Infectious Disease Modelling 5 (2020) 699–713

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Bayesian modeling of COVID-19 cases with a correction to account for under-reported cases☆

Anderson Castro Soares de Oliveira [a], Lia Hanna Martins Morita [a, *], Eveliny Barroso da Silva [a], Luiz André Ribeiro Zardo [a], Cor Jesus Fernandes Fontes [c], Daniele Cristina Tita Granzotto [b]

[a] Departamento de Estatística, Universidade Federal de Mato Grosso - UFMT, CEP: 78060-900, Cuiabá, MT, Brazil
[b] Departamento de Estatística, Universidade Estadual de Maringá - UEM, CEP: 87020-900, Maringá, PR, Brazil
[c] Faculdade de Medicina, Universidade Federal de Mato Grosso - UFMT, CEP: 78060-900, Cuiabá, MT, Brazil

ARTICLE INFO

Article history:
Received 24 May 2020
Received in revised form 14 September 2020
Accepted 20 September 2020
Available online 24 September 2020
Handling editor: Dr. J Wu

Keywords:
COVID-19
Under-reporting
SIR model
Bayesian aproach

ABSTRACT

The novel of COVID-19 disease started in late 2019 making the worldwide governments came across a high number of critical and death cases, beyond constant fear of the collapse in their health systems. Since the beginning of the pandemic, researchers and authorities are mainly concerned with carrying out quantitative studies (modeling and predictions) overcoming the scarcity of tests that lead us to under-reporting cases. To address these issues, we introduce a Bayesian approach to the SIR model with correction for under-reporting in the analysis of COVID-19 cases in Brazil. The proposed model was enforced to obtain estimates of important quantities such as the reproductive rate and the average infection period, along with the more likely date when the pandemic peak may occur. Several under-reporting scenarios were considered in the simulation study, showing how impacting is the lack of information in the modeling.

1. Introduction

The COVID-19 epidemic disease is caused by the new SARS-CoV-2 coronavirus associated with the severe acute respiratory syndrome (SARS) that began in Wuhan, China, late 2019 (Rodríguez-Morales et al., 2020). After the first detected case in China, the disease continued to spread globally with exported cases confirmed in all of the continents worldwide. In a matter of a few months, the disease overtook 80 thousand reported cases until early April 2020. On March 12th, 2020 the World Health Organization (WHO) declared COVID-19 as pandemic disease, when more than 20 thousand cases and almost a thousand deaths were registered in the European Region - the center of this pandemic according to the Europe's Standing Committee (WHO, 2020).

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
E-mail address: liamorita@ufmt.br (L.H.M. Morita).

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2020.09.005

There are still many unknowns about COVID-19 and the lack of evidence complicates the design of appropriate response policies - for example, it is impossible to precisely say something about the mortality rate and determine the disease recurrence rate (Lenzer, 2020).

Despite uncertainties, the frightening speed through which this disease spreads across communities and the collapse that it is capable of causing to the health systems are facts that must be faced. The exponential growth of the cases and the consequent number of deaths had been observed in a short period. In mid-January 2020, a few weeks after the first detected COVID-19 case in the world, the countries that are close to the territory of the virus origin, on the Asian continent, as well in European and American Regions also began to report cases of the disease. Five months later, more than 200 countries and territories around the world have reported over to 3 million confirmed cases of COVID-19 and a death toll of about 200 thousand people.

In Brazil, the first confirmed COVID-19 case occurred on February 25th, 2020. This first case was a 61 years-old male, who stayed from February 9th to February 20th, 2020 in Lombardy - an Italian region were a significant outbreak was ongoing at that time. On March 17th, the health authorities in São Paulo confirmed the Brazilian death from the new coronavirus. The victim, whose identity has not been disclosed, had been hospitalized in São Paulo city.

Preserving due proportions, COVID-19 is not the first experienced significant outbreaks of infections that were declared Public Health Emergencies of International Concern by the WHO. Year after year we also have experimented with the Zika and Chikungunya outbreaks in the last decade and continue facing the huge consequences of dengue. Confronting outbreaks in the large Brazilian territory is a twofold problem. The first is the demographic and territorial size of the country, with an estimated population of 210 million according to the Brazilian Institute for Geography and Statistics and the heterogeneity intrinsic to its extensive territory. Another problem pointed out by the past epidemics run into a recurring problem of under-reporting (de Oliveira et al., 2017; Stoner et al., 2019).

The COVID-19, given its complexity and behavior, exposed the problem of under-reporting disease occurrence not only in Brazil but in several countries worldwide. As a consequence, the lack of information has launched a warning about the researchers of the world concerning models and estimates, since the database available may not be reliable from what had indeed been observed.

Focusing on the modeling and estimating, aiming to preview the behavior and the speed of the COVID-19 growth, this paper presents an approach to address the problem of under-registration of COVID-19 cases in Brazil, proposing methodologies to work on the inaccuracy of the official reported cases. Then, we investigate a general framework for correcting under-reporting data making it possible to perform a model, in a Bayesian framework, which allows great flexibility and leads to complete predictive distributions for the true counts, therefore quantifying the uncertainty in correcting the under-reporting. Several scenarios of under-reporting were considered in a simulation study, presenting the real lack of data impact.

This paper is organized as follows. Section 2 describes the methodology for estimating the reported rates. In Section 3, we introduce the SIR model for modeling epidemics. In Section 4, we introduce the Bayesian framework for the SIR model with a modification to account for under-reporting. In Section 5 we show the model application for COVID-19 cases in Brazil and in Section 6, we present a simulation study of the proposed model. Finally, in Section 7, we give some concluding remarks.

2. Reported rate estimation

Although in the first moment there was a real hunt for the size and the moment of the COVID-19 cases peak, the most important aspects of the outbreak are the growth rate of the infection. Statistical and mathematical models are being used to preview the rates and analyze the growth curve behavior to assist health public managers in decision-making (Cotta et al., 2020).

According to Kim et al. (2020), estimating the case fatality rate (CFR) is a high priority in response to this pandemic. This fatality rate is the proportion of deaths among all confirmed patients with the disease, which has been used to assess and compare the severity of the epidemic between countries. The rates can also be used to assess the healthcare capacity in response to the outbreak. Indeed, several researchers are interested in estimating the CFR in the peak of the outbreak, analyzing its variation among different countries, and check the influence of other features as ages, gender, and physical characteristics in the CFR of the COVID-19.

Aiming to estimate the CFR, first of all, lets set up the Brazilian scenario of COVID-19 case notification: the Brazilian Ministry of Health collects daily all confirmed cases data for Brazil and all its states. Although the data presented by the health authorities are official, they are only from patients with COVID-19 confirmed by blood and/or swab positive tests. Given the scarcity of tests for all the suspected individuals, the notified patients are only those with severe disease or that demanding hospitalization. It is relevant to highlight that no clinically diagnosed patient, even those with symptoms compatible with the disease have been officially counted, evidencing an under-reporting of the case frequency.

Faced with the lack of COVID-19 tests, which naturally leads to the under-reporting data, before any modeling purpose we have the desire to correct and update the current numbers, bringing them as close as possible to reality.

Following Russel et al. (2020), we also based on a delay-adjusted case fatality ratio to estimate under-reporting, using the incidence of cases and deaths to estimate the number of notified cases by

$$\mu_t = \sum_{j=0}^{t} \frac{c_{t-j} f_j}{c_t}, \tag{1}$$

where c_t is the daily incidence of cases at the moment t, f_j is the proportion of cases with a delay between the confirmation and the death, and μ_t represents the underestimation proportion of cases with known outcomes, (Nishiura et al., 2009).

Then, the corrected CFR is given by

$$\mathrm{CFR}_c = \frac{m_t}{\mu_t}, \tag{2}$$

where m_t is the cumulative number of deaths.

To estimate the potential for under-reporting, we assume that the CFR is 1.4% with a 95% confidence interval from 1.2% up to 1.7% found in China (Guan et al., 2020; WHO, 2020). Thus, the potential for reporting rate is given by

$$\eta = \frac{1.4}{\mathrm{CFR}_c}. \tag{3}$$

3. The SIR model

Epidemic models are tools widely used to study the mechanisms by which diseases spread, to predict the course of an outbreak, and to evaluate strategies to control an epidemic disease. Several analyses of an epidemic spreading disease can be found in the literature that applies the time series model (given the historical data), the log-logistic family of models (the Chapman, Richards, among others), and compartments models (Bjørnstad, 2018).

Kermack and McKendrick (1927) proposed a class of compartmental models that simplified the mathematical modeling of infectious disease transmission. Entitled as SIR model, it is a set of general equations which explains the dynamics of an infectious disease spreading through a susceptible population. Essentially, the standard SIR model is a set of differential equations that can suit the Susceptible (if previously unexposed to the pathogen), Infected (if currently colonized by the pathogen), and Removed (either by death or recovery) as follows:

$$\frac{dS}{dt} = -\beta SI,$$

$$\frac{dI}{dt} = \beta SI - \gamma I,$$

$$\frac{dR}{dt} = \gamma I,$$

where S, I and R are the total number of susceptible, infected and removed individuals in the population, respectively, γ is the removal rate and β is the infectious contact rate.

If we are interested in investigating the unknown number of infected individuals at the moment t in an under-reporting scenario, one can establish the following relation

$$\eta I(t) = c_t,$$

where η is as given in (3) and c_t is as given in (1).

It is important to note that

$$\frac{dS}{dt}+\frac{dI}{dt}+\frac{dR}{dt}=0$$

and so, the total population, $S(t)+I(t)+R(t)$ remains constant for all $t \geq 0$.

For the practical point of view, the most interesting issue is to estimate $\frac{1}{\gamma}$, which determines the average infection period, and the basic reproductive ratio R_0. For the simple SIR model, all individuals in the population are susceptible, that is, $S(t) = 1$, then R_0 is defined as the expected number of secondary infections from a single index case and given by the expression $R_0 = \frac{\beta}{\gamma}$ (Keeling & Rohani, 2011).

4. Bayesian approach

The Bayesian methods are used in several works (Gelman et al., 1995); (Paulino et al., 2018). The Bayesian approach in the context of the SIR model is a flexible way to account for uncertainty in the parameters, in the form of the disease transmission dynamic. The Dirichlet-Beta state-space model appears in some papers as Osthus et al. (2017) and Song et al. (2020). The target distribution for inference is the *a posteriori* distribution of the quantities of interest, more specifically β, γ, and R_0: the infectious contact rate, the removal rate, and the propagation rate, respectively. The application of this methodology is through Markov chain Monte Carlo methods (MCMC) through Gibbs Sampling and the Metropolis-Hastings algorithm (Chib & Greenberg, 1995).

The use of Dirichlet distribution for the proportions of susceptible, infected, and removed individuals in the target population are a feasible way to guarantee that the support set of these quantities has boundaries, for example, the number of infected individuals must be always positive.

4.1. Model specification

In this section, we present a modification to account for under-reporting in the context of the Dirichlet-Beta state-space model from Osthus et al. (2017). This adaptation is based on a reparametrization of Beta distribution that includes the reported rate estimate, η, from equation (3).

The Beta distribution, as is well known, is very flexible for proportions modeling since its density can have quite different shapes depending on the values of the two parameters that index this distribution (Ferrari & Cribari-Neto, 2004). For this reason, we made a reparametrization to the Beta model in such a way that we could obtain a regression structure for the means of the response variables associated with a precision parameter.

Let Y_t^I be the reported infected proportion, Y_t^R be the reported removed proportion and $\boldsymbol{\theta}_t = (\theta_t^S, \theta_t^I, \theta_t^R)$ be the true but unobservable susceptible, infectious, and removed proportions of the population, respectively.

Hence, we rewrite the SIR model in terms of these unobservable proportions as the following

$$\begin{aligned}
\frac{d\theta_t^S}{dt} &= -\beta\theta_t^S\theta_t^I, \\
\frac{d\theta_t^I}{dt} &= \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \\
\frac{d\theta_t^R}{dt} &= \gamma\theta_t^I.
\end{aligned} \tag{4}$$

Then, the distributions for Y_t^I, Y_t^R, and $\boldsymbol{\theta}_t$ are given below

$$Y_t^I \left| \theta_t^I, \varphi \tilde{} \text{Beta}\left(\lambda_I \eta\theta_t^I, \lambda_I\left(1-\eta\theta_t^I\right)\right),\right.$$

$$Y_t^R \left| \theta_t^R, \varphi \tilde{} \text{Beta}\left(\lambda_R \eta\theta_t^R, \lambda_R\left(1-\eta\theta_t^R\right)\right),\right.$$

$$\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \varphi \tilde{} \text{Dirichlet}(\kappa f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)),$$

where $\varphi = (\beta, \gamma, \boldsymbol{\theta}_0, \kappa, \lambda_I, \lambda_R)$ is the parameter vector for this model. Since we consider the beta distribution, we are assuming that $E[Y_t^I] = \eta\theta_t^I$ and $E[Y_t^R] = \eta\theta_t^R$ and the parameters $\lambda_I > 0$ and $\lambda_R > 0$ are responsible for controlling of the distribution variance. Besides that, the parameter $\kappa > 0$ controls the variance of the Dirichlet distribution. The solution for the differential equations in (4) is given by $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$, that have the role of propagating the latent state θ_t forward in one time step.

Note that it is necessary to obtain the solutions for the proportions θ_t^S θ_t^I and θ_t^R. These solutions can be found using the Runge-Kutta fourth-order method, in short RK4, for solving non-linear ordinary differential equations (Mathews, 1992) and can be seen in Appendix A.

5. Case study: the COVID-19 Brazilian data

The official Brazilian data consists of daily collections carried out by the national health department with records of infected individuals and deaths in all states and national territory, from February 26th, 2020 when the first case of COVID-19 was registered up to May 20th, 2020.

It is notable in Brazil a lack of testing due to the registry of only severe cases and consequently under-reporting cases of COVID-19. Taking this fact into account, we consider for this research not only the official data but also the estimates of reported rate.

5.1. Reported rate of COVID-19

In order to obtain the estimate of reported rate, assume that delay in confirmation until death follows the same estimated distribution of hospitalization until death. Using data from COVID-19 in Wuhan, China, between December 17th, 2019, and January 22nd, 2020, it has a lognormal distribution with mean of 13, median of 9.1 and standard deviation of 12.7 days (Linton & Kobayashi, 2020). This methodology based on the information of delay from hospitalization until death is reasonable since China was considered as one of the countries that most tested the population for the virus, and consequently, it is supposed to have a tiny under-reporting rate.

Using the methodology presented in section 2 and assuming that c_t in (1) is the daily incidence of official cases reported by the Brazilian Ministry of Health, the reporting rate in Brazil, η, was estimated to be 0.07 with 95% confidence interval from 0.06 up to 0.08. Prado et al. (2020) obtained a reporting rate of 0.08 with data from Brazil until April 10th, 2020. These results are similar to the analysis from Ribeiro and Bernardes (2020), which present a $7.7:1$ Funder-reporting rate, meaning that the real cases in Brazil should be, at least, seven times the published number.

Table 1 presents the rates for all states of Brazil, from which we can observe that Paraíba has the lowest reported rate 0.06 and while Roraima presents the highest reported rate 0.52. Indeed, Prado et al. (2020) found that Paraíba and Pernambuco had a low reporting rate comparing with other states.

Table 1
Reported rate estimates and 95% confidence interval (95% CI) for COVID-19 Brazilian data.

State	Rate ($\widehat{\eta}$)	Lower 95% CI	Upper 95% CI
Acre	0.14	0.12	0.17
Alagoas	0.10	0.09	0.12
Amapa	0.20	0.17	0.24
Amazonas	0.08	0.07	0.10
Bahia	0.20	0.17	0.24
Ceará	0.11	0.10	0.13
Distrito Feral	0.40	0.34	0.48
Espírito Santo	0.19	0.16	0.23
Goiás	0.19	0.17	0.24
Maranhão	0.11	0.09	0.13
Mato Grosso	0.22	0.19	0.27
Mato Grosso do Sul	0.24	0.21	0.30
Minas Gerais	0.19	0.16	0.23
Parā	0.10	0.08	0.12
Paraíba	0.06	0.06	0.08
Paraná	0.15	0.13	0.18
Pernambuco	0.07	0.06	0.09
Piauí	0.10	0.09	0.12
Rio de Janeiro	0.08	0.07	0.10
Rio Grande do Norte	0.14	0.12	0.17
Rio Grande do Sul	0.23	0.20	0.28
Rondônia	0.17	0.15	0.21
Roraima	0.52	0.44	0.63
Santa Catarina	0.30	0.25	0.36
São Paulo	0.09	0.07	0.10
Sergipe	0.12	0.10	0.14
Tocantins	0.17	0.15	0.21

5.2. Estimation: Dirichlet-Beta state-space model

For the adjustment of the Bayesian model, the *prioris* and hyper-parameters are specified:

- γ - We assume that the average infection period is equal to 15 days. Thus, the γ *a priori* belongs to lognormal distribution with mean of 0.07 and variance of 0.01.

$$\gamma \tilde{} \text{LogN}(-3.215, 1.112).$$

- The average infection period ρ comes directly from γ parameter, that is, $\rho = \frac{1}{\gamma}$.
- β - The reproduction number R_0 of the disease is estimated by the ratio $R_0 = \frac{\beta}{\gamma}$. We assume that R_0 *a priori* belongs to lognormal distribution with mean of 3 and variance of 9. Thus β values were obtained from $\beta = R_0\gamma$

$$R_0 \tilde{} \text{LogN}(0.752, 0.693)$$

- The *a priori* distributions for k, λ_I and λ_R and $\boldsymbol{\theta}_0$ were obtained according to Osthus et al. (2017), that is, $k \tilde{}$ Gamma$(2, 0.0001)$,

$$\lambda_I \tilde{} \text{Gamma}(2, 0.0001),$$

$$\lambda_R \tilde{} \text{Gamma}(2, 0.0001),$$

$$\boldsymbol{\theta}_0 \tilde{} \text{Dirichlet}(0.99, 0.001, 0.001).$$

The estimates from *a posteriori* distributions for R_0, β, γ, k, λ_I and λ_R were obtained through MCMC methods, specifically Gibbs sampling (Geman & Geman, 1984). To execute the sampling procedure, we used the R programming language (R Core Team, 2020), with rjags package (Plummer, 2019).

The total number of iterations considered, as well as the discard (burn-in) and the minimum distance between one iteration to another (thin) were obtained through the criterion of Raftery and Lewis (1992) in the analysis of a pilot sample with 10,000 iterations.

The convergence diagnosis of the MCMC procedure was verified using the Geweke (Geweke, 1992) and Heidelberger and Welch (Heidelberger & Welch, 1983) criteria, which are available in the coda package (Plummer et al., 2006).

Table 2 shows the p-values from Geweke, and Heidelberger and Welch convergence diagnostics, from which we conclude that chains reached convergence for all parameters (p-value > 0.05). The inference was made by considering the reported rate estimate in Brazil, $\widehat{\eta} = 0.07$, a chain of 300,000 interactions was generated, with a burn-in of 10,000 and a thin of 300, resulting in a final sample of 1450 values.

The parameter estimates are shown in Table 3, in which $\widehat{\beta} = 0.1125$ and $\widehat{\gamma} = 0.0308$ are the major characteristics from SIR model and $\widehat{k} = 52,535.34$, $\widehat{\lambda_I} = 217,894.30$ and $\widehat{\lambda_R} = 223,431.60$ express the magnitude of the process error for the unknown proportions ($\boldsymbol{\theta}$) in Bayesian approach.

Table 2
P-values for Geweke, and Heidelberger and Welch convergence diagnostics.

parameter	Geweke	Heidelberger and Welch
R_0	0.7222	0.2026
β	0.8900	0.1898
γ	0.8210	0.2611
κ	0.2965	0.1455
λ_I	0.8205	0.2462
λ_R	0.1118	0.2568

Table 3
Point estimates and 95% Credible Interval.

parameter	Mode	95% Credible Interval	
		lower	upper
γ	0.0308	0.0272	0.0343
β	0.1125	0.1067	0.1201
R_0	3.6243	3.3528	4.0335
ρ	32.1667	29.1268	36.7576
κ	52535.34	38384.26	71244.52
λ_I	217894.30	148822.20	310111.60
λ_R	223431.60	147997.80	320880.00

The inference results show that $\widehat{R_0} = 3.6243$ which expresses a high reproductive rate of the virus. Also, $\widehat{\rho} = 32.1667$ days shows that the time for virus infection is very close to one month period.

Furthermore, Fig. B.6 shows the charts of the estimated *a posteriori* densities for the parameters β, γ, R_0, ρ, λ_I, and λ_R, from which we conclude that the curves have a symmetrical shape around its modes.

Using the parameter estimates from Table 3 and the latent proportion ($\boldsymbol{\theta}$), we reached information about the peak from SIR curve for the COVID-19 transmission in Brazil, that is the time when the proportion of infected individuals reaches its maximum. The peak estimate is June 18th, 2020, occurring between June 12th and June 22nd, 2020 and it is shown in Fig. 1.

Finally, to evaluate the goodness of fit from the SIR model, Fig. B.5 on Appendix B shows the chart with observed COVID-19 cases and the cases estimated by the SIR model, along with the corresponding highest probability density (HPD)interval. The cases fitted by the SIR model accounts for under-reporting and the HPD intervals are obtained through Chen and Shao algorithm (Chen & Shao, 1999). From Fig. B.5 we can observe that the SIR estimates for reported cases are close to observed cases reported by the Brazilian Ministry of Health.

5.3. Estimation: Dirichlet-Beta state-space model considering CFR unknown

Additionally, we conducted a Bayesian analysis considering the case fatality rate (CFR) being unknown and assumed to have a uniform distribution. To achieve this goal, we focus our attention on the η parameter, which is the under-reporting rate. It means to say that CFR varies according to the uniform distribution, it is equivalent to saying that the η varies according to the uniform distribution:

$$\eta \tilde{} U(0.0579, 0.0821)$$

The results are shown in Table B5 on Appendix B. We can observe that the point estimates for the model parameters are very similar to the estimates from Table 3. However, the 95% credible intervals for β and R_0 are wider than in Table 3 due to the flexibility that is given to η parameter. Moreover, the peak estimate under this more flexible Bayesian approach is June 16th, 2020, which is very close to the first model when CFR is supposed to be fixed.

6. Simulation study

Concerning to evaluate the effect of the notification rate on the model's estimates, a simulation study was carried out. The model was estimated considering COVID-19 data in Brazil, assuming a reporting rate between 0.05 and 1.00, varying every 0.05. Aiming the practical point of view, we conduct a simulation study to investigate the effects of under-reporting in the parameters of the SIR model and how it impacts on the pandemic curve behavior. For each value of η, a chain of 300,000 interactions was generated, with a burn-in of 10,000 and a thin of 300.

Fig. 2 shows the point estimates and 95% credible intervals for β and γ versus the reported rate values. It can be observed that as reported rate increases, β estimate becomes lower, which means that the infectious contact rate is underestimated when under-reporting is ignored. Additionally, the removal rate γ remains almost constant when the reported rate increases, which means that it is not influenced by the rates.

The graphics with the point estimates and 95% credible intervals for R_0 and infection period ρ versus the reported rates are shown in Fig. 3, from which we observe that R_0 decreases as the reported rate increases and ρ keeps roughly invariant, then we can conclude that the reproduction rate and infection period can be underestimated when under-reporting is ignored, affording an unreal impression on a tiny mean number of secondary individuals that a primary individual can infect, when in fact it is large.

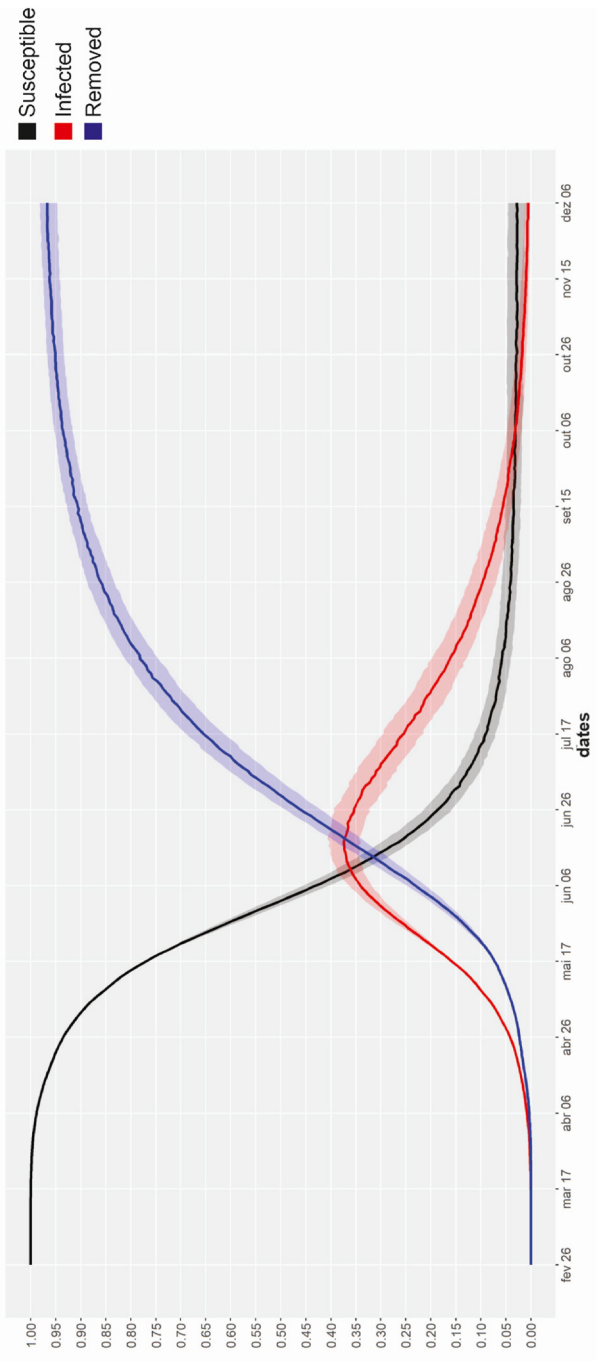

Fig. 1. Estimated SIR curves for COVID-19 Brazilian data from February 26th to May 20th, 2020.

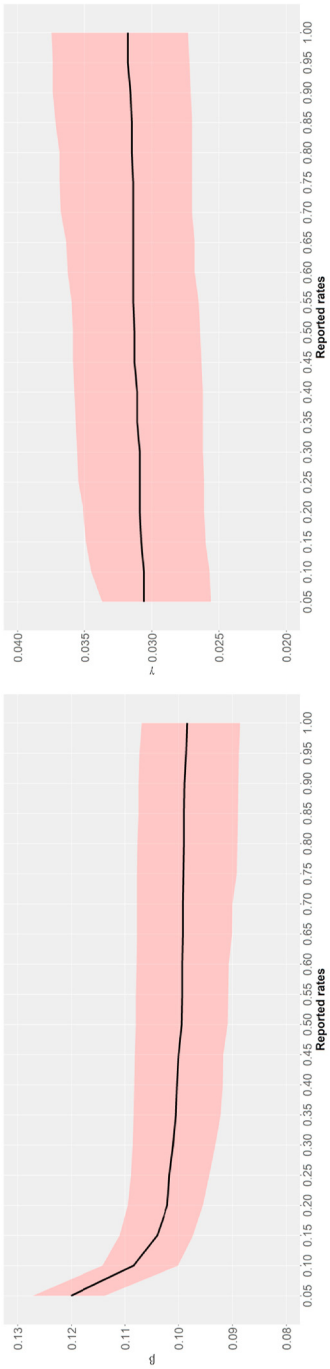

Fig. 2. Point estimates and 95% credible intervals for β and γ versus reported rates.

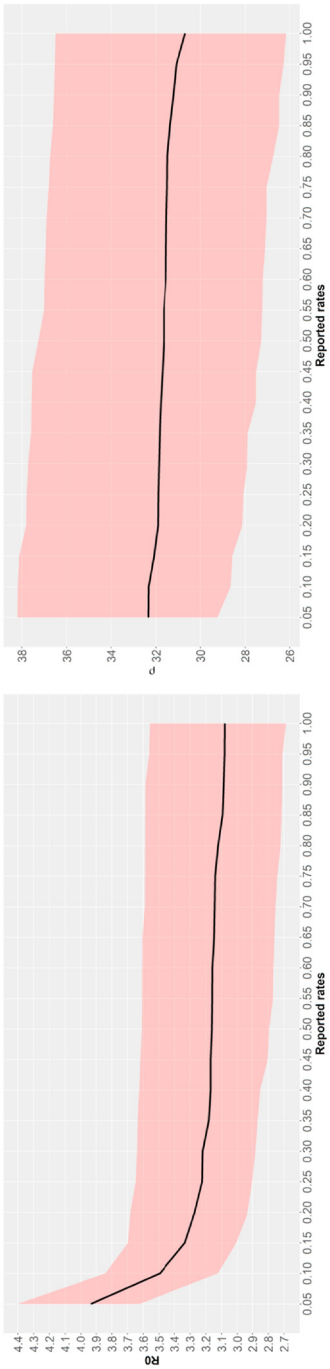

Fig. 3. Point estimates and 95% credible intervals for R_0 and infection period versus reported rates.

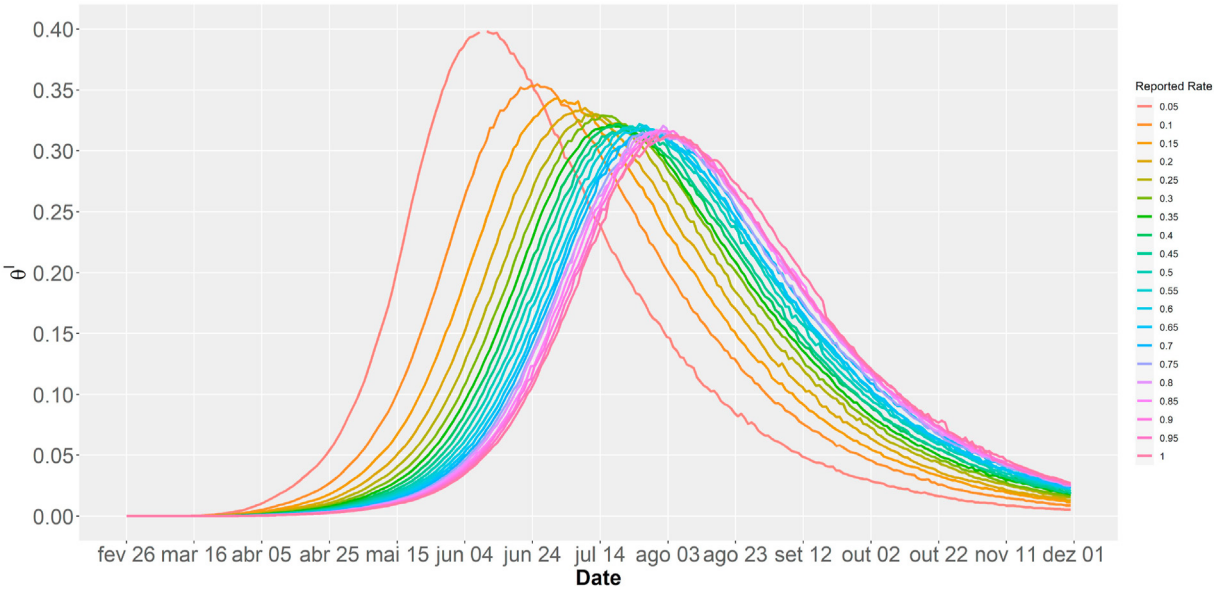

Fig. 4. Estimated SIR curves versus reported rate for COVID-19 Brazilian data.

Table 4
DIC values for COVID-19 Brazilian data.

rate(η)	DIC	rate(η)	DIC
0.05	3197.96	0.55	3260.65
0.10	3183.47	0.60	3266.44
0.15	3208.17	0.65	3272.19
0.20	3237.96	0.70	3277.76
0.25	3241.55	0.75	3279.27
0.30	3248.51	0.80	3283.36
0.35	3249.57	0.85	3295.20
0.40	3258.51	0.90	3296.90
0.45	3259.62	0.95	3306.84
0.50	3259.71	1.00	3307.88

Fig. 4 shows the estimated SIR curves for COVID-19 versus reported rate, from which we observe that the lower the reported rate, the earlier the peak is reached with a higher proportion of infected individuals. It is also observed that the contagion curves become similar to each other as the reported rates increase. These results reveal that the peak estimate of the COVID-19 transmission curve in Brazil is compromised when the presence of under-reporting is ignored.

Finally, Table 4 presents the deviance information criterion (DIC) (Spiegelhalter et al., 2002), which indicates the SIR model with the reported rate of 0.1 as the best one that fitted the simulated data, since its DIC value is the lowest. These results suggest that the notification rate is very low.

7. Concluding remarks

In this paper, we show that the method of adjusting cases by delay can be used to determine the reported rate of COVID-19 cases. Thus, it was possible that the rate of cases reported in Brazil is 0.07 and thus underestimates the real spreading of pandemic in the country.

Thus we proposed a SIR model with correction for under-reporting. The Bayesian approach is a feasible way to deal with the parameters inherent to the SIR model.

The methods reached convergence in the application with the Brazilian COVID-19 data set. Thus, a reproductive rate of 3.6243 was obtained, indicating that the epidemic is still booming in Brazil.

The simulation study revealed that the parameters estimates from the SIR model and the peak estimate which is a concern of several researchers and health authorities are sensitive to reporting rates. Future work may include considering the use of

extended SIR models like the SEIR model (with the compartments of susceptible, exposed, infected, and removed individuals), and further, consider different scenarios of isolation and quarantine for the strategy of the COVID-19 transmission control.

conflicts of interest

None.

Appendix A. Numerical Solution for SIR model

Let $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ be the Runge-Kutta RK4 approximation to the SIR model. Thus,

$$f(\boldsymbol{\theta}_{t-1}, \beta, \gamma) = \begin{bmatrix} \theta^S_{t-1} + \frac{1}{6}\left(k^{S_1}_{t-1} + 2k^{S_2}_{t-1} + 2k^{S_3}_{t-1} + k^{S_4}_{t-1}\right) \\ \theta^I_{t-1} + \frac{1}{6}\left(k^{I_1}_{t-1} + 2k^{I_2}_{t-1} + 2k^{I_3}_{t-1} + k^{I_4}_{t-1}\right) \\ \theta^R_{t-1} + \frac{1}{6}\left(k^{R_1}_{t-1} + 2k^{R_2}_{t-1} + 2k^{R_3}_{t-1} + k^{R_4}_{t-1}\right) \end{bmatrix}$$

where

$$k^{S_1}_t = -\beta\theta^S_t\theta^I_t$$

$$k^{S_2}_t = -\beta\left(\theta^S_t + \frac{1}{2}k^{S_1}_t\right)\left(\theta^I_t + \frac{1}{2}k^{I_1}_t\right)$$

$$k^{S_3}_t = -\beta\left(\theta^S_t + \frac{1}{2}k^{S_2}_t\right)\left(\theta^I_t + \frac{1}{2}k^{I_2}_t\right)$$

$$k^{S_4}_t = -\beta\left(\theta^S_t + k^{S_2}_t\right)\left(\theta^I_t + k^{I_2}_t\right)$$

$$k^{I_1}_t = \beta\theta^S_t\theta^I_t - \gamma\theta^I_t$$

$$k^{I_2}_t = \beta\left(\theta^S_t + \frac{1}{2}k^{S_1}_t\right)\left(\theta^I_t + \frac{1}{2}k^{I_1}_t\right) - \gamma\left(\theta^I_t + \frac{1}{2}k^{I_1}_t\right)$$

$$k^{I_3}_t = \beta\left(\theta^S_t + \frac{1}{2}k^{S_2}_t\right)\left(\theta^I_t + \frac{1}{2}k^{I_2}_t\right) - \gamma\left(\theta^I_t + \frac{1}{2}k^{I_2}_t\right)$$

$$k^{I_4}_t = \beta\left(\theta^S_t + k^{S_3}_t\right)\left(\theta^I_t + k^{I_3}_t\right) - \gamma\left(\theta^I_t + k^{I_3}_t\right)$$

$$k^{R_1}_t = -\gamma\theta^I_t$$

$$k^{R_2}_t = -\gamma\left(\theta^I_t + \frac{1}{2}k^{I_1}_t\right)$$

$$k^{R_3}_t = -\gamma\left(\theta^I_t + \frac{1}{2}k^{I_2}_t\right)$$

$$k^{R_4}_t = -\gamma\left(\theta^I_t + k^{I_3}_t\right)$$

Appendix B. Figures and tables

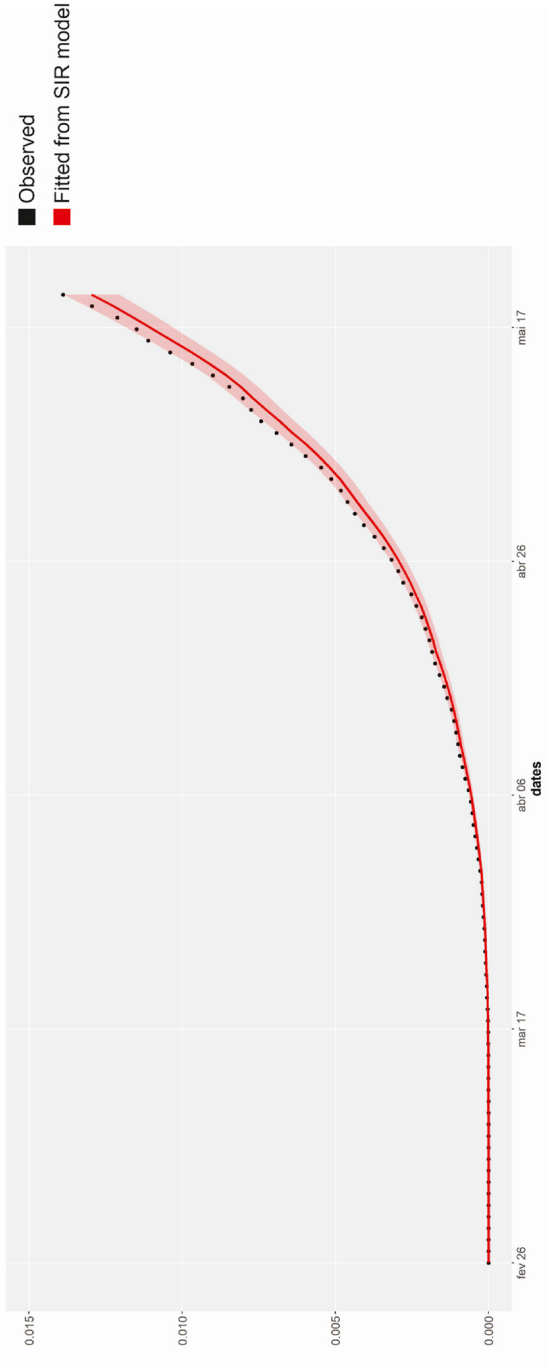

Fig. B.6. Estimated a posteriori densities for the parameters.

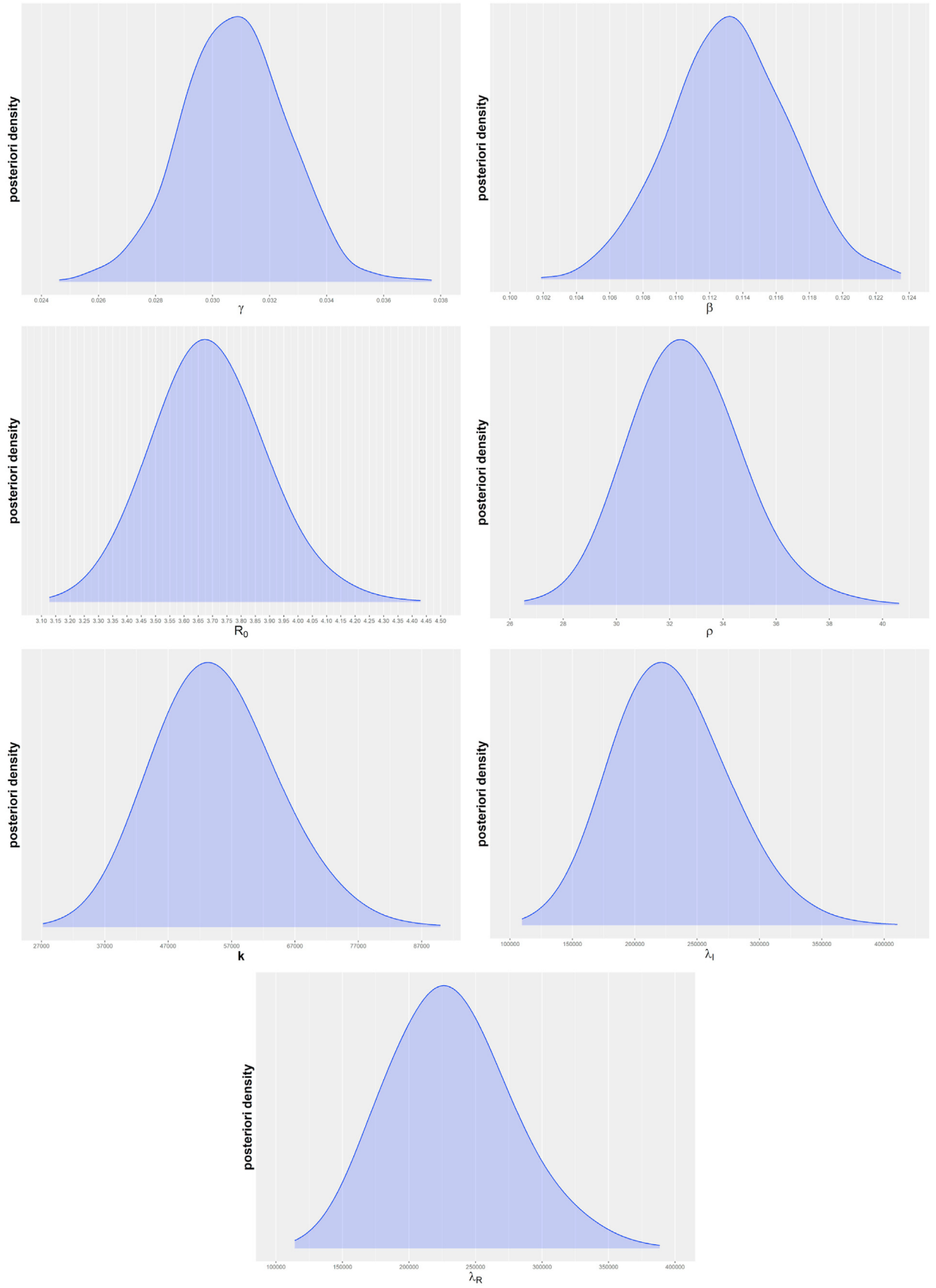

Fig. B.5. Comparison of observed COVID-19 cases and the cases estimated by SIR model with a correction to account for under-reporting.

Table B.5
Point estimates and 95% Credible Interval considering CFR unknown.

parameter	Mode	95% Credible Interval	
		lower	upper
Γ	0.0309	0.0270	0.0341
B	0.1153	0.1078	0.1220
R_0	3.7320	3.3600	4.1200
P	32.300	28.9280	36.4790
K	48478.87	33976.36	67350.45
λ_I	219245.20	142043.10	308019.10
λ_R	219204.00	149087.40	312584.00

References

Bjørnstad, O. N. (2018). *Epidemics: Models and data using R*. Springer.

Chen, M.-H., & Shao, Q.-M. (1999). Monte Carlo estimation of bayesian credible and hpd intervals. *Journal of Computational & Graphical Statistics, 8*, 69–92. http://www.jstor.org/stable/1390921.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician, 49*, 327–335. http://www.jstor.org/stable/2684568.

Cotta, R. M., Naveira-Cotta, C. P., & Magal, P. (2020). *Parametric identification and public health measures influence on the covid-19 epidemic evolution in Brazil. medRxiv.* arXiv:https://www.medrxiv.org/content/early/2020/05/12/2020.03.31.20049130.full.pdf.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*, 799–815.

Gelman, A., Robert, C., Chopin, N., & Rousseau, J. (1995). *Bayesian data analysis*.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Recognition, 6*, 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian statistics* (pp. 169–193). University Press.

Guan, W-j, Ni, Z-y, Hu, Y., et al. (2020). *Clinical characteristics of coronavirus disease 2019 in China, 382* pp. 1708–1720). New England Journal of Medicine, 18, In this issue.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31*, 1109–1144.

Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115*, 700–721.

Kim, D.-H., Choe, Y. J., & Jeong, J.-Y. (2020). Understanding and interpretation of case fatality rate of coronavirus disease 2019. *Journal of Korean Medical Science, 35*.

Lenzer, J. (2020). *Covid-19: US gives emergency approval to hydroxychloroquine despite lack of evidence* (369:m1335). BMJ Publishing Group Ltd. In this issue.

Linton, NM Y. Y.e. a., & Kobayashi, T. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine, 9*, 538.

Mathews, J. H. (1992). *Numerical methods for mathematics, science and engineering* (2nd ed.). Englewood Cliffs: Prentice-Hall International.

Nishiura, H., Klinkenberg, D., Roberts, M., & Heesterbeek, J. A. P. (2009). Early epidemiological assessment of the virulence of emerging infectious diseases: A case study of an influenza pandemic. *PLoS One, 4*.

de Oliveira, G. L., Loschi, R. H., & Assunção, R. M. (2017). A random-censoring Poisson model for underreported data. *Statistics in Medicine, 36*, 4873–4892.

Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D., & Del Valle, S. Y. (2017). Forecasting seasonal influenza with a state-space sir model. *Annals of Applied Statistics, 11*. https://doi.org/10.1214/16-AOAS1000

Paulino, C. D., Amaral Turkman, M. A., Murteira, B., & Silva, G. L. (2018). *Estatística bayesiana* (2nd ed.). Lisboa: Fundação Calouste Gulbenkian.

Plummer, M. (2019). rjags: Bayesian graphical models using MCMC. https://CRAN.R-project.org/package=rjags r package version 4-10.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News, 6*, 7–11. https://journal.r-project.org/archive/.

Prado, M., Bastos, L., Batista, A., Antunes, B., Baião, F., Maçaira, P., Hamacher, S., & Bozza, F. (2020). *Análise de subnotificação do número de casos confirmados da COVID-19 no Brasil*. Technical Report NOIS.

R Core Team. (2020). *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing Vienna. https://www.R-project.org/.

Raftery, A. E., & Lewis, S. M. (1992). [practical Markov chain Monte Carlo]: Comment: One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science, 7*, 493–497. https://doi.org/10.1214/ss/1177011143

Ribeiro, L. C., & Bernardes, A. T. (2020). *Estimate of underreporting of COVID-19 in Brazil by acute respiratory syndrome hospitalization reports*. UFOP: Technical Report UFMG.

Rodríguez-Morales, A. J., MacGregor, K., Kanagarajah, S., Patel, D., & Schlagenhauf, P. (2020). Going global–travel and the 2019 novel coronavirus. *Travel Medicine and Infectious Disease, 33*(101578), 101578. https://doi.org/10.1016/j.tmaid.2020.101578.

Russel, T., Hellewell, J., Abbot, S., et al. (2020). Using a delay-adjusted case fatality ratio to estimate under-reporting. *Available at the Centre for Mathematical Modelling of Infectious Diseases Repository*.

Song, P. X., Wang, L., Zhou, Y., He, J., Zhu, B., Wang, F., Tang, L., & Eisenberg, M. (2020). *An epidemiological forecast model and software assessing interventions on covid-19 epidemic in China. medRxiv.* https://doi.org/10.1101/2020.02.29.20029421. arXiv:https://www.medrxiv.org/content/early/2020/03/03/2020.02.29.20029421.full.pdf https://www.medrxiv.org/content/early/2020/03/03/2020.02.29.20029421.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583–639.

Stoner, O., Economou, T., & Drummond Marques da Silva, G. (2019). A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association*, (1–17).

WHO. (2020). WHO announces COVID-19 outbreak a pandemic. http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic.