Infectious Disease Modelling 7 (2022) 317–332

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

# Estimate of the rate of unreported COVID-19 cases during the first outbreak in Rio de Janeiro

M.S. Aronna [a], R. Guglielmi [b, *], L.M. Moschen [a]

[a] Escola de Matemática Aplicada - EMAp, FGV, Rio de Janeiro, RJ, Brazil
[b] Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada

ARTICLE INFO

*Article history:*
Received 4 October 2021
Received in revised form 18 May 2022
Accepted 5 June 2022
Available online 22 June 2022
Handling editor: Dr Lou Yijun

*Keywords:*
COVID-19 model
Parameter estimation
Identifiability
Least squares
B-splines
Bootstrap method

ABSTRACT

In this work we fit an epidemiological model SEIAQR (*Susceptible - Exposed - Infectious - Asymptomatic - Quarantined - Removed*) to the data of the first COVID-19 outbreak in Rio de Janeiro, Brazil. Particular emphasis is given to the unreported rate, that is, the proportion of infected individuals that is not detected by the health system. The evaluation of the parameters of the model is based on a combination of error-weighted least squares method and appropriate B-splines. The structural and practical identifiability is analyzed to support the feasibility and robustness of the parameters' estimation. We use the Bootstrap method to quantify the uncertainty of the estimates. For the outbreak of March–July 2020 in Rio de Janeiro, we estimate about 90% of unreported cases, with a 95% confidence interval (85%, 93%).



## 1. Introduction

In late December 2019, health professionals in the city of Wuhan (Hubei, China) identified several cases of pneumonia (The 2019 nCoV Outbreak Joint Field Epidemiology Investigation Team and Q. Li, 2020) caused by a new coronavirus, which was named SARS-CoV-2. The disease induced by SARS-CoV-2, called COVID-19, rapidly spread around the world. Most COVID-19 cases are asymptomatic patients or with mild symptoms, but in more severe cases the disease may progress to viral pneumonia and multi-organ failure (World Health Organization, 2020) and can lead to hospitalization and death.

Brazil declared the disease as a public health emergency in February 2020, before the first COVID-19 cases were reported in Brazil. With the aim of protecting the population, several measures were implemented at different administrative levels. In March 2020, the state of Rio de Janeiro declared a public health emergency and ordered to avoid gatherings, closing down schools, restaurants, and theaters, and restricting access to beaches, shopping centers, and non-essential commerce (Dantas et al., 2020). Nevertheless, these measures could not prevent the outbreak of the virus in the region, fueled by the large portion of asymptomatic individuals and by the long incubation period of the virus (Li et al., 2020). The objective of this work is to provide a quantitative estimate of crucial parameters related to the COVID-19 outbreak in the municipality of Rio de Janeiro during the period March–July 2020. The analysis is based on the fitting of the SEIAQR compartmental model

* Corresponding author.
*E-mail addresses:* soledad.aronna@fgv.br (M.S. Aronna), roberto.guglielmi@uwaterloo.ca (R. Guglielmi), lucas.moschen@fgv.edu.br (L.M. Moschen).
Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2022.06.001

introduced in (Aronna et al., 2021) to real data retrieved from the public health agency (Municipal Health Department, City Hall of Rio de Janeiro, 2021). This epidemiological model takes into account isolation, quarantine of confirmed cases, and testing of asymptomatic individuals as non-pharmaceutical strategies to contain the spread of the virus among the population (more details in Section 2.1). The main purpose of the study is to estimate the rate of unreported cases in the municipality of Rio de Janeiro. This is a crucial step to gauge the real extent and impact of the disease on the population, since most of the infections do not result in severe symptoms and are therefore likely to remain undetected by the health department (Canzian, 2020).

The question of analyzing the data from Brazil or some of its cities or states has been addressed by some previous works. In Crokidakis (2020b), the authors fitted an SIQR (Susceptible - Infectious - Quarantined - Recovered) model to the Brazilian national data, while Bastos and Cajueiro (2020) used a SIRASD (Susceptible - Infectious - Recovered - Asymptomatic - Symptomatic - Death) model. The work He et al. (2021) was also dedicated to the issue of inferring how many cases were not detected by the health system. More precisely, they fit an SEIHDR (Susceptible - Exposed - Infectious - Hospitalized - Death - Recovered) model using the mortality data of the city of Manaus and estimated the *attack rate* by October 2020, when a lethal second wave of COVID-19 hit the region (Sabino et al., 2021). In Ritto et al. (2021), the authors applied the Approximate Bayesian Computation (ABC) to calibrate parameters and quantify uncertainties in an SEIAHRD (Susceptible - Exposed - Infectious - Asymptomatic - Hospitalized - Recovered - Death) compartmental model using the data of hospitalization and deaths from the city of Rio de Janeiro. Other works dealing with estimation of parameters for Brazil or some Brazilian region are *e.g.* Morato et al. (2020); Crokidakis (2020a); Portal COVID-19 Brasil (2021). A more quantitative comparison with these and other works is given in the Discussion Section.

The period under consideration in this paper, from March to July 2020, only partially overlaps with the periods considered in the above-mentioned references. Moreover, in this paper we apply a wealth of statistical methods that not only provide an excellent fitting between the data and the model (see Figs. 5 and 6), but also allows to derive a dynamical estimate of the ratio of unreported cases (see Fig. 7), its sensitivity to the parameters of the system (see Table 7), an approximation of time-dependent transmission and mortality rates (see Fig. 9), and the confidence interval of the time-dependent reproduction number (see Fig. 11).

The article is organized as follows: Section 2 collects different properties of the mathematical model; the structure of the data as retrieved for the city of Rio de Janeiro; and the chosen representation for the parameter estimation. Section 3 describes the notions of identifiability analyzed in the paper. Section 4 presents the results of the fitting of the data to the model and the resulting estimates of the parameters. Finally, the last two sections conclude the paper with some discussions and final comments.

### 1.1. Data availability

No new data are released as part of this research. This paper relies on publicly available datasets, with references provided in the text. The code required to reproduce our analysis is available in the GitHub repository (Moschen, 2021).

## 2. Material and methods

### 2.1. The epidemiological model

We recall the key features of the compartmental model from (Aronna et al., 2021). The population is split into the compartments $S$, $E$, $I$, $A$, $Q$, $R$, and $D$, summarized in Table 1. These compartments are linked in the following way (see Fig. 1): people in $S$ move to compartment $E$ when exposed to the virus. After a given latent period, they become infectious and thus pass to compartment $I$. At this stage, an individual in $I$ may either report symptoms and thus move (after testing) to the compartment $Q$, or result in an asymptomatic/paucisymptomatic infection and move to the compartment $A$. Individuals from $E$ and $A$ may move to $Q$ as a result of the testing strategy among the asymptomatic population, as described in more detail in the following. Finally, the compartment $R$ collects the recovered individuals from either $A$ or $Q$, whereas the compartment $D$ counts the COVID-19-related deaths.

**Table 1**
List of state variables of the system (1). The associated flow diagram can be found in Fig. 1.

| Compartment | Description |
|---|---|
| $S$ | susceptible |
| $E$ | exposed |
| $I$ | infectious |
| $A$ | asymptomatic and infectious |
| $Q$ | infected in quarantine (including hospitalized) |
| $R$ | recovered |
| $T$ | total positive tests |
| $D$ | death |

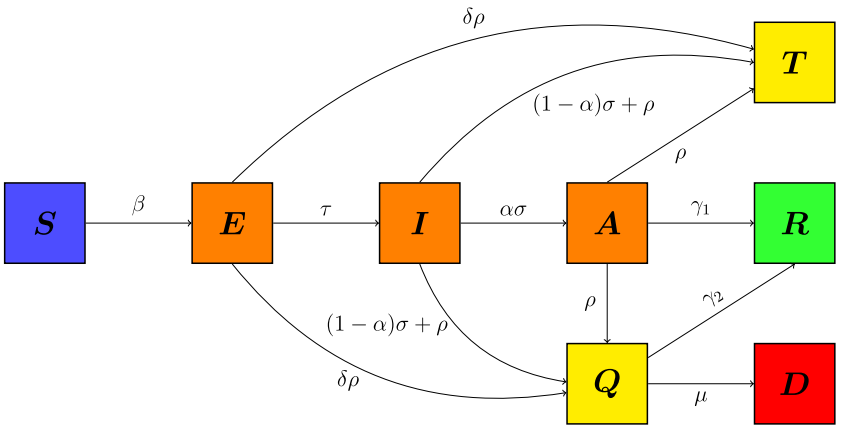


**Fig. 1.** Flow chart for the state variables from system (1), the flux direction between compartments and related parameters. People in $S$ move to compartment $E$ when exposed to the virus. After a given latent period, they become infectious and thus pass to $I$. An individual in $I$ may either report symptoms and thus move (after testing) to the compartment $Q$, or result in an asymptomatic infection and move to $A$. Individuals from $E$ and $A$ may move to $Q$ as a result of the testing strategy among the asymptomatic population. Finally, $R$ collects the recovered individuals from either $A$ or $Q$, whereas the compartment $D$ counts the deaths and $T$ the total positive tests.

By normalizing the total population to 1, the value of each variable $S$, $E$, $I$, $A$, $Q$, $R$, and $D$ represents the proportion of that given compartment in the total population. Moreover, we neglect the birth and natural death rates, given the limited time horizon chosen for the data fitting. The dynamics is described by the following system of differential equations:

$$\begin{aligned}
\dot{E} &= \beta(t)S(I+A) - \rho\delta E - \tau E \\
\dot{I} &= \tau E - \sigma I - \rho I \\
\dot{A} &= \sigma\alpha I - \rho A - \gamma_1 A \\
\dot{Q} &= \sigma(1-\alpha)I + \rho(\delta E + I + A) - \gamma_2 Q - \mu(t)Q \\
\dot{S} &= -\beta(t)S(I+A) \\
\dot{R} &= \gamma_1 A + \gamma_2 Q \\
\dot{D} &= \mu(t)Q \\
\dot{T} &= \sigma(1-\alpha)I + \rho(\delta E + I + A).
\end{aligned} \tag{1}$$

This is a simplified version of the model studied in (Aronna et al., 2021), that takes into account also the lockdown for non-essential workers, thus dividing each compartment $S$, $E$, $I$ and $A$ into two subgroups. In system (1), instead, we assume that the whole population is subject, in average, to the same mobility restrictions. The parameters $\tau$, $\sigma$, $\omega$, $\gamma_1$, $\gamma_2$, and $\mu$ related to the pathogen and induced by the disease are described in Table 2. The function $\beta$ is the *effective contact rate* at time $t$, which depends on the average contact rate – directly affected by social distancing, public policies, use of Personal Protective Equipment (PPE), etc. – and the transmissibility of the virus – probability of infection given a contact between an infected and a susceptible individual. A constant of particular interest in our analysis is the parameter $\alpha \in (0, 1)$, which represents the proportion of asymptomatic infectious individuals. Among these cases, only those found positive through either testing induced by contact tracing or simply random testing are reported to the health system. All the others will remain unreported, being paucisymptomatic (but infectious) cases. In view of the scarcity of testing kits during the first outbreak in Rio de Janeiro,

**Table 2**
Summary of the model parameters from system (1). The parameters $\tau$, $\sigma$, $\omega$, $\gamma_1$ and $\gamma_2$ are related to SARS-CoV-2 virus, while $\mu(t)$, $\beta(t)$ and $\alpha$ to both COVID-19 and the non-pharmaceutical interventions. Finally, $\rho$ and $\delta$ correspond to the testing strategy and the tests' sensitivity, respectively.

| Par. | Description |
|---|---|
| $\tau^{-1}$ | latent period, from exposure to infectiousness |
| $\sigma^{-1}$ | time from infectiousness to possible symptoms onset |
| $\omega^{-1}$ | incubation period (i.e. $\omega^{-1} = \tau^{-1} + \sigma^{-1}$) |
| $\gamma_1$ | recovery rate for paucisymptomatic individuals |
| $\gamma_2$ | recovery rate for detected positive cases |
| $\mu(t)$ | mortality rate among detected positive cases |
| $\beta(t)$ | effective contact rate at time $t$ |
| $\alpha$ | proportion of asymptomatic infectious cases |
| $\rho$ | rate of testing among asymptomatic or paucisymptomatic cases |
| $\delta$ | probability of positive test for exposed and not infectious cases |

we consider the parameter $\alpha$ as a proxy for the *unreported rate* of infections. It is clear that estimating such parameter $\alpha$ is crucial to assess the effective size of the epidemic since such unreported cases may fuel the ongoing outbreak. Moreover, having an approximation of the size of the recovered population is also an essential information for epidemiological management and the comprehension of the virus behaviour. At the same time, estimating $\alpha$ may be very difficult, since it accounts for those cases not detected by the radar of the health system (Nogrady, 2020).

As mentioned above, the unreported rate is strictly related to the availability of testing kits and to the effectiveness of tracing, tracking, and testing in place during the outbreak. Such measures are described in model (1) by both the detection rate $1-\alpha$ and the parameter $\rho$, the latter representing the *rate of testing among asymptomatic or paucisymptomatic individuals*. We point out that the value of the unreported rate $\alpha$ is not sensitive to small variations of values of $\rho$ (see Section 2.3.1).

For clarity of exposition, in model (1) we do not describe important features such as the *sensitivity* and the *specificity* of the testing kits, although these attributes may play a crucial role in consideration of new variants of the virus (see (Aronna et al., 2021, Remark 2.3) for more details). With this premise, in our model we assume that a person in the compartments $I$ or $A$ will always test positive, in $S$ always negative, and in $E$ positive with a probability $\delta \in (0,1)$. Finally, the variable $T$ in (1) acts as a counter for the total positive tests.

### 2.1.1. The basic reproduction number

Assuming a constant contact rate $\beta$, the basic reproduction number $\mathcal{R}_0$ associated with the model (1) is given by the relation

$$\mathcal{R}_0 = \frac{1}{2}\left(\varphi + \sqrt{\varphi^2 + \frac{4\sigma\alpha}{\rho+\gamma_1}\varphi}\right), \tag{2}$$

where

$$\varphi = \frac{\beta\tau}{(\rho\delta+\tau)(\sigma+\rho)}$$

(see (Aronna et al., 2021)). However, as the epidemic evolves, the recovered and immune portion of the population becomes more relevant, impacting the force of the spread. For this reason we introduce the effective (time-dependent) reproduction number $\mathcal{R}_t$, which decreases as the susceptible population $S(t)$ decreases, and it is expressed by the same relation (2) of $\mathcal{R}_0$ with $\varphi$ replaced by

$$\varphi(t) = \frac{\beta(t)\tau S(t)}{(\rho\delta+\tau)(\sigma+\rho)}.$$

## 2.2. Data description

We retrieve data from the public health agency (Municipal Health Department, City Hall of Rio de Janeiro, 2021), collecting information on COVID-19 confirmed cases and deaths in the municipality of Rio de Janeiro from March 01, 2020 to July 31, 2020. The format of the data is displayed in Table 3.

The column Symptom onset date denotes the symptom onset date reported by the patient, and Outcome date marks the date on which the person recovers or dies (respectively recovered or death in the column Outcome), thus ceasing to be an active case. The column Outcome date was incomplete having about 3.7% of empty fields, so we were not able to use the recovery date as an input in our data fitting and missing death dates were imputed randomly according to the empirical distribution of the time difference between notification and evolution dates. From the Symptom onset date column, we obtain the curve of daily new cases grouping the individuals by date. The daily new COVID-19-related deaths curve comes from the Outcome data column, grouping the individuals with outcome death by date. Finally, we normalize these curves by the size of

**Table 3**
Extract of the raw data: each row represents a reported case. The displayed information shows the patient's neighborhood, gender, age group and ethnicity, together with the dates of notification, symptom onset and outcome. With outcome, it means the current state of the individual, the possible states being: recovered, death or active.

| Notification date | Symptom onset date | Neighborhood | Gender | Age group | Outcome | Outcome date | Ethnicity |
|---|---|---|---|---|---|---|---|
| 05/18/20 | 05/03/20 | PACIENCIA | M | 50–59 | death | 05/17/20 | Black |
| 04/25/20 | 04/02/20 | BARRA DA TIJUCA | M | 80–89 | death | 05/01/20 | White |
| 05/06/20 | 05/06/20 | CACHAMBI | M | 70–79 | death | 05/07/20 | Ignored |
| 06/12/20 | 06/02/20 | BARRA DA TIJUCA | M | 70–79 | death | 06/12/20 | White |
| 06/13/20 | 04/26/20 | MARECHAL HERMES | M | 60–69 | death | 06/16/20 | Ignored |

the population in Rio de Janeiro, which we assume to be 6.7 millions (IBGE, 2021). The resulting curves of daily new cases and new deaths are the blue curves in Figs. 2 and 3.

We point out that the paper focuses only on data from the municipality of Rio de Janeiro. It is clear that the metropolitan area of Rio de Janeiro has a considerably larger population, with many commuting to/from the municipality of Rio de Janeiro. From a practical point of view, considering data from the whole metropolitan area would require to deal with very different epidemiological data, different protocols, delays, and features collected. At the same time, we point out that mobility among different municipalities was restricted at the beginning of the pandemic Ávila (2020); Moreira and Pelegi (2020). For this reason, we assume that the limited effects of the population entering and exiting the city of Rio de Janeiro from surrounding ones balance out over the period under consideration.

*2.2.1. Smoothed and cumulative curves*

There is a weekly seasonality in the number of confirmed cases and deaths during the outbreak, with a negative deviation on weekends. This variation affects the model since it has no seasonal adjustment. For this reason, the data is smoothed using a centered *moving average* with 7 days. Thus, given a set of data $\{x_s\}_{1\le s\le n}$ of confirmed positive cases or deaths, we replace it by

$$\hat{x}_t = \frac{1}{2k+1}\sum_{s=-k}^{k} x_{t+s}, \tag{3}$$

where $2k+1=7$. Although this choice may seem arbitrary, it is directly related to the weekly periodicity of the data. In particular, in Fig. 2 we can observe that the first wave ends at the end of July, which justifies our choice to set the horizon of the data fitting on July 31st. Moreover, we calculate the curves of cumulative positive cases and deaths by evaluating

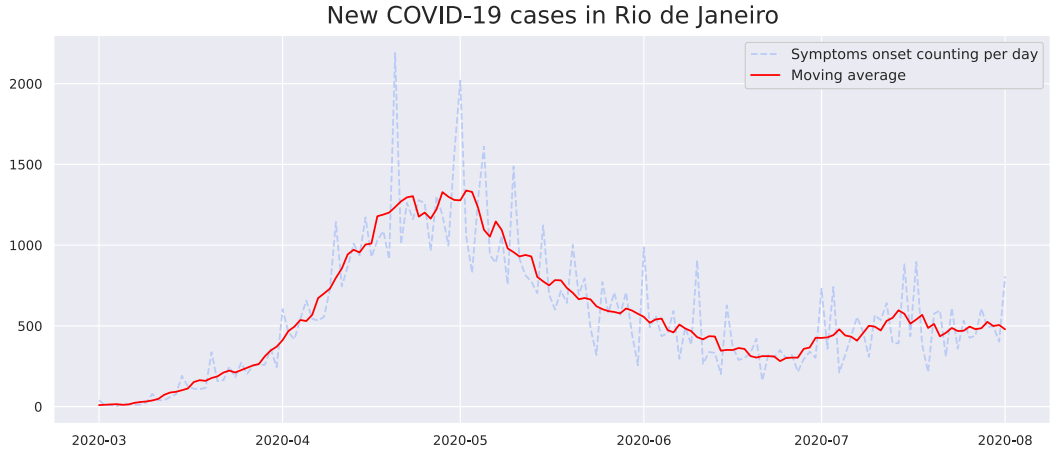


**Fig. 2.** The dashed blue line is the time series of reported daily new COVID-19 cases in the city of Rio de Janeiro between March and August 2020. The red curve is the 7-day centered moving average following equation (3) with $2k+1=7$.

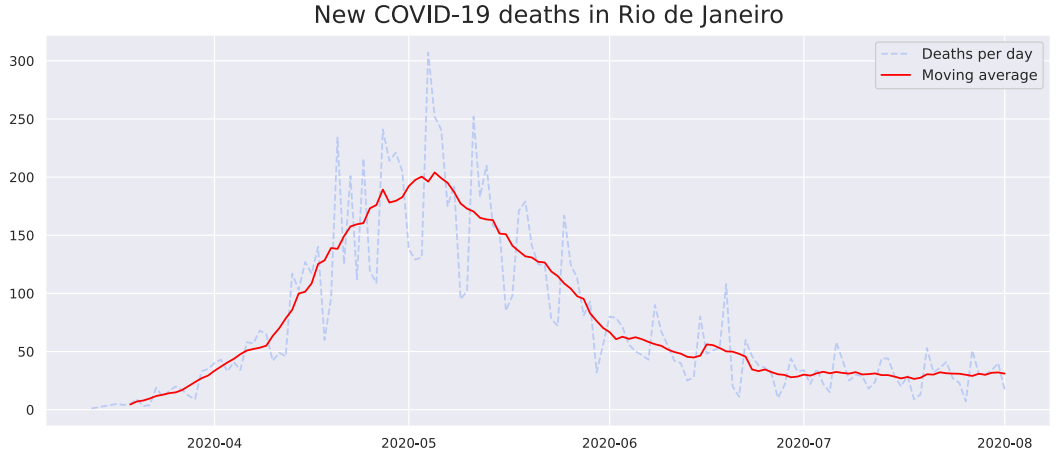


**Fig. 3.** The dashed blue line is the time series of reported daily COVID-19-related deaths in the city of Rio de Janeiro between March and August 2020. The red curve is the 7-day centered moving average following equation (3) with $2k+1=7$.

**Table 4**
Parameter estimates retrieved from the indicated literature and used in our model as fixed values.

| Par. | Value | Reference |
|---|---|---|
| $\omega^{-1}$ | 5.74 days | Rai et al. (2021) |
| $\tau^{-1}$ | 3.69 days | Li et al. (2020) |
| $\sigma^{-1}$ | $\omega^{-1} - \tau^{-1}$ | Li et al. (2020) |
| $\gamma_1^{-1}$ | 7.5 days | Byrne et al. (2020) |
| $\gamma_2^{-1}$ | 13.4 days | Byrne et al. (2020) |
| $\delta$ | 0.01 | Kucirka et al. (2020) |

$$y_t = \sum_{i=-14}^{t} \hat{x}_i, \quad (t = 0, \ldots, n).$$

The curves of cumulative cases, according to the self-reported symptoms onset, and cumulative deaths will be compared to the compartments $T$ and $D$ from model (1).

### 2.3. Choice and modeling of parameters

We split the set of parameters between a first group, consisting of the epidemiological constants $\tau$, $\sigma$, $\gamma_1$ and $\gamma_2$, together with the testing-related parameters $\delta$ and $\rho$, whose values we either retrieve from the literature (see Table 4) or deduce from the testing data (see Table 5), and a second group of parameters composed by $\beta(t)$, $\alpha$ and $\mu(t)$, which are estimated by a curve fitting procedure. The model used to represent the effective contact rate $\beta(t)$ and the mortality $\mu(t)$ is described in Section 2.3.2. We start by explaining in next subsection 2.3.1 how the value for $\rho$ is chosen.

#### 2.3.1. Estimation of $\rho$

In Table 2, we define $\rho$ as the rate of testing among asymptomatic or paucisymptomatic individuals, due to tracing, tracking and testing. We are aware that this kind of tests were marginal during the period under consideration, due to lack of testing kits and the lack of efficient tracing and tracking protocols. Therefore, in this section we provide only an upper estimate of the parameter $\rho$. We start from the testing data from the state of Rio de Janeiro (IBGE, 2020), summarized in Table 5: until the end of July 2020, 6.8% of the total population had been tested. Thus, assuming that testing started on March 15, 2020, the daily testing rate until the end of July is given by $6.8\%/135 \approx 0.05\%$. We thus deduce that the parameter $\rho$ is bounded by $\rho_{\max} = 0.0005$. However, the data does not allow us to discern the tests due to symptom onset from the testing of asymptomatic cases due to the effectiveness of the trace, track, and test strategy, which we assume to be only a minimal part of the total daily tests. For this reason, we assign to $\rho$ the value of $\rho = 10^{-4}$, which accounts for 20% of total daily tests. In any case, in Section 4 (see Table 7), we verify that our findings are robust and consistent for $\rho$ in the range of values $[0, \rho_{\max}]$.

#### 2.3.2. Representation of $\beta$ and $\mu$

The parameter $\beta$ in model (1) varies in time according to the different public policies in place in Rio de Janeiro in different periods (Official Journal of the State of Rio de Janeiro, 2020b; Estadão, 2020), and according to the compliance of the population with these measures. For this reason, we approximate $\beta$ by *B-splines* in the form

$$\beta(t) \approx \sum_{j=1}^{s} \beta_j B_{j,k}(t),$$

where $\beta_j$ are the coefficients to be estimated and $\{B_{j,k}(t)\}_{j=1}^{s}$ is the basis of functions of order $k$ (De Boor, 1978). A similar representation is used for the mortality rate $\mu(t)$,

$$\mu(t) \approx \sum_{j=1}^{r} \mu_j B_{j,k}(t),$$

which has to be understood as relative to the number of confirmed deaths. This reflects the fact that the relative mortality may vary in periods of distress for the health system and lack of testing kits. We therefore define the vector of $s + r + 1$ parameters to be estimated

**Table 5**
Data gathered from IBGE (2020). The first row represents the cumulative percentage of people tested for COVID-19 in the population of the state of Rio de Janeiro from July to October 2020, while the second indicates the cumulative percentage of positive tests. The data released by IBGE (2020) start from July 2020.

| | July | August | September | October |
|---|---|---|---|---|
| Percentage (%) | 6.8 | 8.6 | 10.2 | 11.9 |
| Positive rate (%) | 1.2 | 1.5 | 1.9 | 2.4 |

$$\theta = (\alpha, \beta_1, \ldots, \beta_s, \mu_1, \ldots, \mu_r). \tag{4}$$

We assume that the *knots*, i.e., the time points where the polynomials connect, are equally spaced. The number of knots is equal to $s + k + 1$ for $\beta$ and $r + k + 1$ for $\mu$, where $s$ and $r$ are the number of parameters for $\beta$ and $\mu$, respectively, and $k$ is the order of the B-spline.

To determine the order $k$ and the number of coefficients $s$, $r$, for the B-spline approximation, we use the *Akaike Information Criterion (AIC)* (Liang et al., 2010), which under the normality hypothesis of the errors is given by the formula

$$AIC = n\ln\left(\frac{RSS}{n}\right) + 2(s+r+1),$$

such that *RSS* is the sum of the squares of the model's residuals. For the numerical experiments in Section 4, we compare models with 3 and 4 coefficients, as models with less than 3 coefficients are unable to capture temporal variations over the timeframe of the outbreak, while models with $s$ and $r$ greater than 4 are computationally difficult to handle.

### 2.4. Parameter estimation

For the parameter estimation, we follow the methods applied in (Cao et al., 2012; Liang et al., 2010; Ramsay et al., 2007). The first 15 days of March were used to estimate the initial conditions (see Section 2.4.1) and thus set up the model starting from March 16th, when the first containment measures were imposed (Official Journal of the State of Rio de Janeiro, 2020a). Denoting by $\hat{x}_i^{(1)}$ and $\hat{x}_i^{(2)}$ the daily new cases and deaths, respectively, on the $i$-th day, where $i = 0$ and $i = n$ indicate March 16 and July 31, 2020, respectively, we assume that

$$\hat{x}_i^{(1)} = (T(i) - T(i-1)) + \epsilon_i^{(1)}, \quad i = -14, \ldots, n, \tag{5}$$

$$\hat{x}_i^{(2)} = (D(i) - D(i-1)) + \epsilon_i^{(2)}, \quad i = -14, \ldots, n, \tag{6}$$

where the sequences $\{\epsilon_i^{(1)}\}_{-14\le i\le n}$ and $\{\epsilon_i^{(2)}\}_{-14\le i\le n}$ are independent and normally distributed random variables with unknown variances $\sigma_k^2$, $k = 1, 2$. Summing up, we get

$$y_i^{(1)} = T(i) + \xi_i^{(1)}, \quad i = 0, \ldots, n, \tag{7}$$

$$y_i^{(2)} = D(i) + \xi_i^{(2)}, \quad i = 0, \ldots, n, \tag{8}$$

for the cumulative quantities, where $\xi_i^{(k)} = \sum_{j=-14}^{i} \epsilon_j^{(k)}$ for $k = 1, 2$. Therefore, for any $i \le j \le n$ and $k = 1, 2$, the *covariance matrix* is defined by

$$\mathrm{Cov}(\xi_i^{(k)}, \xi_j^{(k)}) = (i+15)\sigma_k^2. \tag{9}$$

At this point, we introduce the notations $\hat{T}(\theta)$ and $\hat{D}(\theta)$ for the numerical approximations of the functions $T$ and $D$, obtained by integrating the corresponding equations in (1) through the Runge-Kutta method and corresponding to the parameter vector $\theta$ introduced in (4).

#### 2.4.1. Initial values estimation

Considering $S \approx 1$ in the first 15 days of the epidemic, and focusing on the compartments $E$, $I$, $A$, and $T$, the system (1) reduces to

$$\begin{bmatrix} \dot{E} \\ \dot{I} \\ \dot{A} \\ \dot{T} \end{bmatrix} = \begin{bmatrix} -\tau & \beta & \beta & 0 \\ \tau & -\sigma & 0 & 0 \\ 0 & \sigma\alpha & -\gamma_1 & 0 \\ 0 & \sigma(1-\alpha) & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} E \\ I \\ A \\ T \end{bmatrix},$$

whose solution is a linear combination of exponential functions. The daily testing of asymptomatic individuals at the beginning of the pandemic is assumed to be negligible, thus $\rho \approx 0$, while the parameters $\gamma_1$, $\tau$, and $\sigma$ are taken from the existing literature (see Table 4). The parameters $\alpha$ and $\beta$ are fixed in this period. The value $T(-14)$ corresponds to the confirmed cases on March 2 and $T(-15)$ is set to zero. Therefore, the parameters to be estimated in this period are reduced to $\theta_0 = (\alpha, \beta, E(-14), I(-14), A(-14))$. Considering $\hat{T}(\theta_0)_i$ the approximation for $T$ as function of $\theta_0$ at day $i$, we aim to minimize the expression

$$\sum_{i=-14}^{0} w_i \left(y_i^{(1)} - \hat{T}(\theta_0)_i\right)^2,$$

where the weights $w_i = \frac{i+14}{14}$ are chosen to give less importance to the initial days. With these estimates, we obtain the values $(E(0), I(0), A(0))$. There are no reported COVID-19-related deaths before March 15, which allows to set $\mu = 0$ until that day and so $D(0) = 0$. Since $R(-14) = 0$, we get that $Q(-14) = T(-14)$, and thus we can integrate the curves $Q$ and $R$ to obtain the values $Q(0)$ and $R(0)$. Finally, we deduce $S(0) = 1 - E(0) - I(0) - A(0) - Q(0) - R(0) - D(0)$.

**Remark 2.1**. In order to analyze the sensitivity of the curves $\hat{T}$ and $\hat{D}$ with respect to the initial guess of the parameter $\theta_0$, we perform a series of random realizations and compare the values $\hat{T}(\theta_0)_n$ and $\hat{D}(\theta_0)_n$ for all guesses, where $n$ denotes the last day of the considered period.

#### 2.4.2. Curve fitting

We use the *weighted least squares* method to estimate the unknown parameters by the data of daily confirmed cases and deaths given by equations (7) and (8). This approach is based on the solution of a constrained nonlinear minimization problem. We consider the following objective functional

$$F(\theta) = (y^{(1)} - \hat{T}(\theta))^T \Sigma^{-1} (y^{(1)} - \hat{T}(\theta)) + \psi (y^{(2)} - \hat{D}(\theta))^T \Sigma^{-1} (y^{(2)} - \hat{D}(\theta)), \tag{10}$$

where $M^T$ denotes the transpose matrix of $M$, $\sigma_1^2\Sigma$ and $\sigma_2^2\Sigma$ are the covariance matrices given by equation (9), and $\psi$ is a weight proportional to the ratio of the variances $\sigma_1^2/\sigma_2^2$. We solve the minimization problem by means of the L-BFGS-B algorithm (Byrd et al., 1995), which combines the gradient projection method and the BFGS algorithm with optimized use of computational memory, implemented in the SciPy Python library (Virtanen et al., 2020).

### 2.5. Uncertainty quantification over the parameters

Uncertainty is intrinsic in the parameters' estimation, owing to a combination of different factors, such as the natural variability of the data, the measurement errors of the data collection, and the biases of the estimation method. In order to tackle this problem, we construct confidence intervals for each unknown parameter. We rely on the *Bootstrap method* (Efron & Tibshirani, 1986), which involves a constructive approach based on the data. Starting from a series $Y$, it generates replicated data $Y_1^*, \ldots, Y_N^*$ and performs the estimations for each one. The confidence interval for the value of interest is then given by the corresponding percentile of the $N$ replicated samples (Joshi et al., 2006).

As a consequence of the error structure presented in equation (5), we generate each replicated curve $T^{(B)}$ with the initial value $T^{(B)}(0) = y_0^{(1)}$ and satisfying $T^{(B)}(i+1) = T^{(B)}(i) + \epsilon_{i+1}$ for any $i \geq 1$, where $\epsilon_{i+1} \sim \mathcal{N}(\hat{x}_{i+1}^1, \hat{\sigma}_1^2)$ and $\hat{\sigma}_1^2$ (see Appendix A) is an estimate for $\sigma_1^2$. An analogous construction is carried out for equation (6). In the estimation process, in order to avoid local minima, we randomize the initial guess of the optimization algorithm: for every $j = 1, \ldots, r+s+1$, we select initial guesses $\theta_j^{\text{initial}}$ randomly chosen with uniform distribution in a prescribed range $(l_j, u_j)$, that represents the interval of admissible values for that parameter. After $m$ iterations of this process, we pick the best solution minimizing the objective functional (10).

## 3. Theory

### 3.1. Identifiability

From a structural-theoretical point of view, we are interested in analyzing the *identifiability* of model (1). In general, this analysis identifies whether unknown parameters can be estimated in a unique way from the available measurements on the system (Audoly et al., 2001; Saccomani & Thomaseth, 2019). There are two conceptually different ways to develop this analysis: the structural approach (a priori) and the practical approach (a posteriori). The former is a theoretical property that resides in the structure of the model itself: the parameters are *structurally identifiable* if they can be (globally) uniquely identified from the available measurements; they are called *locally structurally identifiable* if they are structurally identifiable within a neighborhood of the solution. On the other hand, the practical approach is based on the outcome of the data fitting, and it is evaluated by means of the *correlation matrix* of the parameters.

The mathematical description of the structural identifiability is placed in Appendix B. Without simplifying the model or assuming we have access to additional data, as the curve of recovered cases $R$, we cannot guarantee structural identifiability. Therefore, we proceed to analyze *practical identifiability.*

### 3.2. Practical identifiability

Structural identifiability is based on two assumptions: the model structure is accurate and measurement errors are absent. However, they are not valid in practice and, therefore, it is necessary to assess whether the parameters can be reliably and accurately estimated from noisy data (Miao et al., 2011). Let $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\mu}_r)$ be the vector of parameters estimated from the

data fitting, and $\hat{T}$, $\hat{D}$, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ the model approximations for the confirmed cases, deaths, and their respective estimated variances, as introduced in Section 2.4. The correlation matrix quantifies the interdependence between model parameters, and can be computed as follows: starting from the *Fisher Information matrix (FIM)*

$$FIM = \left| \frac{1}{\hat{\sigma}_1^2} \left( \frac{\partial \hat{T}}{\partial \theta} \right) \right|_{\theta=\hat{\theta}}^{T} \Sigma^{-1} \left( \frac{\partial \hat{T}}{\partial \theta} \right)_{\theta=\hat{\theta}} + \left| \frac{1}{\hat{\sigma}_2^2} \left( \frac{\partial \hat{D}}{\partial \theta} \right) \right|_{\theta=\hat{\theta}}^{T} \Sigma^{-1} \left( \frac{\partial \hat{D}}{\partial \theta} \right)_{\theta=\hat{\theta}},$$

we compute the *covariance matrix C* as the inverse of *FIM*. Then, the element $r_{ij}, 1 \leq i, j \leq r+s+1$ of the *correlation matrix R* is defined as

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}$$

which measures the correlation between $\hat{\theta}_i$ and $\hat{\theta}_j$. A value close to 1 indicates that the parameters are strongly interconnected, and that each of them varies according to the other.

## 4. Results

### 4.1. Data fitting

We first set the initial conditions according to Section 2.4.1 and we approximate the curve of the first two weeks of the outbreak, as represented in Fig. 4. Following Remark 2.1, the fitting with respect to the initial guess for $\theta_0$ is robust: the final values on July 31 range from 0.0127 to 0.0131. As summarized in Table 6, the best fitting over the outbreak period March–July 2020 is obtained by choosing an approximation with 4 coefficients for the two parameters $\beta$ and $\mu$, and B-spline of order 2 for $\beta$ and order 1 for $\mu$.

We fit the available data using the L-BFGS-B algorithm explained in Section 2.4.2. Figs. 5 and 6 show a very good matching between the available data and the fitting curves.

### 4.2. The estimate of the unreported rate $\alpha$

To analyze the robustness of the estimate of $\alpha$ with respect to the epidemiological parameters, we perform the following analysis: choosing a grid of values based on the confidence interval of each parameter, we estimate the parameters for each possible combination and we determine the variation in the estimated value of $\alpha$, as reported in Table 7.

In the paper, we have used $\alpha$ as a proxy of the proportion of unreported positive cases, and $\rho$ as the rate of testing among asymptomatic or paucisymptomatic individuals. However, it must be noticed that, owing to the structure of model (1), the testing among asymptomatic individuals may lead to detect positive cases that would have otherwise remained unreported. This is specifically due to the term $-\rho A$ in the dynamics of $A$ in (1). Indeed, this term entails that some individuals in $A$ are detected as positive and thus moved into $Q$. In principle, this may reduce the effective number of total unreported cases. To verify that our results are not undermined by this issue, we analyze the evolution of the ratio between the number of *non-tested* positive cases, accounted for by the curve $1-(S+T)$, and the total number of positive cases (obtained as the cumulative number of individuals that move out of $S$ over the period under consideration). As described by the plot in Fig. 7, the rate of

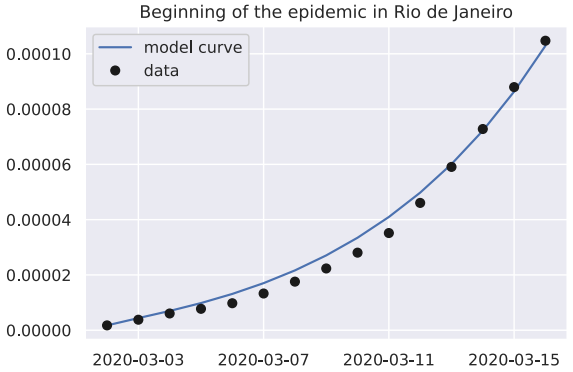


**Fig. 4.** Results of the model fitting at the beginning of the outbreak. The black dots represent the time series of reported cumulative COVID-19 cases from March 2 to March 16, 2020, and the blue line is the fitted model curve following Section 2.4.1.

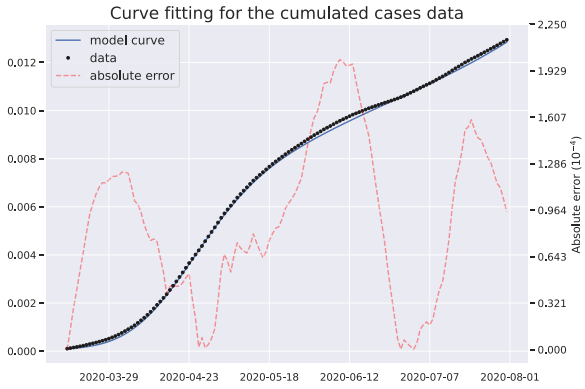


**Fig. 5.** Results of the model fitting. The black dots represent the time series of reported cumulative COVID-19 cases from March 16 to July 31, 2020, and the blue line is the fitted model curve $T$, following Section 2.4.2. The dashed red line shows the absolute error of the model in the scale $10^{-4}$.

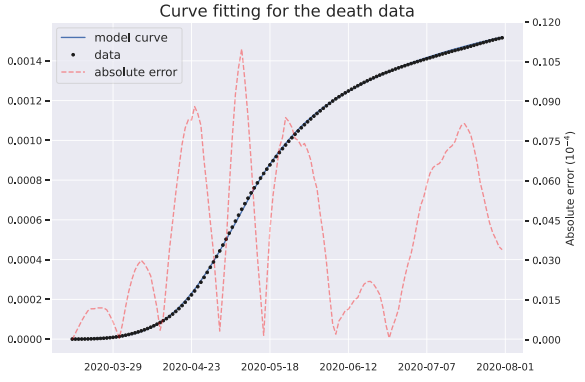


**Fig. 6.** Results of the model fitting. The black dots represent the time series of reported COVID-19-related deaths from March 16 to July 31, 2020, and the blue line is the fitted model curve $D$, following Section 2.4.2. The dashed red line shows the absolute error of the model in the scale $10^{-4}$.

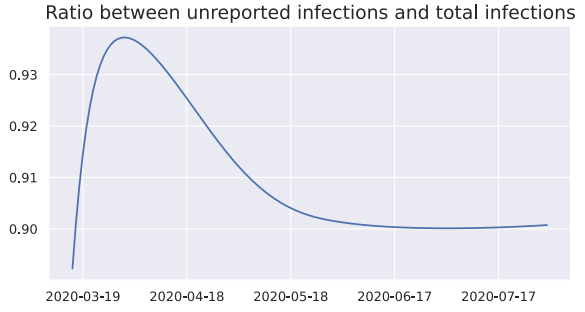


**Fig. 7.** Curve of proportion of unreported cases among infections, in the period of consideration. This curve is defined for each $t$ as $1 - \frac{T(t)}{1-S(t)}$, that is the proportion of individuals who already got the disease and were not tested until time $t$, divided by the proportion of individuals who already got the disease.

unreported cases is consistently above 90%. In particular, it is worth noticing that the ratio of unreported cases is higher during the first month of the outbreak, reaching up to 94% of unreported cases, and then stabilizes around a steady value above 90%.

### *4.3. Do model residuals approximate the errors?*

Since we have assumed the errors are normal-distributed, we expect that the residuals have a similar behaviour. We plot the histograms of the residuals and the *Q-Q plots* (comparison between the quantiles of the normal distribution and the

**Table 6**
Results of model selection with *Akaike Criterion* (AIC) in the scale $10^3$. Each tuple $(r, k)$ includes the B-spline representation hyperparameters of $\beta$ and $\mu$ independently, as explained in Section 2.3.2, such that $r$ is the number of coefficients to be estimated and $k$ is the order of the B-spline. The best model has the lowest value highlighted in **bold**.

| B-splines | | Parameter $\mu$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **(3,0)** | **(3,1)** | **(3,2)** | **(4,0)** | **(4,1)** | **(4,2)** | **(4,3)** |
| Parameter $\beta$ | **(3,0)** | −2.021 | −2.029 | −2.023 | −2.025 | −2.028 | −2.032 | −2.001 |
| | **(3,1)** | −2.228 | −2.241 | −2.266 | −2.288 | −2.298 | −2.311 | −2.298 |
| | **(3,2)** | −2.246 | −2.230 | −2.253 | −2.294 | −2.337 | −2.331 | −2.309 |
| | **(4,0)** | −1.967 | −1.979 | −1.979 | −1.998 | −2.002 | −2.000 | −1.987 |
| | **(4,1)** | −2.262 | −2.304 | −2.277 | −2.302 | −2.372 | −2.348 | −2.338 |
| | **(4,2)** | −2.346 | −2.243 | −2.262 | −2.300 | **−2.374** | −2.357 | −2.332 |
| | **(4,3)** | −2.213 | −2.233 | −2.251 | −2.294 | −2.363 | −2.347 | −2.323 |

quantiles of the residuals sampling distribution) in Fig. 8. In addition to this visual analysis, we apply statistical tests to verify correlation and normality: the *Ljung-Box test* (Ljung & Box, 1978) returns a p-value 0 which indicates that the residuals are not uncorrelated; the *Jarque-Bera test* (Jarque & Bera, 1980) does not reject the null hypothesis at the 5% level, which is an evidence supporting the normality of the residuals.

### 4.4. Time-varying transmission and mortality rate

Another important outcome of the study is that it provides time-varying curves of transmission and mortality rate, both depicted in Fig. 9. We observe a fast decay of the transmission rate $\beta$ over the month following the imposition of restrictions, which drives down the effective (time-varying) reproduction number $\mathcal{R}_t$ curve, as well (see Fig. 11). Moreover, the temporary growth of the mortality rate reflects the increase of under-reporting in the first weeks of the outbreak (see Fig. 7).

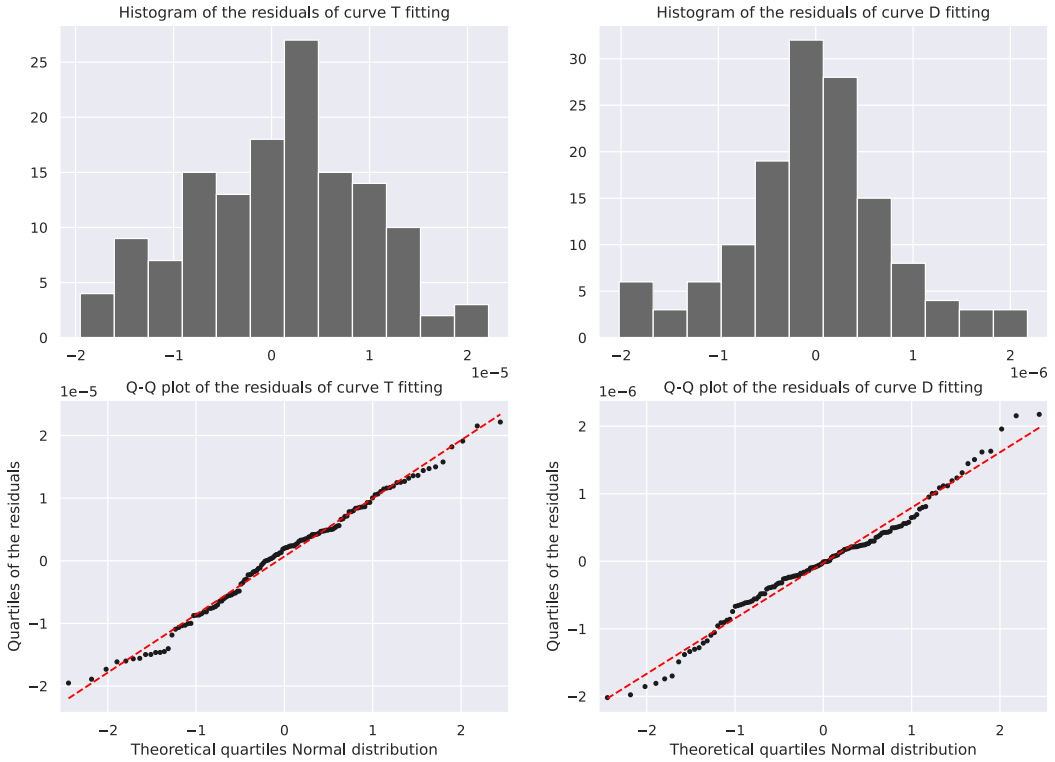


**Fig. 8.** Graphical analysis of the model's residuals, considering the difference between the cumulative and the model curves produced by the system (1) with the parameters' estimates. The first row presents the histograms for the fitting of $T$ and $D$, respectively. The second row presents the Q-Q plots, in which the red line is the line that best approximates the black dots. They represent the comparison between the theoretical quantiles from the Normal distribution and the empirical quantiles from the residuals.

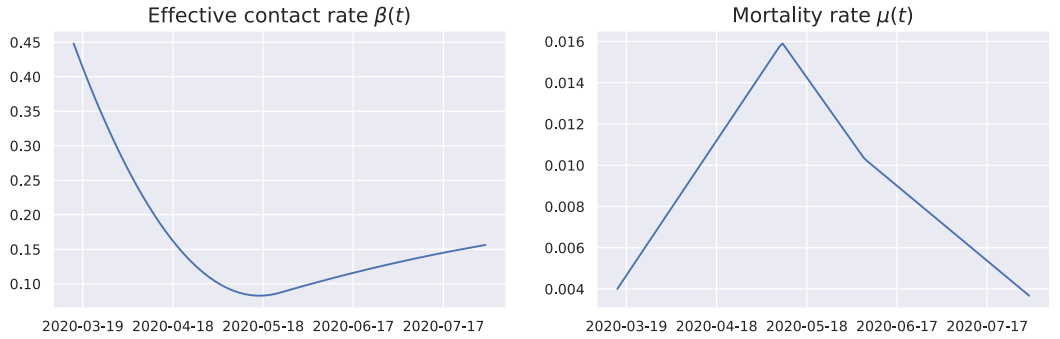


**Fig. 9.** The effective contact rate $\beta$ and the mortality rate $\mu$ as functions of the time $t$. The curves are the B-spline approximations defined by the estimated coefficients. The B-spline is of order 2 for $\beta$ and of order 1 for $\mu$, according to Subsection 4.1. It can be observed that the significant rise in the mortality rate coincides with the peak of the infections occurred at the beginning of May 2020. See Fig. 5 for the curve of infections.

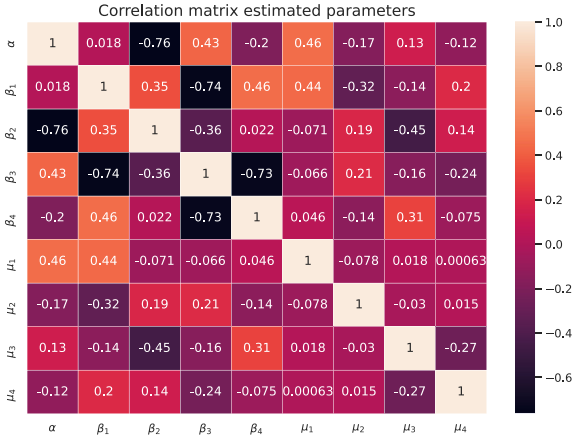


**Fig. 10.** Correlation matrix of the estimated model's parameters, following Section 3.2. The higher the absolute value is, the more the parameters are correlated, indicating a relation in the estimation process.

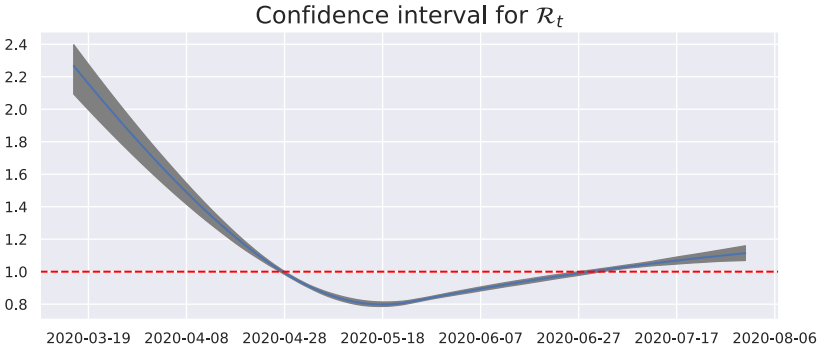


**Fig. 11.** Estimation results of the (time-dependent) reproduction number $\mathcal{R}_t$, following the Bootstrap method from Section 2.5. The blue curve is the median of the estimated curves and the gray area represents the confidence interval for each time. The dashed red line is the threshold 1 for $\mathcal{R}_t$.

### *4.5. Correlation and confidence intervals for the parameters' estimates*

The correlation matrix among the estimated parameters (Section 3.2) is depicted in Fig. 10. It is worth noticing that the parameter $\alpha$ and the second coefficient of the B-spline of $\beta$ have a strong inverse correlation.

Relying on the Bootstrap method from Section 2.5 we derive confidence intervals for the parameters of the fitting: after $N = 500$ simulations with $m = 10$, we conclude that the 95% confidence interval for the unreported rate $\alpha$ is (0.849, 0.931). Fig. 11 displays the confidence interval of the effective (time-varying) reproductive number $\mathcal{R}_t$ and the estimated curve in blue. Fig. 12 provides a scatter plot and histogram visualization for the estimated correlations in Fig. 10. Finally, Table 8 summarizes the estimates for all parameters.

## 5. Discussion

The estimated range of values for the unreported rate $\alpha$ is in line with the results of other works on this issue. In Rio de Janeiro state and up to April 20, 2020, Prado et al. (2020) estimated the notification at 7.2%, while they found the value of 9.2% at a national level. In the period February 26 – March 25, 2020, and also at a Brazilian level, Crokidakis (2020b) estimated that

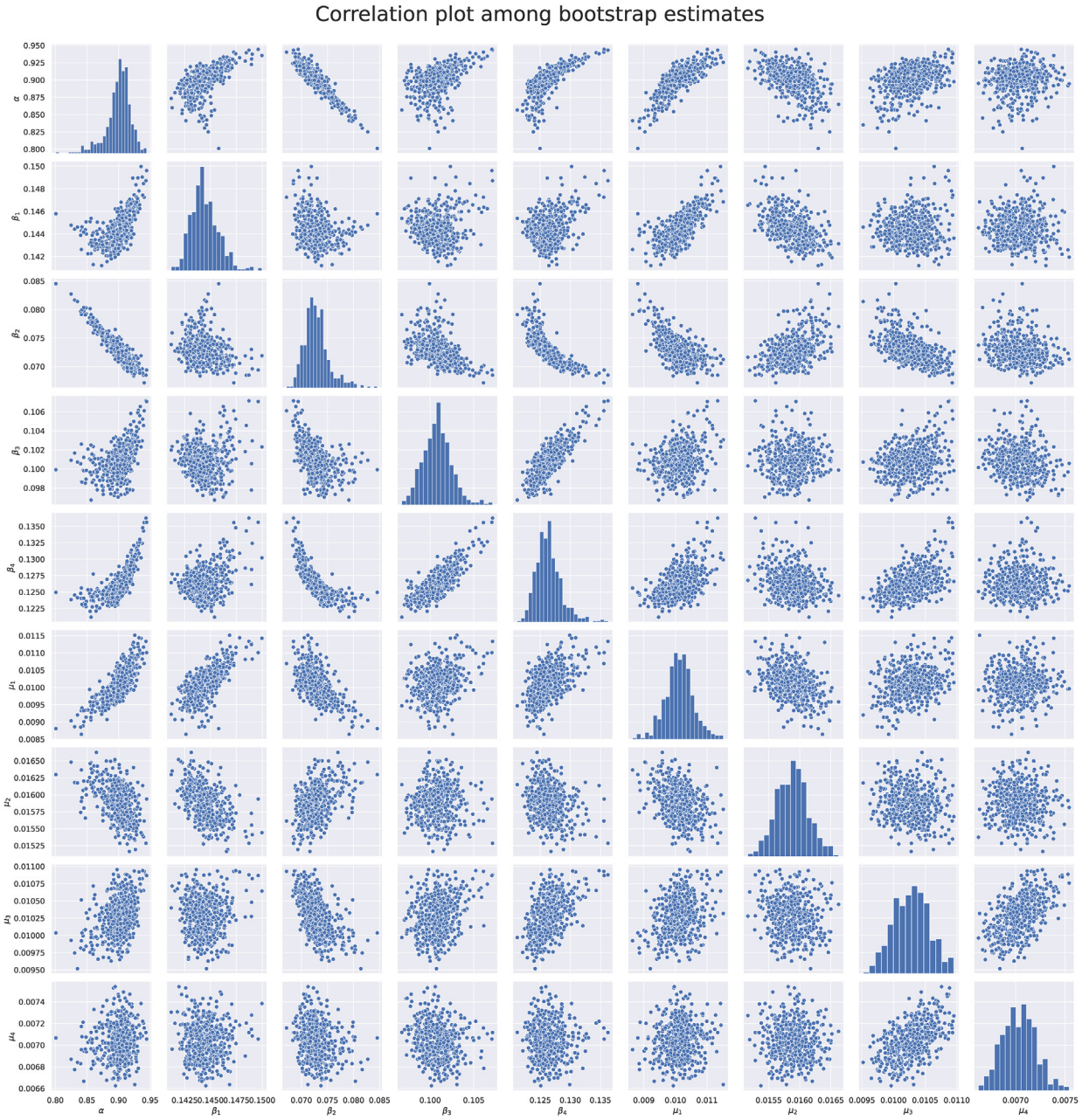


**Fig. 12.** Scatter plots of estimates considering each Bootstrap sample for each pair of parameters, and the histogram of the estimates for each parameter.

**Table 7**
Analysis of the robustness of the estimate of $\alpha$ with respect to the parameters $\tau$, $\sigma$, $\rho$, $\gamma_1$ and $\gamma_2$. For each parameter, the Interval indicates a confidence interval for its estimate, according to the references of Table 4; the Range of values of $\alpha$ is the resulting interval (convex hull) encompassing all the estimated values of $\alpha$.

| Parameters | Interval | Range of values of $\alpha$ |
|---|---|---|
| $\tau^{-1}$ | [2, 4] | [0.897, 0.902] |
| $\sigma^{-1}$ | [2, 4.5] | [0.897, 0.9] |
| $P$ | $[0, 5 \cdot 10^{-4}]$ | [0.898, 0.899] |
| $\gamma_1^{-1}$ | [6.5, 9.5] | [0.894, 0.903] |
| $\gamma_2^{-1}$ | [11, 16] | [0.896, 0.898] |

**Table 8**
Estimates of the median and the 95%-confidence interval for each parameter, following the Bootstrap approximation defined in Section 2.5.

| Parameter | Median | Confidence interval |
|---|---|---|
| $\alpha$ | 0.903 | [0.849, 0.93] |
| $\beta_1$ | 0.144 | [0.142, 0.147] |
| $\beta_2$ | 0.073 | [0.069, 0.079] |
| $\beta_3$ | 0.101 | [0.098, 0.104] |
| $\beta_4$ | 0.126 | [0.123, 0.132] |
| $\mu_1$ | 0.0101 | [0.0092, 0.011] |
| $\mu_2$ | 0.0159 | [0.0154, 0.0164] |
| $\mu_3$ | 0.0103 | [0.0098, 0.0108] |
| $\mu_4$ | $7.03 \cdot 10^{-3}$ | $[6.73 \cdot 10^{-3}, 7.38 \cdot 10^{-3}]$ |

*"for each patient in quarantine, approximately ten infectious individuals are present in the population"*, which is in line with our findings for the city of Rio de Janeiro. Bastos and Cajueiro (2020) analyzed the outbreak at a national level over the same period as in Crokidakis (2020b), finding that approximately 70% of the cases were not reported. The estimate of Ritto et al. (2021) of the asymptomatic rate for the period from January 24 to July 28, 2021 was 0.83, which also agrees with our results.

It is also interesting to compare the estimates of $\mathcal{R}_t$ over time: in (Mellan et al., 2020), the point estimate for May 9, 2020 from the state of Rio de Janeiro was 1.1 with 95%-confidence interval [0.9, 1.3], which includes our range of values. The curve estimated by (Observatório COVID-19 BR, 2021) for the city of Rio de Janeiro has a very similar trend as in Fig. 11: at the beginning of May, the $\mathcal{R}_t$ has an estimated value smaller than 1 and it grows again until it is greater than 1 at the end of July. These values corroborate our fitting.

The correlation between $\alpha$ and $\beta_2$ represents a limitation in terms of identifiability of the system. In order to achieve a structural identifiability result it is necessary to know the recovered curve (see Appendix B), whereas for the practical identifiability it would be useful to approximate the transmission and mortality functions with a different model.

## 6. Conclusions

In this work, we combine tools from the analysis of differential equations, statistics, and optimization to estimate the unreported rate of COVID-19 in the city of Rio de Janeiro during the first outbreak in March–July 2020. We determine that the rate of unreported positive cases is about 90%, with a confidence interval between 85% and 93%. This means that every case reported by the health system corresponds to about 9–10 cases that were not detected. This estimation shall be considered as a statistical approximation of the unreported rate: in addition to the lack of testing kits during the period under evaluation, other delays and errors at all stages of the notification process may generate bad inferences. Data analysis techniques have been deployed to deal with the low quality of data.

Concerning the modeling chosen for the fitting, it is of great importance to underline that the initialization of the parameters has little influence on the final estimation of $\alpha$, supporting the robustness of our analysis. However, the fact that the residuals of the estimation are normally distributed but correlated needs more attention and will be addressed in future works.

Finally, the outcome of this work provides a description of the evolution in time of the disease reproductive number. This estimation provides a useful tool to determine periods of growth or decrease of the epidemic force at a geo-localized level, and thus to inform policymakers in their decision process to limit the spread of the virus and its consequences on the population.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors thank the referees for their comments, that helped improving the first version of the paper.

The authors wish to thank Luiz Max Carvalho (FGV EMAp) and Marcelo Fernandes (FGV EESP) for fruitful discussions about several statistical methods applied in this research.

The first and third authors were supported by FAPERJ and CNPq, Brazil. The second author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2021-02632. The third author thanks the Center for the Development of Mathematics and Science (FGV CDMC) for their support.

## Appendix A. Variance estimate

For the variances $\sigma_1^2$ and $\sigma_2^2$, we follow the estimates from (Seber & Wild, 2005, p. 21–28) given by

$$\hat{\sigma}_1^2 = \frac{1}{n-K}(y^{(1)} - \hat{T}(\hat{\theta}))^T \Sigma^{-1} (y^{(1)} - \hat{T}(\hat{\theta})),$$

where $K = r + s + 1$ and $n$ is the number of data points. In a similar way,

$$\hat{\sigma}_2^2 = \frac{1}{n-K}(y^{(2)} - \hat{D}(\hat{\theta}))^T \Sigma^{-1} (y^{(2)} - \hat{D}(\hat{\theta})).$$

## Appendix B. Structural identifiability

Consider a dynamical system

$$\begin{aligned} \dot{x} &= f(x(t), \theta), \quad x(0) = x_0 \\ y(t) &= h(x(t), \theta), \end{aligned} \tag{B.1}$$

where $x(t) \in \mathbb{R}^n, y(t) \in \mathbb{R}^m$ and $f$ and $h$ are rational functions of the variable $x$ and $\theta \in \Theta \subset \mathbb{R}^p$, $\Theta$ being the set of admissible values of the parameter $\theta$. The variable $y$ is the output of the system, that is the observable component. In the model, $y$ is the number of confirmed cases and deaths. When $x(0) = x_0$, the measurement with initial value $x_0$ is $y = \psi_{x_0}(\theta)$. The structural identifiability (Ljung & Glad, 1994) of system (B.1) at a fixed $\theta^* \in \Theta$ is related to the number of solutions of the equation

$$\psi_{x_0}(\theta) = \psi_{x_0}(\theta^*). \tag{B.2}$$

The system (B.1) is *globally identifiable* at $\theta^*$ if equation (B.2) has a unique solution $\theta = \theta^*$, or, equivalently if the mapping $\psi_{x_0}$ is invertible. System (B.1) is *locally identifiable* if $\psi_{x_0}$ is invertible in a neighborhood of $\theta^*$.

We used the DAISY software (Bellu et al., 2007) as a computational method to check the structural identifiability of the model. DAISY – Differential Algebra for Identifiability of SYstems – is a software based on the programming language *Reduce*, specific to the identifiability problem in dynamical systems. DAISY requires certain specific conditions on the system, that allow to apply algebraic methods to the model under consideration. In order to rewrite our system (1) in terms of the DAISY algorithm, we treat all model parameters' in Section 2.1 as constant in time. If we assume to know the curve $R$ of recovered cases, we verify that model (1) is globally identifiable. However, the knowledge of $R$ may sometimes be only partially available, as discussed in Section 2.2. Without access to these data, DAISY has not been able to certify the structural identifiability of the system.

## References

Aronna, M. S., Guglielmi, R., & Moschen, L. M. (2021). A model for COVID-19 with isolation, quarantine and testing as control measures. *Epidemics, 34*.

Audoly, S., Bellu, G., D'Angio, L., Saccomani, M., & Cobelli, C. (2001). Global identifiability of nonlinear models of biological systems. *IEEE Transactions on Biomedical Engineering, 48*(1), 55–65.

Ávila, E. (2020). *RJ decide ampliar restrições no transporte público e isolar a cidade do Rio (Rio de Janeiro decides to increase restrictions in the public transportation to isolate the city)*. globo Rio de Janeiro. Available at https://g1.globo.com/rj/rio-de-janeiro/blog/edimilson-avila/noticia/2020/03/19/rj-decide-ampliar-restricoes-no-transporte-publico-e-isolar-a-cidade-do-rio.ghtml.

Bastos, S. B., & Cajueiro, D. O. (2020). Modeling and forecasting the early evolution of the COVID-19 pandemic in Brazil. *Scientific Reports, 10*(1), 1–10.

Bellu, G., Saccomani, M. P., Audoly, S., & D'Angiò, L. (2007). DAISY: A new software tool to test global identifiability of biological and physiological systems. *Computer Methods and Programs in Biomedicine, 88*(1), 52–61.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing, 16*(5), 1190–1208.

Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., et al. (2020). Inferred duration of infectious period of SARS-CoV-2: Rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open, 10*(8).

Canzian, F. (2020). *Estados e municípios no país relatam subnotificação gigantesca de casos (States and cities alert on a huge underreporting of cases)*. folha de São Paulo Newspaper. Available at https://www1.folha.uol.com.br/equilibrioesaude/2020/04/estados-e-municipios-no-pais-relatam-subnotificacao-gigantesca-de-casos.shtml.

Cao, J., Huang, J. Z., & Wu, H. (2012). Penalized nonlinear least squares estimation of time-varying parameters in ordinary differential equations. *Journal of Computational & Graphical Statistics, 21*(1), 42–56.

Crokidakis, N. (2020a). COVID-19 spreading in Rio de Janeiro, Brazil: Do the policies of social isolation really work? Chaos. *Solitons & Fractals, 136*, Article 109930.
Crokidakis, N. (2020b). Modeling the early evolution of the COVID-19 in Brazil: Results from a susceptible–infectious–quarantined–recovered (SIQR) model. *International Journal of Modern Physics C, 31*(10), Article 2050135.
Dantas, G., Siciliano, B., França, B. B., da Silva, C. M., & Arbilla, G. (2020). The impact of COVID-19 partial lockdown on the air quality of the city of Rio de Janeiro, Brazil. *Science of the Total Environment, 729*, Article 139085.
De Boor, C. (1978). *A practical guide to splines* (Vol. 27). New York: Springer Verlag.
Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.
Estadão. (2020). Governo do Rio cria classificação em 3 bandeiras para flexibilizar isolamento (Government of Rio creates classification in 3 flags to make isolation more flexible). Available at https://revistapegn.globo.com/Noticias/noticia/2020/05/pegn-governo-do-rio-cria-classificacao-em-3-bandeiras-para-flexibilizar-isolamento.html.
He, D., Artzy-Randrup, Y., Musa, S. S., Gräf, T., Naveca, F., & Stone, L. (2021). The unexpected dynamics of COVID-19 in Manaus, Brazil: Was herd immunity achieved? *medRxiv*.
IBGE. (2020). Brazilian Institute for Geography and Statistics. Brazilian National Household Survey sample - PNAD COVID-19. Available at https://www.ibge.gov.br/estatisticas/sociais/trabalho/27946-divulgacao-semanal-pnadcovid1.html?=&t=downloads.
IBGE. (2021). Brazilian Institute for Geography and statistics. Cities and states. Available at https://www.ibge.gov.br/cidades-e-estados/rj/rio-de-janeiro.html.
Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters, 6*(3), 255–259.
Joshi, M., Seidel-Morgenstern, A., & Kremling, A. (2006). Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering, 8*(5), 447–455.
Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., & Lessler, J. (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine, 173*(4), 262–267.
Liang, H., Miao, H., & Wu, H. (2010). Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model. *Annals of Applied Statistics, 4*(1), 460.
Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science, 368*(6490), 489–493.
Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303.
Ljung, L., & Glad, T. (1994). On global identifiability for arbitrary model parametrizations. *Automatica, 30*(2), 265–276.
Mellan, T. A., Hoeltgebaum, H. H., Mishra, S., Whittaker, C., Schnekenberg, R. P., Gandy, A., Unwin, H. J. T., Vollmer, M. A., Coupland, H., Hawryluk, I., et al. (2020). *Subnational analysis of the COVID-19 epidemic in Brazil*. MedRxiv.
Miao, H., Xia, X., Perelson, A. S., & Wu, H. (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Review, 53*(1), 3–39.
Morato, M. M., Bastos, S. B., Cajueiro, D. O., & Normey-Rico, J. E. (2020). An optimal predictive control strategy for COVID-19 (SARS-CoV-2) social distancing policies in Brazil. *Annual Reviews in Control, 50*, 417–431.
Moreira, W., & Pelegi, A. (2020). Coronavírus: Decreto isola estado do Rio de Janeiro e transporte é afetado (coronavirus: Decree isolates the state of Rio de Janeiro and public transport is affected). Available at https://diariodotransporte.com.br/2020/03/19/coronavirus-decreto-isola-estado-do-rio-de-janeiro-e-transporte-e-afetado/, diário do Transporte.
Moschen, L. M. (2021). Repository COVID-19. GitHub. Available at https://github.com/lucasmoschen/covid-19-model.
Municipal Health Department, City Hall of Rio de Janeiro. (2021). *Dados individuais dos casos confirmados de COVID-19 no município do Rio de Janeiro (Individual data of COVID-19 confirmed cases in the city of Rio de Janeiro)*. Available at https://www.arcgis.com/home/item.html?id=f314453b3a55434ea8c8e8caaa2d8db5.
Nogrady, B. (2020). What the data say about asymptomatic COVID infections. *Nature*.
Observatório COVID-19 BR. (2021). R efetivo no Rio de Janeiro (effective R in Rio de Janeiro. Available at https://covid19br.github.io/municipios.html?aba=aba3&uf=RJ&mun=Rio_de_Janeiro.
Official Journal of the State of Rio de Janeiro. (2020a). Ordinance number 46.973, March 16th 2020. Available at https://pge.rj.gov.br/comum/code/MostrarArquivo.php?C=MTAyMjI.
Official Journal of the State of Rio de Janeiro. (2020b). Law number 8859, June 3rd, 2020. Available at http://www.aeerj.net.br/file/04-06-2020-leiestadomascara.pdf.
Portal COVID-19 Brasil. (2021). COVID-19 BRASIL. Available at https://ciis.fmrp.usp.br/covid19/.
Prado, M. F.d., Antunes, B. B.d. P., Bastos, L.d. S. L., Peres, I. T., Silva, A.d. A. B.d., Dantas, L. F., Baião, F. A., Maçaira, P., Hamacher, S., & Bozza, F. A. (2020). *Análise da subnotificação de COVID-19 no Brasil (Analysis of COVID-19 underreporting in Brazil)*. Revista Brasileira de Terapia Intensiva (ahead).
Rai, B., Shukla, A., & Dwivedi, L. K. (2021). incubation period for COVID-19: A systematic review and meta-analysis. *Journal of Public Health*, 1–8.
Ramsay, J. O., Hooker, G., Campbell, D., & Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society: Series B, 69*(5), 741–796.
Ritto, T. G., Cunha, A., Jr., & Barton, D. A. (2021). Parameter calibration and uncertainty quantification in an SEIR-type COVID-19 model using approximate Bayesian computation. In *3rd Pan American congress on computational mechanics (PANACM 2021)*. hal–03425932.
Sabino, E. C., Buss, L. F., Carvalho, M. P., Prete, C. A., Crispim, M. A., Fraiji, N. A., Pereira, R. H., Parag, K. V., da Silva Peixoto, P., Kraemer, M. U., et al. (2021). Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *The Lancet, 397*(10273), 452–455.
Saccomani, M. P., & Thomaseth, K. (2019). Calculating all multiple parameter solutions of ODE models to avoid biological misinterpretations. *Mathematical Biosciences and Engineering, 16*(6), 6438–6453.
Seber, G., & Wild, C. (2005). *Nonlinear regression. Wiley series in probability and statistics*. Wiley. URL https://books.google.com.br/books?id=YBYlCpBNo_cC.
The 2019 nCoV Outbreak Joint Field Epidemiology Investigation Team and Q. Li. (2020). *An out-break of NCIP (2019-nCoV) infection in China - Wuhan, Hubei Province, 2019 - 2020*. Available at http://weekly.chinacdc.cn/en/article/id/e3c63ca9-dedb-4fb6-9c1c-d057adb77b57.
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272.
World Health Organization. (2020). Coronavirus disease (COVID-19). Available at https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19.