeta \sim \mathcal{N}(0, 1)$; $f \sim \mathcal{LN}(\log 1, 1)$; and $\theta \sim \mathcal{G}(10, 1)$. Similarly to Subsection 2.1, this means that $\mathbb{E}(\theta) = 10$ and $\mathbb{V}(\theta) = 10$ *a priori*. Again, these values of mean and variance for θ express our prior belief of overdispersion for the daily occurrences in the considered data.

3. Application

In this section, we apply the methodology developed previously to analyze the daily deaths and the daily reported cases of COVID-19 data in Espírito Santo, Brazil. The data were obtained by systematically accessing https://coronavirus.es.gov.br/painel-covid-19-es and monitoring and recording the daily changes in the provided data.

3.1. Daily deaths

The constant parameters for the delay correction model were estimated and are presented in Table 1. The negative values for the overall mean parameter μ indicate that the contributions to delayed notifications are mostly from the time index t and the delay lag j. Note the relatively small dispersion parameter ϕ, indicating that daily deaths notifications are moderately overdispersed.

Fig. 4 display the estimated coefficients. In Figs. 4a and 4b the shaded areas represent the Highest Posterior Density (HPD) intervals with 0.95 credibility. As expected, Fig. 4a indicates that delayed notification increases with time. On the other hand, Fig. 4b shows a non-monotonous behavior, increasing for small delay lags and decreasing after a peak around a delay lag of 10 days after the corresponding day. Most delayed notifications seems to occur until 30 days after the respective day. This led us to choose the correction window as $L = 30$. Note the resemblance of the shapes displayed in Figs. 2 and 4c. Fig. 4c indicates that delay lag increment experiences an increase for more recent days.

Fig. 5 shows the results of the proposed method for delay correction. Shaded areas are the 0.95 credibility intervals for delay correction. In Fig. 5a, the 14 more recent days were discarded in order to visually inspect

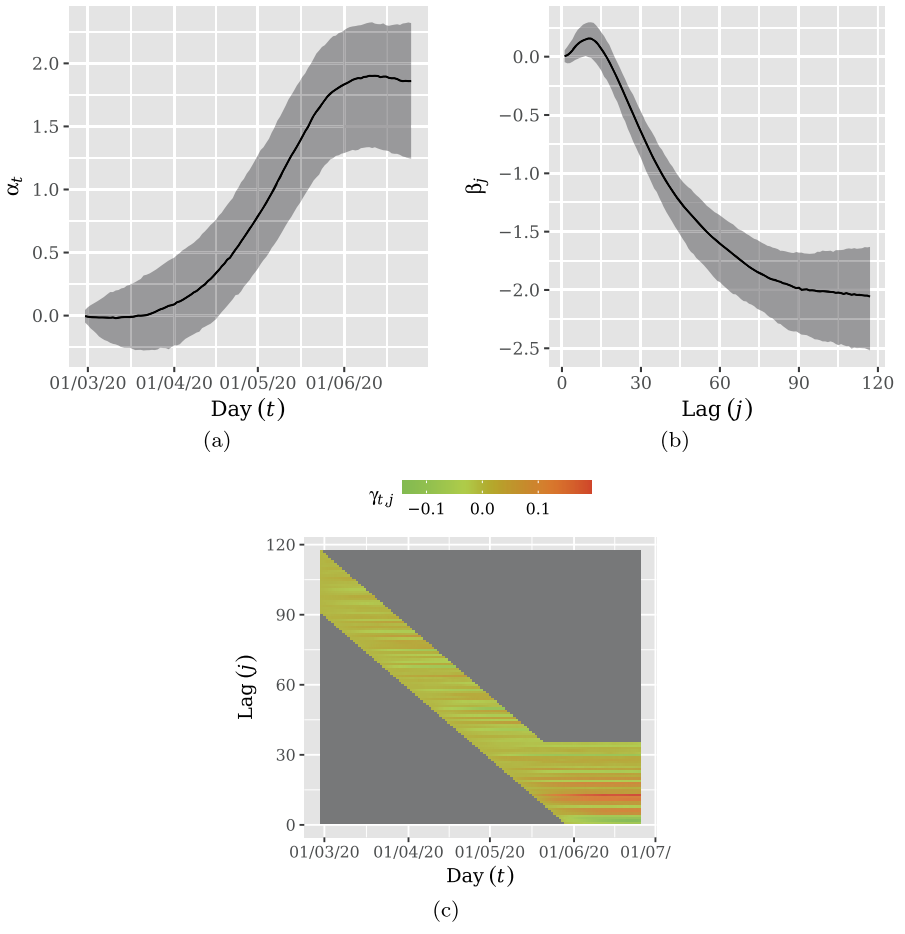


Fig. 4. Coefficient evolution: (a) temporal increment, α_t; (b) delay lag increment, β_j; (c) temporal increment in the delay lag effect, $\gamma_{t,j}$. Shaded areas in (a) and (b) represent the Highest Posterior Density (HPD) intervals with 0.95 credibility. Frame (a) indicates that delayed notification increases with time. Frame (b) shows a non-monotonous behavior, increasing for small delay lags and decreasing after a peak around a delay lag of 10 days after the corresponding day. Frame (c) indicates that delay lag increment experiences an increase for more recent days.

the performance of delay correction. We note that 14 days after delay correction, the updated data tends to be inside the credibility interval. Fig. 5b presents the result of delay correction considering the whole sample, which will be used for forecasting. This plot evidences the high degree of impact caused by delayed notifications.

The 14 more recent days were discarded in order to visually inspect the performance of delay correction. We note that 14 days after delay correction, the updated data tends to be inside the credibility interval. Fig. 5b presents the result of delay correction considering the whole sample, which will be used for forecasting. This plot evidences the high degree of impact caused by delayed notifications.

We now turn to investigation of the forecasting for daily deaths. We applied the methodology in Subsection 2.2 to the delay corrected daily deaths data presented in Fig. 5b. In this study, *a priori*, we tried to be conservative by considering bad scenarios when fitting the model. The $\mathscr{M}_0$ value was taken to be around the double of the maximum observed daily deaths, which gives $\mathscr{M}_0 = 70$. In addition, considering a lethality of 4%, a underreporting percentage guess of 10%, a 50% contamination to slow down the spread of COVID-19 and the Espírito Santo state population of ≈ 3800000, *a priori*, we take the total of deaths as $\mathscr{C}_0 = 7600$. Considering a symmetric behavior ($f_0 = 1$), we compute the initial value c_0 using Equation (5), which gives $c_0 = \frac{70}{7600}(\frac{1+1}{1})^{1+1} \approx 0.0368$. The estimated constant parameters are presented in Table 2. For the ζ parameter, following [18], according to the 0.95 HPD interval, the number of daily deaths is not statistically

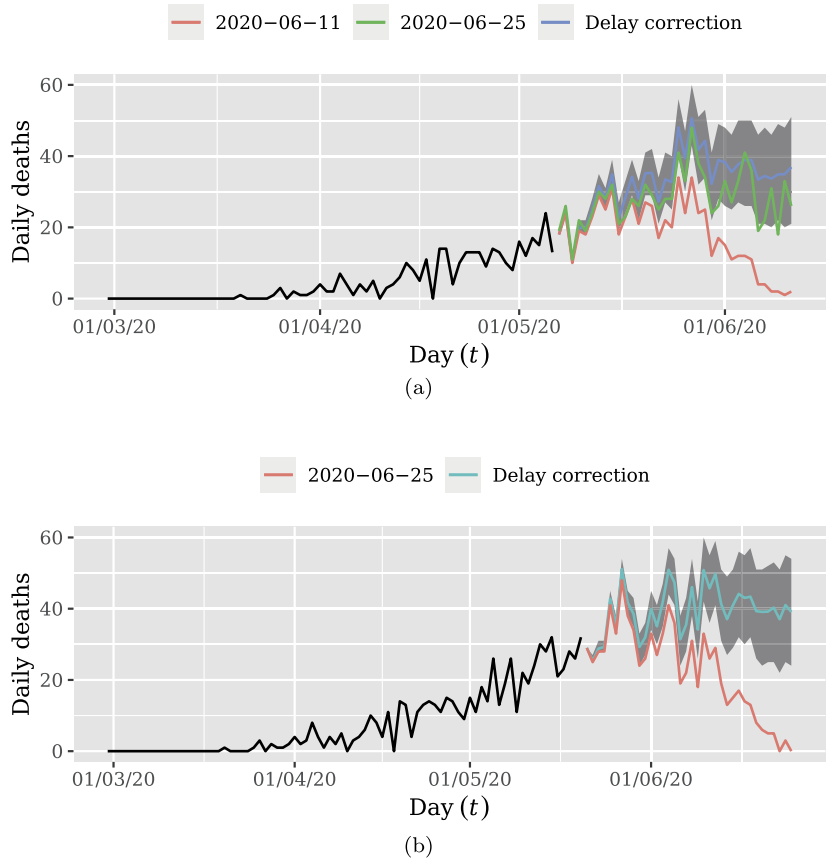


Fig. 5. Delay correction of daily deaths: (a) discarding the 14 more recent days; (b) the whole data. Frame (a) shows that the methodology is able to accurately correct the total of occurrences of the period for the unobserved delayed notifications. This also evidence the high impact of disregarding future delayed notifications from the analysis. Frame (b) shows the complete delayed corrected dataset. The corrected data displayed in Frame (b) will be used to perform the forecasting.

Table 2
Means and 0.95 HPD credibility intervals of the constant parameters.

Measure	ζ	f	θ
Lower	-0.2925	1.3109	12.3850
Mean	-0.1522	1.7692	19.8102
Upper	0.0062	2.3110	26.8390

affected by the week days and days off, since that, given the observed data, the null effect lie within the interval with the most plausible values of ζ. The estimated f indicates a right skew shape of the curve, that is the decay of the daily deaths will be slower than the growth stage.

The forecast is displayed in Fig. 6. This figure shows that the peak of daily deaths was not reached yet and will occur between July 2, 2020 and August 10, 2020 with 0.95 credibility.

3.2. Daily reported cases

The constant parameters for the delay correction model were estimated and are presented in Table 3. Note the small dispersion parameter ϕ, indicating that daily reported cases are dramatically overdispersed.

Fig. 7 display the evolution of the estimated coefficients. In Figs. 7a and 7b the shaded areas represent the HPD 0.95 credibility intervals. Again, Fig. 7a indicates that delayed notification increases with time.

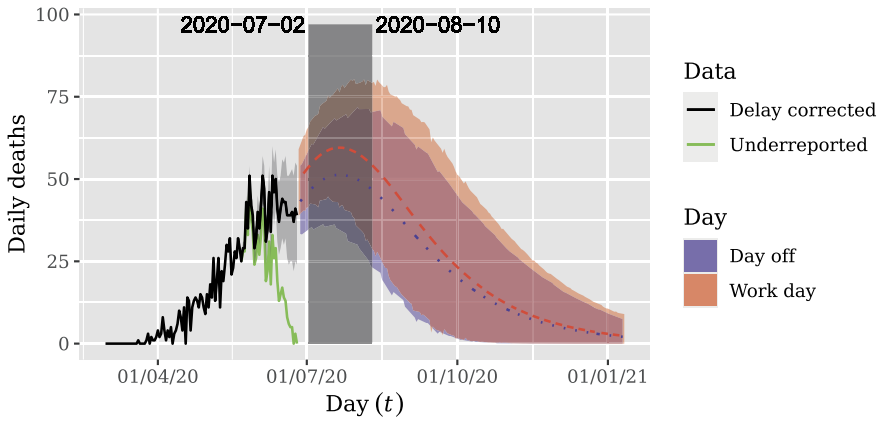


Fig. 6. Long-term forecasts for delay corrected daily deaths. Shaded areas represent 0.95 HPD forecast intervals. Note that the peak of daily deaths was not reached yet and will occur between July 2, 2020 and August 10, 2020.

Table 3
Means and 0.95 HPD credibility intervals of the constant parameters.

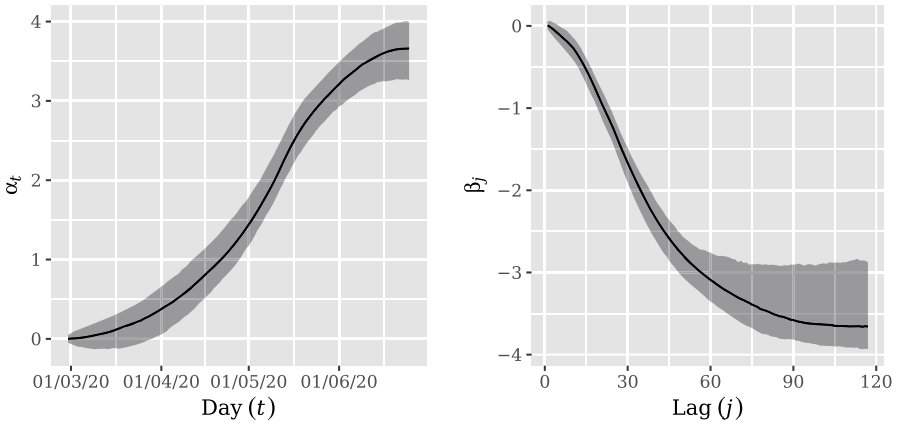
Measure	μ	ϕ
Lower	0.0288	0.3490
Mean	0.3119	0.3877
Upper	0.6266	0.4300

Unlike the daily deaths data, in this case, delayed notifications present a monotonous decreasing behavior in function of delay lag (Fig. 7b). Similarly to daily deaths, most delayed notifications seem to occur until 30 days after the corresponding day. This led us to choose the correction window as $L = 30$. Note the resemblance of the shapes displayed in Figs. 2 and 7c. Fig. 7c indicates that delay lag increment experiences an increase for more recent days.

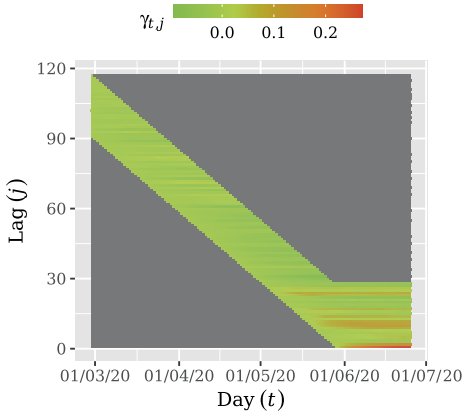
Fig. 8 presents the results of the proposed method for delay correction. Shaded areas are the HPD 0.95 credibility intervals for delay correction. In Fig. 8a, the 14 more recent days were discarded in order to visually inspect the performance of delay correction. We note that 14 days after delay correction, the updated data tend to be inside the credibility interval. Fig. 8b presents the result of delay correction considering the whole sample, which will be used for forecasting. This plot illustrates the major impact caused by delayed notifications.

We now turn to investigation of the forecasting for daily reported cases. The methodology of Subsection 2.2 was applied to the corrected data in Fig. 8b. For the daily reported cases, we use similar arguments to specify the initial values. The $\mathscr{M}_0$ value was taken to be around the double of the maximum observed daily reported cases, which gives $\mathscr{M}_0 = 2500$. In addition, considering a underreporting percentage guess of 10%, a 50% contamination to slow down the spread of COVID-19 and the Espírito Santo state population of ≈ 3800000, *a priori*, we take the total of reported cases as $\mathscr{C}_0 = 190000$. Considering a symmetric behavior ($f_0 = 1$), we compute the initial value c_0 using Equation (5), which gives $c_0 = \frac{2500}{190000}(\frac{1+1}{1})^{1+1} \approx 0.0526$. The estimated constant parameters are presented in Table 4. At a 0.95 credibility level, the HPD interval for the ζ parameter shows a strong evidence to support a negative effect of day offs in reported cases. The estimated f indicates a right skewed curve, that is the decay of the daily reported cases will be slower than the growth stage.

The forecast is displayed in Fig. 9. This figure shows that the peak of daily deaths was not reached yet and will occur between June 29, 2020 and July 31, 2020 with 0.95 credibility.



(a) (b)



(c)

Fig. 7. Coefficient evolution: (a) temporal increment, α_t; (b) delay lag increment, β_j; (c) temporal increment in the delay lag effect, $\gamma_{t,j}$. Shaded areas in (a) and (b) represent the Highest Posterior Density (HPD) intervals with 0.95 credibility. Frame (a) indicates that delayed notification increases with time. Frame (b) shows a monotonous decreasing behavior of the delay lag increment, that is, the greater the lag the smaller the increment in the delayed notifications mean. Frame (c) indicates that delay lag increment experiences an increase for more recent days.

Table 4
Means and 0.95 HPD credibility intervals of the constant parameters.

Measure	ζ	f	θ
Lower	-0.7429	1.1358	16.2158
Mean	-0.6388	1.3019	24.2250
Upper	-0.5306	1.5000	31.7097

4. Final remarks

This paper focuses in the correction of notification delay and predictions of daily COVID-19 cases and deaths. The proposed models were estimated from a Bayesian point of view. In both methods, we resorted to the negative binomial distribution in order to accommodate the overdispersion caused by the usual population heterogeneity.

The first methodology has presented good performance and has been able to capture delayed notifications. It was observed that daily death notifications are moderately overdispersed. Additionally, delayed notifications increase with time. The model was also able to show the high impact caused by delayed noti-

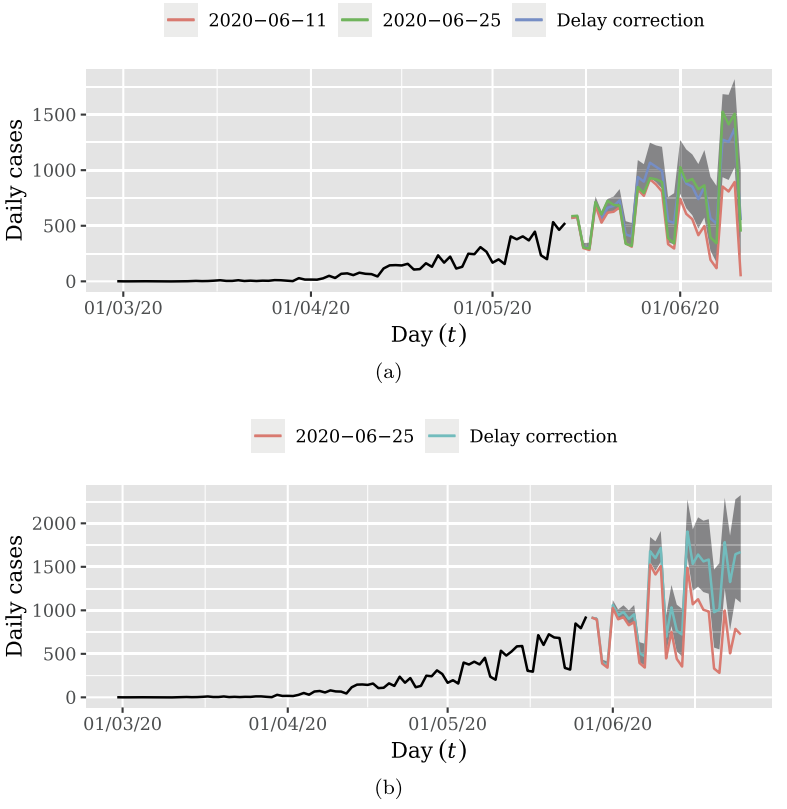


Fig. 8. Delay correction of daily reported cases: (a) discarding the 14 more recent days; (b) the whole data. Frame (a) shows that the methodology is able to accurately correct the total of occurrences of the period for the unobserved delayed notifications. This also evidence the high impact of disregarding future delayed notifications from the analysis. Frame (b) shows the complete delayed corrected dataset. The corrected data displayed in Frame (b) will be used to perform the forecasting.

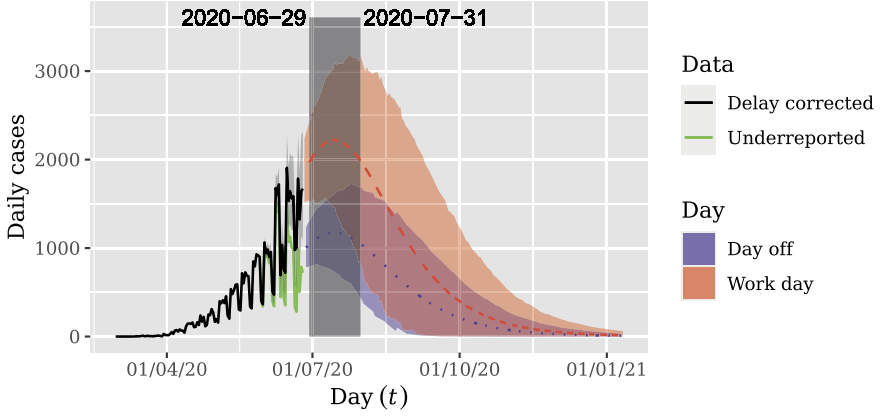


Fig. 9. Long-term forecasts for delay corrected daily reported cases. Shaded areas represent 0.95 HPD forecast intervals. Note that the peak of daily reported cases was not reached yet and will occur between June 29, 2020 and July 31, 2020.

fications. Another interesting result was the finding of the skewness of the curve, that is, the decay of the daily deaths will be slower than the growth stage.

The functional form and the inclusion of an explanatory variable regarding week days and days off, adopted for the prediction method, was able to explain satisfactorily the data dynamics and to provide posterior inference for maximum number of occurrences and for the peak of the occurrences. The model showed that the reported cases are highly overdispersed. Unlike the daily deaths, delayed notifications show

a monotonous decreasing behavior in function of the delay lag. At last, there was strong evidence on the effect of the day in reported cases.

Although the results in this paper indicate that the proposed methods are promising, we envision as potential way of improving the results to consider the impact of media in COVID-19 dynamics. This impact has recently been considered by [17] and it would be interesting to extend our model in a similar manner.

References

[1] H.A. Adamu, M. Muhammad, A.M. Jingi, M.A. Usman, Mathematical modelling using improved SIR model with more realistic assumptions, Int. J. Eng. Appl. Sci. 6 (1) (2019) 64–69.

[2] L. Avilez, Coronavírus no ES: pacientes relatam longa espera por resultado de teste, A Gazeta (2020), Vitória (ES/BR), accessed in April 11, 2020, https://www.agazeta.com.br/es/cotidiano/coronavirus-no-es-pacientes-relatam-longa-espera-por-resultado-de-teste-0620.

[3] L.S. Bastos, T. Economou, M.F. Gomes, D.A. Villela, F.C. Coelho, O.G. Cruz, O. Stoner, T. Bailey, C.T. Codeço, A modelling approach for correcting reporting delays in disease surveillance data, Stat. Med. 38 (22) (2019) 4363–4377.

[4] M.H.A. Biswas, L.T. Paiva, M. d, R. de Pinho, A SEIR model for control of infectious diseases with constraints, Math. Biosci. Eng. 11 (4) (2014) 761–784.

[5] P. Borges, L.G. Godoi, Pólya–Aeppli regression model for overdispersed count data, Stat. Model. 19 (4) (2019) 362–385.

[6] F. Brauer, Compartmental models in epidemiology, in: F. Brauer, P. van den Driessche, J. Wu (Eds.), Mathematical Epidemiology, Springer Berlin Heidelberg, 2008, pp. 19–79.

[7] T. Britton, F. Ball, P. Trapman, A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2, Science 369 (6505) (2020) 846–849.

[8] D. Chen, Modeling the spread of infectious diseases: a review, in: D. Chen, B. Moulin, J. Wu (Eds.), Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases, John Wiley & Sons, 2014, pp. 19–42.

[9] CovidLPTeam, CovidLP: short and long term prediction for COVID-19, Tech. Rep, Statistics Department, Federal University of Minas Gerais, Brazil, 2020, accessed in June 06, 2020, http://est.ufmg.br/covidlp/home/en/.

[10] Y. Fang, Y. Nie, M. Penny, Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: a data-driven analysis, J. Med. Virol. 92 (6) (2020) 645–659.

[11] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, M. Colaneri, Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, Nat. Med. 26 (2020) 855–860.

[12] A. Grant, Dynamics of COVID-19 epidemics: SEIR models underestimate peak infection rates and overestimate epidemic duration, medRxiv, https://doi.org/10.1101/2020.04.02.20050674, 2020.

[13] D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley-Meza, M. Moussaid, A. Johansson, J. Krause, S. Schutte, et al., Saving human lives: what complexity science and information systems can contribute, J. Stat. Phys. 158 (3) (2015) 735–781.

[14] J. Hinde, C.G. Demétrio, et al., Overdispersion: models and estimation, Comput. Stat. Data Anal. 27 (2) (1998) 151–170.

[15] S. Khajanchi, K. Sarkar, Forecasting the daily and cumulative number of cases for the COVID-19 pandemic in India, chaos: an interdisciplinary, J. Nonlinear Sci. 30 (7) (2020) 071101.

[16] S. Khajanchi, S. Bera, T.K. Roy, Mathematical analysis of the global dynamics of a HTLV-I infection model, considering the role of cytotoxic T-lymphocytes, Math. Comput. Simul. 180 (2021) 354–378.

[17] S. Khajanchi, K. Sarkar, J. Mondal, Dynamics of the COVID-19 pandemic in India, arXiv:2005.06286, 2021.

[18] D.V. Lindley, Introduction to Probability and Statistics from a Bayesian Viewpoint (Part 2), Cambridge University Press, 1965.

[19] N. Moradian, H.D. Ochs, C. Sedikies, M.R. Hamblin, C.A. Camargo, J.A. Martinez, J.D. Biamonte, M. Abdollahi, P.J. Torres, J.J. Nieto, et al., The urgent need for integrated science to fight COVID-19 pandemic and beyond, J. Transl. Med. 18 (1) (2020) 1–7.

[20] S.K. Panda, Applying fixed point methods and fractional operators in the modelling of novel coronavirus 2019-nCoV/SARS-CoV-2, Results Phys. 19 (2020) 103433.

[21] R.W. Paterson, R.L. Brown, L. Benjamin, R. Nortley, S. Wiethoff, T. Bharucha, D.L. Jayaseelan, G. Kumar, R.E. Raftopoulos, L. Zambreanu, et al., The emerging spectrum of COVID-19 neurology: clinical, radiological and laboratory findings, Brain 143 (10) (2020) 3104–3120.

[22] M. Perc, N. Gorišek Miksić, M. Slavinec, A. Stožer, Forecasting COVID-19, Front. Phys. 8 (2020) 127.

[23] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/.

[24] Redação Folha Vitória, Com falta de insumos para exame de COVID-19 no ES, 3 mil amostras são encaminhadas para o Paraná, Folha Vitória (2020), Vitória (ES/BR), accessed in April 30, 2020, https://www.folhavitoria.com.br/geral/noticia/05/2020/com-falta-de-insumos-para-exame-de-covid-19-no-es-3-mil-amostras-sao-encaminhadas-para-o-parana.

[25] Redação Folha Vitória, Coronavírus: com atraso na entrega de resultados, capixabas sofrem sem saber quadro clínico, Folha Vitória (2020), Vitória (ES/BR), accessed in June 02, 2020, https://www.folhavitoria.com.br/geral/noticia/06/2020/coronavirus-com-atraso-na-entrega-de-resultados-capixabas-sofrem-sem-saber-quadro-clinico.

[26] P. Samui, J. Mondal, S. Khajanchi, A mathematical model for COVID-19 transmission dynamics with a case study of India, Chaos Solitons Fractals 140 (2020) 110173.

[27] S. Sanche, Y. Lin, C. Xu, E. Romero-Severson, N. Hengartner, R. Ke, High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2, Emerg. Infect. Dis. 26 (7) (2020) 1470–1477.

[28] K. Sarkar, S. Khajanchi, J.J. Nieto, Modeling and forecasting the COVID-19 pandemic in India, Chaos Solitons Fractals 139 (2020) 110049.
[29] F.L. Schumacher, C.S. Ferreira, M.O. Prates, A. Lachos, V.H. Lachos, A robust nonlinear mixed-effects model for COVID-19 deaths data, Stat. Interface 14 (1) (2021) 49–57.
[30] J. Whitworth, COVID-19: a fast evolving pandemic, Trans. R. Soc. Trop. Med. Hyg. 114 (4) (2020) 241–248.
[31] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, Lancet 395 (10225) (2020) 689–697.