Nonlinear Dyn (2022) 107:1919–1936
https://doi.org/10.1007/s11071-021-07069-9

ORIGINAL PAPER

Framework for enhancing the estimation of model parameters for data with a high level of uncertainty

Gustavo B. Libotte · Lucas dos Anjos · Regina C. C. Almeida · Sandra M. C. Malta · Renato S. Silva

Received: 8 May 2021 / Accepted: 15 November 2021 / Published online: 7 January 2022

Abstract Reliable data are essential to obtain adequate simulations for forecasting the dynamics of epidemics. In this context, several political, economic, and social factors may cause inconsistencies in the reported data, which reflect the capacity for realistic simulations and predictions. In the case of COVID-19, for example, such uncertainties are mainly motivated by large-scale underreporting of cases due to reduced testing capacity in some locations. In order to mitigate the effects of noise in the data used to estimate parameters of models, we propose strategies capable of improving the ability to predict the spread of the diseases. Using a compartmental model in a COVID-19 study case, we show that the regularization of data by means of Gaussian process regression can reduce the variability of successive forecasts, improving predictive ability. We also present the advantages of adopting parameters of compartmental models that vary over time, in detriment to the usual approach with constant values.

Keywords Gaussian process regression · Regularization of data · Uncertainty quantification · Noisy data · Time-dependent parameters

G. B. Libotte (✉) · L. dos Anjos · R. C. Almeida · S. M. C. Malta · R. S. Silva
National Laboratory for Scientific Computing, Getúlio Vargas Av., 333, Quitandinha, Petrópolis, Rio de Janeiro, Brazil
e-mail: glibotte@lncc.br

1 Introduction

Since the onset of the novel coronavirus (SARS-CoV-2) pandemic, at the end of 2019, a wealth of research has been carried out from across the globe, aiming to understand the dynamics and transmission patterns of the disease. More than a year after the notification of the first case, the number of infected individuals keeps rising significantly worldwide. In the meantime, the confirmation of reinfections and identification of seasonal immunity [17] reinforces the need for actions to contain the disease, even in locations where the epidemic would be under control.

Governmental decisions to mitigate the spread of the disease, such as the introduction of lockdown and social distancing measures, are usually based on computational simulations whose preeminent objective is to predict the way the disease spreads in the population, considering continuously reported data [15]. However, there are several factors associated with natural, economic, and social aspects that make it difficult to adequately predict the spread of the disease and, consequently, the definition of a comprehensive policy for prevention and control of the disease [21,36,46]. The heterogeneity of the population concerning attributes such as demographic diversity, age-dependent characteristics, and randomness related to the mobility and interaction of individuals makes it hardly possible to create a model capable of incorporating all these features together (see, for instance, Refs. [7,10,12,16]).

Asymptomatic people also play a significant role in the ongoing pandemic. Oran and Topol [35] presented a comprehensive bibliographic review on the estimation of asymptomatic cases of COVID-19 in different parts of the world and concluded that the proportion of asymptomatic individuals may vary from 40 to 45% in relation to the total number of reported cases. The scenario of widespread underreporting coupled with a deficient screening and testing capacity leads to significant uncertainty in relation to the reported data of infected individuals. The impact of such uncertainties was analyzed by Ioannidis [22], Li et al. [26], and Wu et al. [58].

Turning attention to the impact of the pandemic in Brazil, a country with a continental dimension, such problems tend to become even more evident [28]. Socioeconomic inequalities [43,55] and cultural factors [14,38] have a direct impact on access to information and health services, which translates into high rates of infection and, as a consequence, underreporting cases. Veiga e Silva et al. [56] also analyzed presumptive inconsistencies in the data collected and made available by the Ministry of Health in Brazil. They reported that there may be a difference of approximately 41% in the number of deaths caused by COVID-19 related complications.

In an attempt to describe the spread of COVID-19 on different population groups, several models have been proposed by means of integrating typical features related to the disease, such as the quarantine period, lockdown, social distancing, and hospitalization. Massonis et al. [29] bring together several of these models, classifying them hierarchically in relation to the number of coupled features. In general, even the most complex models, that is, those with supposedly more capacity to associate knowledge about the spreading dynamics of the disease, tend to experience some difficulties in identifying the behavior of noisy data in the long term, as shown by Alberti and Faranda [2] and Roda et al. [46].

The major motivation of this work is to provide alternatives to enhance the capacity of estimating parameters in compartmental models and predicting the spread of COVID-19, taking into account data with a high level of uncertainty, such as the number of new cases reported in the state of Rio de Janeiro since March 05, 2020. The data do not have well-defined behavior, in such a way that the variations in subsequent days are caused, in part, by the factors that deepen the disparities in the notifications of cases of infection. In addition, the Brazilian government estimates the COVID-19 data considering the daily count in the municipalities—Brazil is made up of 5570 municipalities, of which 92 make up the state of Rio de Janeiro—which are autonomous in relation to population testing policy and do not follow a common strategy to prevent the disease. As in some locations data are not reported on weekends, there are sudden drops in the number of new infections, which afterward cause unexpected increases when data are reported at the beginning of the following week.

In this context, aiming at expanding the predictive capacity of compartmental models subject to scenarios of great uncertainty, the objectives of this work are to propose the use of strategies capable of reducing the variability of the estimation of model parameters and to discuss the advantages of considering time-dependent parameters. We propose that the noisy data set be regularized by means of Gaussian process regression (GPR). This approach allows a set of data to be smoothed, so as to decrease its noise level without significantly changing its behavior. To confirm this assumption, we compared a subgroup of reported data on dead and infected individuals in the state of Rio de Janeiro to simulations produced by the SEIRPD-Q model (which is defined in the due section), whose parameters are estimated using regularized data. The results obtained are also compared to the simulations calculated in the usual way, without regularizing the data, in order to quantify the predictive capacity in relation to known data and the gain in relation to the parameter estimation approach with unchanged data.

A recent review of the literature on the subject found that, to date, few works that incorporate the concepts of Gaussian process (GP) applied to the epidemiological modeling of COVID-19 have been published. These works are briefly presented below. Ketu and Mishra [24] proposed the Multi-Task GPR model, aiming to predict the COVID-19 outbreak worldwide. Zhou and Ji [60] proposed a model for transmission dynamics of COVID-19 considering underreporting of cases (what they called undocumented infections) and estimated the time-varying disease transmission rate using GPR and a Bayesian approach. Arias Velásquez and Mejía Lara [3] demonstrated the correlation between industrial air pollution and infections by COVID-19 before and after the quarantine in Peru, by presenting a classification model using Reduced-Space GPR. This methodology is used by the same authors in Ref. [3] to

report a long-term forecast for COVID-19 in the USA. In turn, Ribeiro et al. [44] compared the predictive capacity of various machine learning regression and statistical models, considering short-term forecasting of COVID-19 cumulative cases in Brazil. As far as we are aware, this is the first time that GPR is employed for the regularization of COVID-19 data, which are subsequently analyzed using a compartmental model.

Other techniques can be used to regularize the data as proposed in this work. One of the approaches that come closest to GPs is the Kalman filter (see Ref. [20] for details). Kalman filters can be seen as a special case of GPs, although the former is not thought of as a nonparametric model. It is widely used for time-series forecasting and can be more efficient than GPs when the problem is described by a Markovian process or linear observation models [42], which is not our case. Other methods used in smoothing noisy data are spline models, support vector machines, and auto-regressive moving average models, for instance (for more details, see Refs. [8,40]). In the context of the spreading dynamics of COVID-19, these techniques were also employed to help predict the number of cases and dead individuals, as seen in Refs. [1,4,50,59]. However, it is important to emphasize that our objective is not to compare numerical regularization techniques but employ GPs as a data regularization technique, in order to improve the predictive capacity of models, through data noise reduction. GPs provide all the necessary framework for the analyses proposed in this work, which is reasonable to support our choice.

The objective of the present paper is to answer the following issues: (*i*) how the regularization of data using GPR can affect the parameter estimation problem? (*ii*) what is the influence of time-varying parameters in terms of improving the descriptive capacity of models? In our analysis, we partition the data sets between training and test data and carry out successive parameter estimations varying the proportion between these types of data, in order to show the behavior of the obtained parameter set. We show that some of these parameters can be approximated by functions and, considering this possibility, we analyze the influence of adopting variable parameters over time. We use both a deterministic approach, in terms of least squares, to estimate the gain related to the use of regularized data and time-varying parameters, and a Bayesian approach, in order to analyze parameter uncertainties, which are propagated to the model to quantify uncertainties on the model outcomes.

2 Materials and methods

2.1 Compartmental model description

Following the models proposed by Jia et al. [23] and Volpatto et al. [57], we develop an extension of the susceptible–exposed–infected–removed (SEIR) model, termed SEIRPD-Q model, which further considers Positively Diagnosed (P) and Dead (D) groups of individuals. To account for social distancing measures, we implicitly consider a mechanism that isolates individuals from the virus transmission ($-Q$) rather than defining a specific compartment for individuals in quarantine. A schematic description of the SEIRPD-Q model is presented in Fig. 1. We consider a population susceptible to a viral outbreak, whose rate of transmission per contact is given by β. At the beginning of a COVID-19 infection, infected individuals pass through a latency period in which they are not capable of transmitting the disease to another individual, becoming infectious only after this stage, even without showing any sign of disease. Individuals in such conditions are said to be exposed, with incubation period given by $1/\tilde{\sigma}$. The infectious individuals are then divided into infected and positively diagnosed compartments, based on the premise that a large number of individuals who contract the virus are not diagnosed. This is due to the reduced testing capacity in some locations so that the diagnosis is made, primarily, in individuals who have severe symptoms or are hospitalized. The proportion

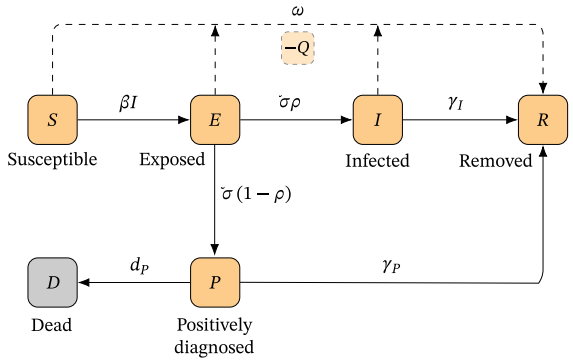

Fig. 1 Schematic description of the SEIRPD-Q model

of infected individuals, given by ρ, is related to the majority of people that only suffer mild symptoms and get recovered without significant complications. On the other hand, the complement of this group is those individuals who, in fact, have been positively diagnosed. In turn, individuals who recover from the disease after being in the infected compartment are moved to the removed compartment at a rate of γ_I. The same goes for positively diagnosed individuals, who are removed at a rate of γ_P. In addition, it is reasonable to assume that most of the individuals who died from complications caused by the disease had severe symptoms and were tested or hospitalized. Therefore, we do not consider that individuals in the infected compartment die from virus-related causes without being diagnosed, and the mortality rate of positively diagnosed individuals is given by d_P.

Quarantine measures are also incorporated in this model, affecting the susceptible, exposed, and infected compartments. Individuals in these compartments are kept in quarantine at a rate of ω, and are not assumed to be infectious considering restrictive quarantine measures. The quarantine compartment is implicitly modeled and, therefore, the removed compartment includes individuals who have undergone quarantine, along with those who have been infected and have recovered from the disease. The formulation of the SEIRPD-Q model is given by the following system of ordinary differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta S I - \omega S \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \beta S I - \tilde{\sigma} E - \omega E \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \tilde{\sigma}\rho E - \gamma_I I - \omega I \\
\frac{\mathrm{d}P}{\mathrm{d}t} &= \tilde{\sigma}\,(1-\rho)\,E - d_P P - \gamma_P P \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma_I I + \gamma_P P + \omega\,(S+E+I) \\
\frac{\mathrm{d}D}{\mathrm{d}t} &= d_P P.
\end{aligned} \tag{1}$$

The SEIRPD-Q model presents some fundamental differences in relation to those on which it was based: first, Jia et al. [23] consider that only susceptible individuals are subject to quarantine measures, which is modeled using an explicit compartment and also considering an additional parameter that controls the social distancing relaxation; second, we disregard the asymptomatic compartment, according to Volpatto et al. [57], due to the lack of this information associated with limited testing of the population; third, we consider only the mortality rate of positively diagnosed individuals, unlike Volpatto et al. [57].

2.2 Data

The data used in this work are the daily number of infected (positively diagnosed) and dead individuals in the state of Rio de Janeiro. The Brazilian Ministry of Health reports the data daily, which are synthesized and made available as shown in Ref. [13]. The analyzed data refer to the period between March 10, 2020, and October 5, 2020, consisting of 210 records. Although the number of recovered individuals is also available, it may be wise not to use these data to estimate the parameters of the model since, due to the unseemly policy of testing the population, there is much uncertainty about these data.

In the following, the observable quantities are denoted by the time series $\mathcal{D}_i^{(j)} = \mathcal{D}^{(j)}(t_i)$, for $i = 1, \ldots, p$ and $j \in \{P, D\}$. Overall, let $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_p\}$ be the finite set of real-valued measurements collected from successive observations of the daily number of positively diagnosed and dead individuals at different times t_i.

2.3 Deterministic approach for parameter estimation

Here, we want to solve the so-called *inverse problem*, i.e., to determine the model parameters that generate outputs as close as possible to the observable data. To this end, let $\boldsymbol{\theta}$ be the m-dimensional vector of model parameters, and $y_i^{(j)} = y^{(j)}(t_i, \boldsymbol{\theta})$ be the model responses at different times t_i, $i = 1, \ldots, p$ and $j \in \{P, D\}$, as used previously. Additionally, let $Y_i^{(j)}$ denote the cumulative sum of all model responses given by $y_i^{(j)}$.

The objective of the inverse problem is, therefore, to find the vector $\hat{\boldsymbol{\theta}}$ (an estimate of $\boldsymbol{\theta}$) that produces outputs $\hat{y}_i^{(j)}$ capable of fitting the available observations. The best fit between the responses of the model $\hat{\mathbf{y}} \in \mathbb{R}^p$ and the observed data can be estimated in terms of the *residuals*, the difference between observed and predicted measurements, given by $\mathbf{r}(\boldsymbol{\theta}) = \hat{\mathbf{y}} - \boldsymbol{\mathcal{D}}$. The solution

to an inverse problem is, in other words, the data fitting whose objective is to calculate an estimate $\hat{\boldsymbol{\theta}}$ that minimizes some error norm $\|\mathbf{r}\|$ of the residuals [45,52]. The *least squares* fitting calculates the vector of optimal parameters by taking the mean squared error, given by

$$\mathcal{E}(\boldsymbol{\theta}) = \frac{1}{p}\mathbf{r}(\boldsymbol{\theta})^{\top}\mathbf{r}(\boldsymbol{\theta}) = \frac{1}{p}\sum_{i=1}^{p} r_i(\boldsymbol{\theta})^2, \tag{2}$$

and the estimate $\hat{\boldsymbol{\theta}}$ is the vector that minimizes this quantity:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}).$$

Equation (2) is usually called the *objective function* (or *cost function*). When $\mathcal{E} \to 0$, the estimate $\hat{\boldsymbol{\theta}}$ generates an output vector $\hat{\mathbf{y}}$ that has a high level of agreement with the observed data $\boldsymbol{\mathcal{D}}$, that is, the residuals are minimized. In general, real problems do not admit $\mathcal{E} = 0$, since the noise that affects the model cannot be predicted with such accuracy.

2.4 Bayesian approach for parameter estimation

Bayesian inference provides another perspective for estimating the value of a set of parameters that best characterizes the output of a model, given a set of data. Bayesian inference differs from the deterministic approach because in addition to calibrate the parameter values, it measures their uncertainties, which is one of the focuses of this work. To conduct Bayesian inference, we need some familiarity with a few basic concepts of probability. Here, we give a brief overview of such concepts. A more detailed description is provided by Refs. [5,27,54].

Bayes theorem provides a formulation to estimate the posterior probability of the model parameters given a set of observations $\boldsymbol{\mathcal{D}}$, based on the likelihood of the event of interest occurring given the prior knowledge on the parameters. The theorem is stated as

$$p_{\text{post}}(\boldsymbol{\theta} \mid \boldsymbol{\mathcal{D}}) = \frac{p_{\text{like}}(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\theta})\, p_{\text{prior}}(\boldsymbol{\theta})}{p_{\text{evid}}(\boldsymbol{\mathcal{D}})}, \tag{3}$$

where $p_{\text{like}}(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\theta})$ is the *likelihood* function, $p_{\text{prior}}(\boldsymbol{\theta})$ is the *prior* information or beliefs on $\boldsymbol{\theta}$, $p_{\text{evid}}(\boldsymbol{\mathcal{D}})$ is the *evidence* related to the observations $\boldsymbol{\mathcal{D}}$, and $p_{\text{post}}(\boldsymbol{\theta} \mid \boldsymbol{\mathcal{D}})$ is the *posterior* distribution associated with $\boldsymbol{\theta}$.

Prior knowledge can be thought of as the probability density function over the feasible values of the model parameters, the current knowledge on their values. In turn, the likelihood assumes the role of estimating the probability of characterizing the available data, given a set of parameters. In other words, the likelihood function measures how well the data are being explained by the model. In this work, we assume a Gaussian likelihood, which has the form

$$p_{\text{like}}(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\theta}) = \prod_j \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\sum_{i=1}^{p} \frac{\left(\mathcal{D}_i^{(j)} - y_i^{(j)}\right)^2}{2\sigma_j^2}\right),$$

where σ_j^2 is a measure of the uncertainty (encompassing data errors) related to each quantity j, for $j \in \{P,\ D\}$. Here, both σ_P^2 and σ_D^2 are considered hyperparameters to be estimated together with the set of parameters $\boldsymbol{\theta}$. The evidence, also referred to as marginal likelihood, is the integral of the likelihood over the prior and is considered as a normalization constant. Thus, we actually evaluate $p_{\text{post}}(\boldsymbol{\theta} \mid \boldsymbol{\mathcal{D}}) \propto p_{\text{like}}(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\theta})\, p_{\text{prior}}(\boldsymbol{\theta})$ to produce the posterior distribution $p_{\text{post}}(\boldsymbol{\theta} \mid \boldsymbol{\mathcal{D}})$ over the parameters, that is, an updated belief about $\boldsymbol{\theta}$, given $\boldsymbol{\mathcal{D}}$.

We conducted the Bayesian inference using the probabilistic programming library PyMC3 [47]. We adopt the Transitional Markov Chain Monte Carlo method proposed by Ching and Chen [11] for parameter estimation (which is not described here, for the sake of brevity).

2.5 Gaussian process regression

Here, we describe the general GP framework, with special attention to regression problems using noisy observations. By definition, a GP is a collection of random variables, any finite number of which have a joint Gaussian distribution [40]. In other words, it is an extension of the multivariate Gaussian distributions to infinite dimensionality. GPRs take place directly in the space of functions, defining priors over functions that, once we have seen some data, can be converted into posteriors over functions [32]. Thus, a GPR model is a Bayesian nonlinear regression model that takes into account the

GP prior and whose posterior is the desired regression function that belongs to an infinite dimension random function space [49].

To introduce the GP, denote by $\mathbf{t} = \left(t_1,\ \ldots,\ t_p\right)^\top$ the time training points associated with a finite set of p observations $\boldsymbol{\mathcal{D}} = \left(\mathcal{D}_1,\ \ldots,\ \mathcal{D}_p\right)^\top$. We assume that each observation $\mathcal{D}$ at location t is a random variable associated with the GP stochastic process

$$f(t) \sim \mathcal{GP}\left(m(t),\ k\left(t,\ t'\right)\right)\ ,$$

which is completely defined by its *mean function* $m(t) = \mathbb{E}[f(t)]$, the expected value of all functions in the distribution evaluated for an arbitrary input t, and $k\left(t,\ t'\right)$, the *covariance function* of $f(t)$, which describes the dependence between the function values for a pair of arbitrary input time points t and t', given by $k\left(t,\ t'\right) = \mathbb{E}\left[\left(f(t) - m(t)\right)\left(f\left(t'\right) - m\left(t'\right)\right)\right]$. We now consider the regression problem

$$\mathcal{D}_i = f\left(t_i\right) + \varepsilon \tag{4}$$

with ε being an additive Gaussian *noise* with zero mean and variance σ^2. The GPR begins by assuming the vector-value function $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ as the prior distribution, where $\mathbf{K}$ is the $p \times p$ *covariance matrix* whose entries are $k\left(t,\ t'\right)$. Considering this prior and noise in the time training set, as defined in Eq. (4), the joint distribution taking into account new input time points $\mathbf{t}^*$ and their associated output $\boldsymbol{\mathcal{D}}^*$ is given by

$$\begin{bmatrix} \boldsymbol{\mathcal{D}} \\ \boldsymbol{\mathcal{D}}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0},\ \begin{bmatrix} \mathbf{K}\left(\mathbf{t},\ \mathbf{t}\right) + \sigma^2\mathbf{I} & \mathbf{K}\left(\mathbf{t},\ \mathbf{t}^*\right) \\ \mathbf{K}\left(\mathbf{t}^*,\ \mathbf{t}\right) & \mathbf{K}\left(\mathbf{t}^*,\ \mathbf{t}^*\right) \end{bmatrix}\right),$$

where $\mathbf{I}$ stands for the $p \times p$ identity matrix. Therefore, the predictive equations for GPR are derived from the conditional distribution property for the multivariate Gaussian distribution [40]. Considering the Schur complement (for more details on Schur complements, refer to Puntanen and Styan [39]), the posterior predictive distribution is the multivariate Gaussian distribution

$$p\left(\boldsymbol{\mathcal{D}}^* \,|\, \mathbf{t}^*,\ \mathbf{t},\ \boldsymbol{\mathcal{D}}\right) = \mathcal{N}\left(\boldsymbol{\mu}^*,\ \boldsymbol{\Sigma}^*\right), \tag{5}$$

with mean

$$\boldsymbol{\mu}^* = \mathbf{K}\left(\mathbf{t}^*,\ \mathbf{t}\right)\left(\mathbf{K}\left(\mathbf{t},\ \mathbf{t}\right) + \sigma^2\mathbf{I}\right)^{-1}\boldsymbol{\mathcal{D}} \tag{6}$$

and covariance matrix

$$\boldsymbol{\Sigma}^* = \mathbf{K}\left(\mathbf{t}^*,\ \mathbf{t}^*\right) - \mathbf{K}\left(\mathbf{t}^*,\ \mathbf{t}\right)\left(\mathbf{K}\left(\mathbf{t},\ \mathbf{t}\right) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{K}\left(\mathbf{t},\ \mathbf{t}^*\right). \tag{7}$$

Therefore, the calculation of Eqs. (6) and (7) is sufficient to predict $\mathcal{D}^*$. Note that this involves first calculating the four covariance matrices. Furthermore, the covariance depends only on the time training set ($\mathbf{t}$) and the new input points ($\mathbf{t}^*$), and not on the observation measures vector ($\boldsymbol{\mathcal{D}}$). By applying the GPR to the observation set $\boldsymbol{\mathcal{D}}^{(j)}$, for $j \in \{P,\ D\}$, we then obtain $p\left(\boldsymbol{\mathcal{D}}^{(j)^*} \,|\, \mathbf{t}^*,\ \mathbf{t},\ \boldsymbol{\mathcal{D}}^{(j)}\right)$, whose corresponding mean values at the time training points are the regularized data used in this work.

The covariance function is commonly called the *kernel* of the GP. This function maps a pair of general input vectors $t,\ t' \in \mathbf{t}$ into $\mathbb{R}$. The idea behind the kernel is that if t and t' are said to be similar, it is expected that the function output (observations) at these points will also be similar. The main attribute of the kernel is to avoid the computation of an explicit nonlinear mapping function that relates input and output data, obtaining the identification of the mapping in the space where the number of parameters to be optimized, the so-called *hyperparameters*, is smaller [25]. Thus, the choice of an appropriate kernel is usually based on prior knowledge related to the behavior of the training data, as for example the occurrence of periodic oscillations, and assumptions such as smoothness [48]. Finding suitable properties for the kernel function is one of the main tasks for defining an appropriate GP.

The kernel can be any function that relates two input vectors, on the assumption that it can be formulated as an inner product, producing a positive semi-definite matrix [8]. Different functions can be combined to produce kernels with a variety of features, generally by both adding and multiplying kernels. Rasmussen and Williams [40] present a comprehensive analysis on the construction of kernel functions. Considering the ability of GPR to approximate the behavior of the used data, we adopt the radial basis function (RBF) $k\left(t,\ t'\right) = k_{\mathrm{RBF}}\left(t,\ t'\right)$, which is explicitly defined as

$$k_{\mathrm{RBF}}\left(t,\ t'\right) = \exp\left(-\frac{|t - t'|^2}{2\ell^2}\right),$$

where ℓ is the length scale of the kernel.

Rasmussen and Williams [40] present an algorithm for GPR employing Cholesky factorization to solve the

matrix inversion required by Eqs. (6) and (7). Practical implementations of GPR, such as those provided by the Scikit-learn library [37], which we have adopted in this work, use the Cholesky factorization instead of directly inverting the matrix $\left(\mathbf{K}\left(\mathbf{t},\ \mathbf{t}\right)+\sigma^2\mathbf{I}\right)$, since the procedure is faster and numerically more stable. For the Cholesky decomposition, the computational complexity is $p^3/6$. For our particular case where the training data sets to be regularized are small ($p = 210$, see Sect. 2.2), there are no stability issues. Furthermore, studies on GPR stability are provided in Refs. [6,18]. Hyperparameters are calculated using a computational routine internal to the library, which uses the L-BFGS-B algorithm [34] to obtain the optimal values. The optimizer is restarted 50 times in order to increase the chances of convergence to the optimal set of hyperparameters. In fact, the number of restarts is an arbitrary choice and, as the GPR is executed only once, the computational cost associated with this procedure is negligible.

2.6 Parameter and numerical experiment setups

In models with more compartments and which, in general, have more parameters, finding a unique set of parameter values that best fits some data may be unattainable. Different combinations of parameters can produce similar results for data fitting. This fact is a characteristic of the non-identifiability of parameters [41]. To overcome this issue, some of the biologically known parameters are fixed, namely the incubation period $\tilde{\sigma} = 1/5.8\ \text{day}^{-1}$ [30], the proportion of symptomatic infected individuals $\rho = 0.6$ [35], and both the recovery rate of infected and positively diagnosed individuals, defined as $\gamma_I = \gamma_P = 1/16.7\ \text{day}^{-1}$ [53]. Therefore, the parameters to be estimated are the rate of transmission β, rate of removal due to quarantine measures ω, and mortality rate d_P. When analyzing the influence of data regularization on the predictive capacity of the model, we considered a constant mortality rate. Later, we specifically analyze the gain of adopting d_P as a time-varying parameter, where the corresponding parameters (d_0 and d_1) must be estimated.

In order to define the initial conditions that make it possible to solve Eq. (1), first consider that the population of the state of Rio de Janeiro is approximately equal to $N = 17{,}264{,}943$ individuals, according to the last demographic census conducted by the Brazilian Institute of Geography and Statistics [9]. Using the reported data, it is possible to define the initial conditions for the number of infected and positively diagnosed individuals, which we assume to be the same at the beginning of the time series. The number of dead individuals on the first day considered in this analysis is also sufficient to define the initial condition for D. In addition, it is reasonable to assume $R(0) = 0$, since it is not expected to have recovered individuals at the outbreak of COVID-19. Therefore, the only initial condition for which there is no information from the reported data is related to the exposed individuals. Therefore, we assume $E(0)$ as a parameter to be estimated and the initial condition for the number of susceptible individuals is given by $S(0) = N - (E(0) + I(0) + P(0) + R(0) + D(0))$.

We partition data sets into two subsets, which we call training data and test data. The key idea behind this approach is to analyze the predictive capacity of the model, by comparing test data with short-term simulations, which are calculated using parameters estimated with training data. In this way, it is possible to compare the gain of using regularized data in the parameter estimation procedure, analyzing the results considering an actual scenario. Furthermore, we only consider short-term predictions in our analyses since the simulations are compared with original data. Therefore, we analyze scenarios in which we adopt 14 data points in the test set. Of note, other time windows could also be used.

Arbitrarily, we choose the minimum size of the training data set equal to 60. We set up 14 values in the test data set and calculate successive estimates of the parameters of the model, gradually increasing the proportion between training and test data. After each run, new data are added to the training set, so that the test set is composed of the next 14 values in the time series. As data for 210 days are available, 136 sets of parameters are estimated, using the deterministic approach described in Sect. 2.3. This procedure is performed both using training data as it stands and after regularization. The simulations using the optimal parameters are compared to the corresponding test set (without being regularized), by the computation of the normalized root-mean-square error (NRMSE) [19], considering both the cumulative number of infected and dead individuals—in this step, the cumulative data are adopted, with the purpose of minimizing the influence of noise. The root-mean-square error is normalized by the difference between the highest and lowest values in the corresponding data set.

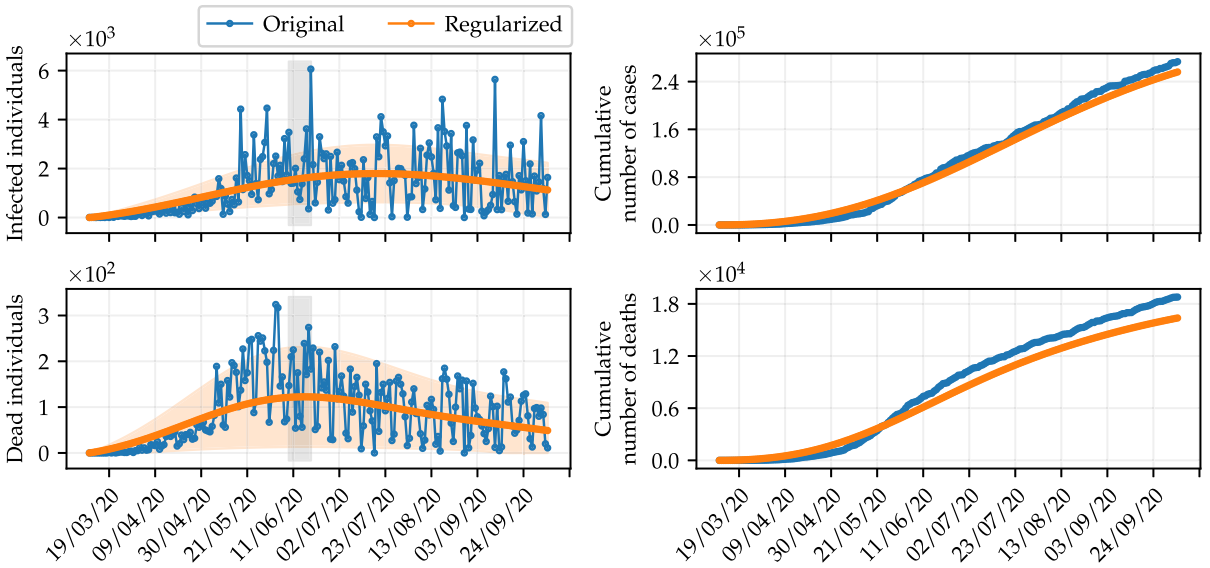

Fig. 2 Results of the regularization of data using GPR. The blue dots are the original data and the orange dots represent the resulting regularized data. Shaded areas indicate two standard deviations from the corresponding regularized data. We also show the cumulative data, for comparison purposes

All optimal parameters are calculated by combining the Differential Evolution [51] and the Nelder–Mead Simplex [33] methods. For each problem, the solution is estimated by Nelder–Mead Simplex and refined by Differential Evolution, which searches for the optimal parameters in the vicinity of the previously obtained point. The solution to each problem—the best individual in the population—is taken as an initial estimate for the next problem. Nelder–Mead Simplex runs with coefficients of reflection, expansion, contraction, and shrinkage equal to 1, 2, 0.5, and 0.5, respectively (as commonly adopted in the literature). In turn, Differential Evolution runs with 20 individuals in the population, amplification factor equal to 0.6, and crossover probability equal to 0.95. The search space is bounded by $0 \leq \beta \leq 10^{-6}$, $0 \leq \omega \leq 1$, $0 \leq d_P \leq 1$, and $0 \leq E(0) \leq 10^4$ in appropriate units (a.u.).

3 Results and discussion

3.1 Influence of regularization of data

The GPR was employed to generate regularized data. The values of the hyperparameters tuned for the RBF kernel are $\ell = 68.9$ and $\ell = 50.2$ for daily data of infected and dead individuals, respectively. Considering these values, we then employ the mean as the regularized data, and the variance provides the uncertainty range of the regularization (refer to Eqs. (6) and (7), respectively). Figure 2 shows an illustrative comparison between the original data for daily infected and dead individuals, alongside the corresponding, regularized data. Shaded areas represent two standard deviations from the fitting data. We also show the respective cumulative data, in order to allow a visual inspection of the agreement of the data resulting from the regularization, but smoothing out the noise. The mean values resulting from the regularization are used to estimate the model parameters.

Next, the influence of estimating the parameters of the SEIRPD-Q model, presented in Sect. 2.1, using the regularized data set, in relation to the approach that adopts the original data is analyzed. Our focus is to assess the gain related to the regularization of data within the scope of compartmental models and, therefore, we consider the simulations using the model adopted in this analysis. It is worth mentioning that the proposed analysis can be extended to any compartmental model with a structure similar to the model given by Eq. (1). The analysis is performed by evaluating the NRMSE between model predictions and test data.

Results concerning the optimal parameters calculated by Differential Evolution and Nelder–Mead Sim-

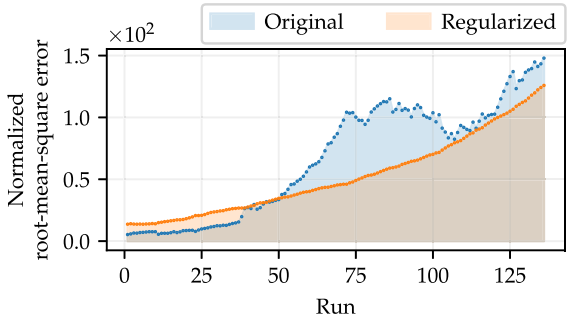

Fig. 3 NRMSEs computed by comparing simulations of the SEIRPD-Q model, performed using parameters that best fit the training data, where $\boldsymbol{\theta} = (\beta,\ \omega,\ d_P,\ E(0))$, in relation to the test set (composed of 14 points). The area under the curve represents the total deviation relative to the test data in all runs, where the number of training data varies. (Color figure online)

plex (Sect. 2.6) are shown in Fig. 3. For each set of parameters obtained in the 136 runs using the methodology described in Sect. 2.6, considering each type of data (original and regularized), the model is simulated and we calculate the corresponding NRMSE. Then we compute the area under the curve, which is composed of the values of the NRMSEs. Analyzing Fig. 3, the effect of data regularization on parameter estimation is clear: for most of the estimated parameter sets, the corresponding simulations have better agreement with the test data. Translating into numbers, regularized data resulted in more reliable predictions in 64.71% of runs. The area under the curve for original data is approximately equal to 899.66 u. a. (unit of area), whereas for regularized data it is 705.70 u. a. This represents an average improvement of approximately 21.56%. Note that the effect of the regularization is more prominent when the data set to be fitted is larger since the influence of the noise tends to become more intensive.

The implication of using regularized data is even more straightforward when we analyze the variability of simulations resulting from the optimal parameters corresponding to each point in Fig. 3. For this purpose, consider the results shown in Fig. 4, in which the shaded area represents the range of the simulations related to the daily number of infected and dead individuals, for both original and regularized data, whose parameters are estimated by varying the amount of training data, as previously described. In turn, the points represent the corresponding data, which together with the box-and-whisker diagrams, aim to demonstrate the variability of the simulations on specific days. We also show the corresponding cumulative values of the results obtained, in order to provide better conditions for comparison. It is worth mentioning that the set of training data, even presenting variable size considering each set of parameters that are estimated, is completely shown in all cases.

It is clear that simulations resulting from parameters estimated using data with no regularization have more variability than after regularization. Considering that on October 5, 2020 (the last day of the simulation), Rio de Janeiro had 273,335 confirmed cases, the boundary values of the shaded area on this day, for the cumulative number of infected individuals obtained with original data are 146,763.4 and 36,518,946.5, whereas the quantities obtained with regularized data vary between 80,762.4 and 253,387.8. In the case of dead individuals, the variation is between 8,066.9 and 3,579,976.6 for original data, and between 6,950.6 and 17,528.7 for regularized data, with the actual cumulative number of dead individuals on this day being 18,780.

The dispersion of these results can be alternatively analyzed using a box-and-whisker diagram. The boxes are bounded by the lower and upper quartiles (which we call Q1 and Q3, respectively) of the full model prediction values, and the horizontal line inside the box represents the median. Whiskers, the vertical lines bounded by perpendicular dashes, extend from the minimum value of the data to the first quartile, and from the third quartile to the maximum value of the data. Whiskers express the variability outside these quartiles. They range from the lowest to the highest values of each set, both for original and regularized data, since no simulation can be considered an outlier. The statistical results associated with the box-and-whisker diagrams in Fig. 4 are summarized in Table 1. It is clear that regularization using GPR preserves the behavior of the data, although substantially reducing the variability of the simulations. The reduction in variability provides evidence that regularizing the data with the GPR does not cause problems related to overfitting. Simulations influenced by overfitting would deviate significantly from the real data. In this case, parameter estimation using regularized data provides a means of reducing noise without negatively influencing forecasts [31].

Now consider a forecast of the epidemic in terms of the cumulative number of infected and dead individuals, aiming to show the difference of the simulations in relation to the analyzed data. For this purpose, we adopt the Bayesian approach for parameter estimation, presented in Sect. 2.4. We arbitrarily choose the training

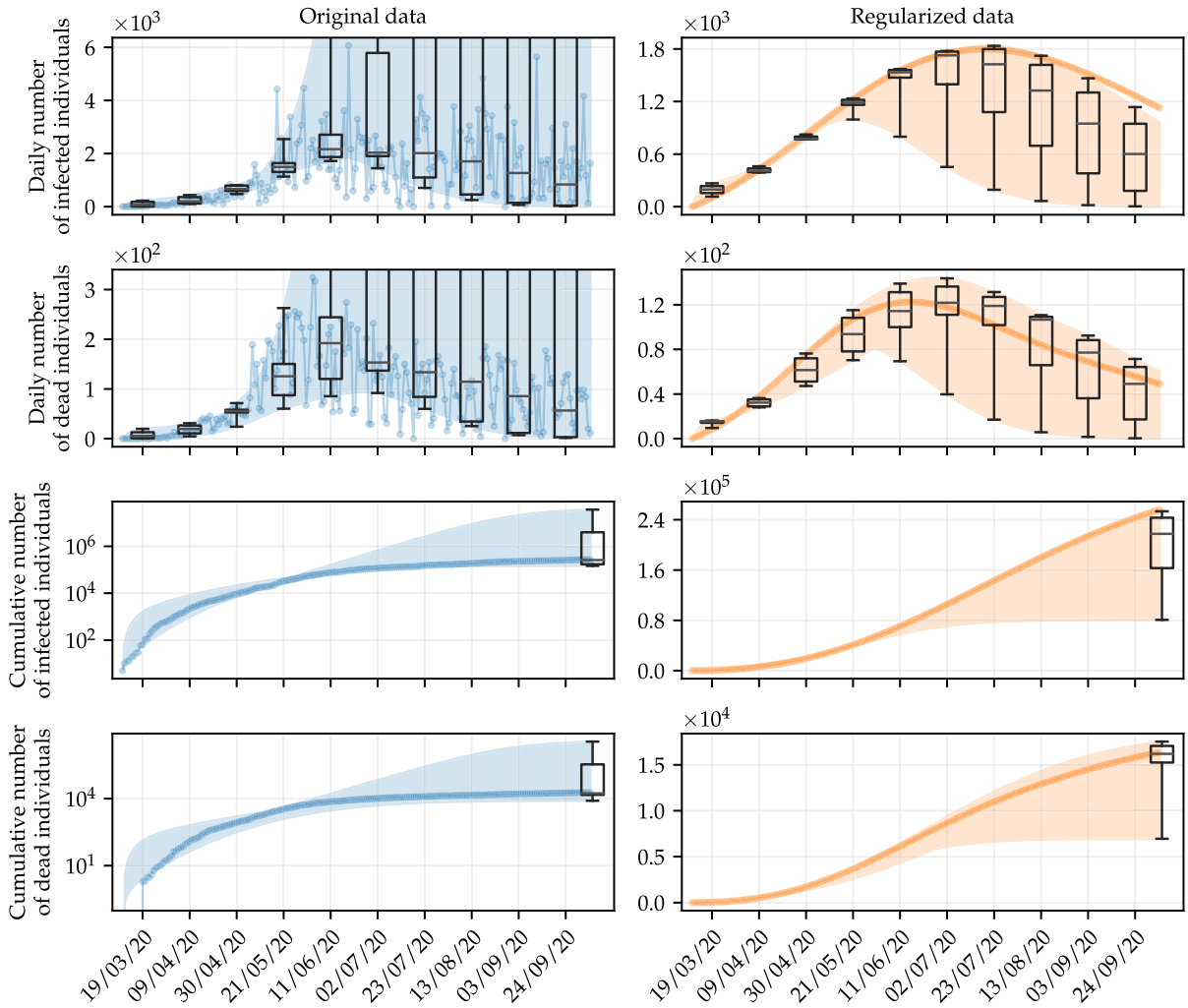

Fig. 4 Simulations of the daily number of infected and dead individuals using optimal parameters obtained with the deterministic approach, by varying the amount of data in the training set. The shaded areas represent the variation range of the simulations, whereas the points are the fitted data. Box-and-whisker diagrams are used to show the variability of the simulations on specific days. We also show the corresponding results for cumulative data, where the noise is less effective, for comparison purposes. The results follow the same color scheme as in Fig. 2, blue for original data, and orange for regularized data. (Color figure online)

data sets for the daily number of infected and dead individuals, both original and regularized, with 196 values, the maximum number of elements that the training set can contain. In turn, the test set is made up of the next 14 values in the time series. By this analysis, we present a visual perspective of the benefit of using regularized data in the parameter estimation procedure, providing a way of comparing the agreement between simulations and test data.

The parameters to be estimated are the same as in the previous analysis, and are assumed to be uniformly distributed, in such a way that $\beta \sim \mathcal{U}\left(0,\ 10^{-6}\right)$, $\omega \sim \mathcal{U}\left(0,\ 1\right)$, $d_P \sim \mathcal{U}\left(0,\ 1\right)$, and $E(0) \sim \mathcal{U}\left(0,\ 10^{4}\right)$, in a.u. Table 2 shows the *maximum a posterior* (MAP) estimates of the parameters, as well as the corresponding 95% credible interval (CI) [5]. Likewise, all other parameters of the SEIRPD-Q model are taken as biological parameters and have the same values previously reported (see Section 2.6). Figure 5 shows the pos-

Table 1 Statistical results of all 136 simulations performed using the optimal parameters. The results refer to the cumulative number of infected and dead individuals, that is, $\mathbf{Y}_p^{(j)}$ for $j \in \{P,\ D\}$, obtained using both original and regularized data on the last day that the model was simulated, October 5, 2020. On this day, Rio de Janeiro accumulated 273,335 confirmed cases and 18,780 deaths

	Original data		Regularized data	
	$\mathbf{Y}_p^{(P)}$	$\mathbf{Y}_p^{(D)}$	$\mathbf{Y}_p^{(P)}$	$\mathbf{Y}_p^{(D)}$
Min	146,763.4	8,066.9	80,762.4	6,950.6
Max	36,518,946.5	3,579,976.6	253,387.8	17,528.7
Median	259,406.5	17,313.5	217,621.5	16,196.7
Q1	172,065.9	14,796.5	163,064.3	15,249.4
Q3	3,970,094.4	342,994.5	243,274.8	17,059.5

Table 2 MAP values and 95% CIs of the parameters estimated using Bayesian calibration (in a.u.)

	Data type	
	Original	Regularized
β	1.5248×10^{-8}	1.2419×10^{-8}
	$(1.2795,\ 1.7991) \times 10^{-8}$	$(1.2206,\ 1.2639) \times 10^{-8}$
ω	0.007121	0.004859
	(0.005704, 0.008576)	(0.004738, 0.004996)
d_P	0.07126	0.0625
	(0.06215, 0.08161)	(0.06030, 0.06475)
$E(0)$	1190.8021	1236.5740
	(655.0367, 1818.3565)	(1168.6603, 1305.0816)

terior distribution of the estimated parameters of the SEIRPD-Q model, obtained when original (blue bins) and regularized (orange bins) data are employed for calibration. In the right frame, violin plots express the variance of the inferred parameters using original and regularized data. Of note, the latter is much narrower than the former. This fact is reflected in the stochastic simulation of the SEIRPD-Q model shown in Fig. 6, along with the training and test data sets (note that the latter is never regularized). The solid curves represent the model responses when the free parameters are set to be the respective MAP values, shown in Table 2. In turn, the shading around the curves represents the 95% CI, the discrete points are the available data, and the ranges of the training and test sets are colored in gray for daily and cumulative data, respectively. Results related to infected individuals are shown in red, whereas dead individuals are shown in green.

Analyzing Fig. 6, it is clear that the Bayesian inference obtained suitable results, using both original and regularized data. The deviation between the simulations and the cumulative data in the test data range when regularized data are used for calibration is due to the regularization calculated using GPR, as can be seen in Fig. 2. Nevertheless, the regularization of the training data seems to make the predictive capacity of the model less unstable. Thus, the model could better capture the behavior of the data, eventually leading to more appropriate predictions. Of note, the narrower posterior distributions obtained by performing the calibration with the regularized data resulted in less uncertainty about the values predicted by the model, as can be seen in Fig. 6. This is because the posterior distributions obtained when data are regularized tend to have a smaller standard deviation, as can be seen in the right frame of Fig. 5 and in the CIs shown in Table 2.

3.2 Influence of time-varying parameters

A further feature that may improve the predictive capacity of compartmental models is the adoption of time-varying parameters. In an effort to analyze this aspect, consider the 136 sets of optimal parameters, obtained using the deterministic approach that led to the results shown in Fig. 3. Figure 7 shows the values of the calibrated parameters for each corresponding run, both using original and regularized data. Initially, note that the variability of the set of parameters obtained by fitting the original data is much higher than those referring to regularized data. This is a consequence of the high level of noise in the data. Note, for instance, the

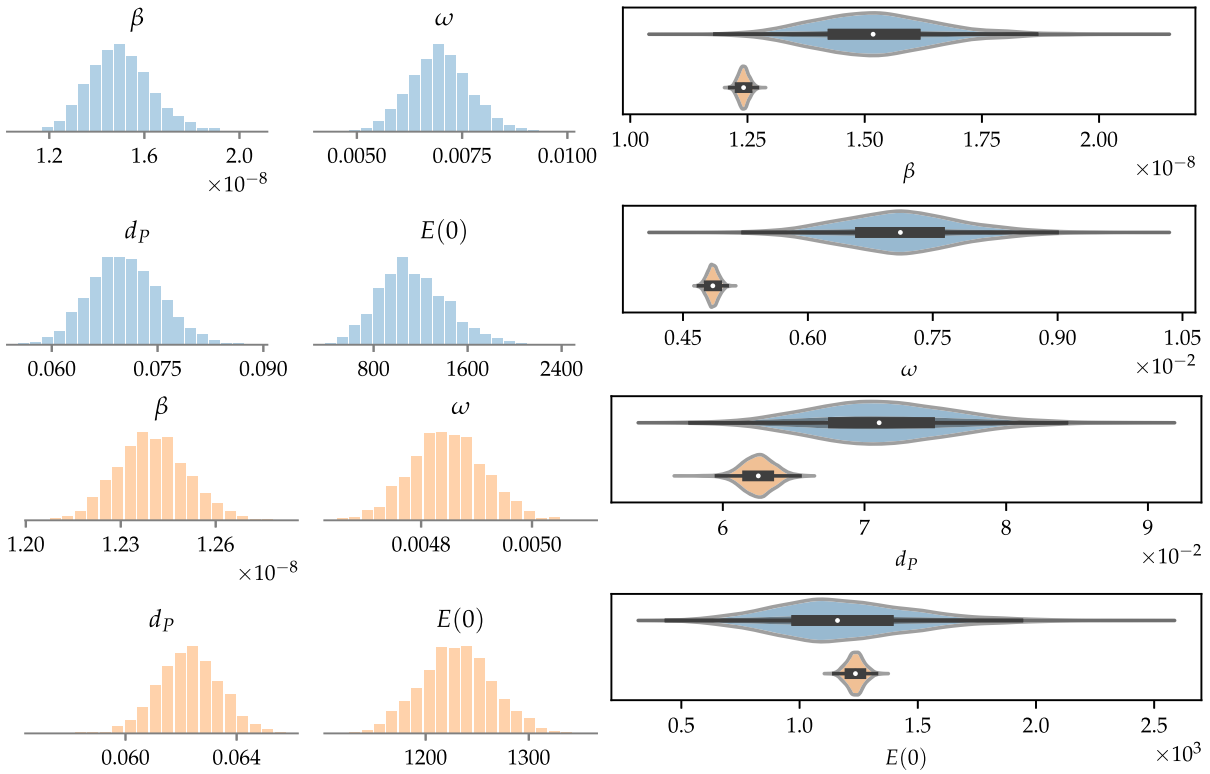

Fig. 5 The left frame shows the posterior distribution of the parameters obtained in the Bayesian inference for original (in blue) and regularized (in orange) data, by fitting the daily number of infected and dead individuals; the right frame illustrates the variance of each parameter in a comparative way, where the same color scheme is adopted. (Color figure online)

gray shaded areas in Fig. 7. They refer to the same runs for all parameters and correspond to the shaded area in the same color in Fig. 2. In this time window, the data show very large variations on subsequent days, especially those of the daily number of infected individuals. These variations influence the values of the parameters obtained, giving rise to a great difference in the value of the set of parameters that best fits the data just by adding a new single value to the training set, which does not occur with regularized data.

The optimal parameters in Fig. 7 express some meaningful facts: first, the behavior of the initial conditions for exposed individuals reveals that defining this value just as a fixed proportion of the population size can undermine the capacity of the method for solving the system of differential equations; second, the parameters β and ω exhibit similar behaviors (especially when considering the results obtained with regularized data). Over time, the rate of contact between susceptible and infected individuals decreases (assuming the hypothesis that recovered individuals are immune for some time), so that quarantine measures end up being eased, which is reflected in the reduction of isolation measures. Definitely, political and social interests also play a significant role in this behavior.

The inspection of the behavior of d_P in Fig. 7 indicates that the mortality rate of positively diagnosed individuals basically only decreases after a certain run, around the gray shaded area. This behavior suggests that the parameter can be approximated by a function. In this case, we propose to represent d_P as a function of the form

$$d_P(t) = d_0 \exp(-d_1 t). \tag{8}$$

This approach assumes an inherent error related to the first runs, at expense of better representing the parameter behavior in later times. Thus, d_P becomes variable over time, and the vector of parameters to be estimated is formed by $\boldsymbol{\theta} = (\beta,\ \omega,\ d_0,\ d_1,\ E(0))$.

We then repeated the methodology used previously to investigate the impact of using the time-dependent

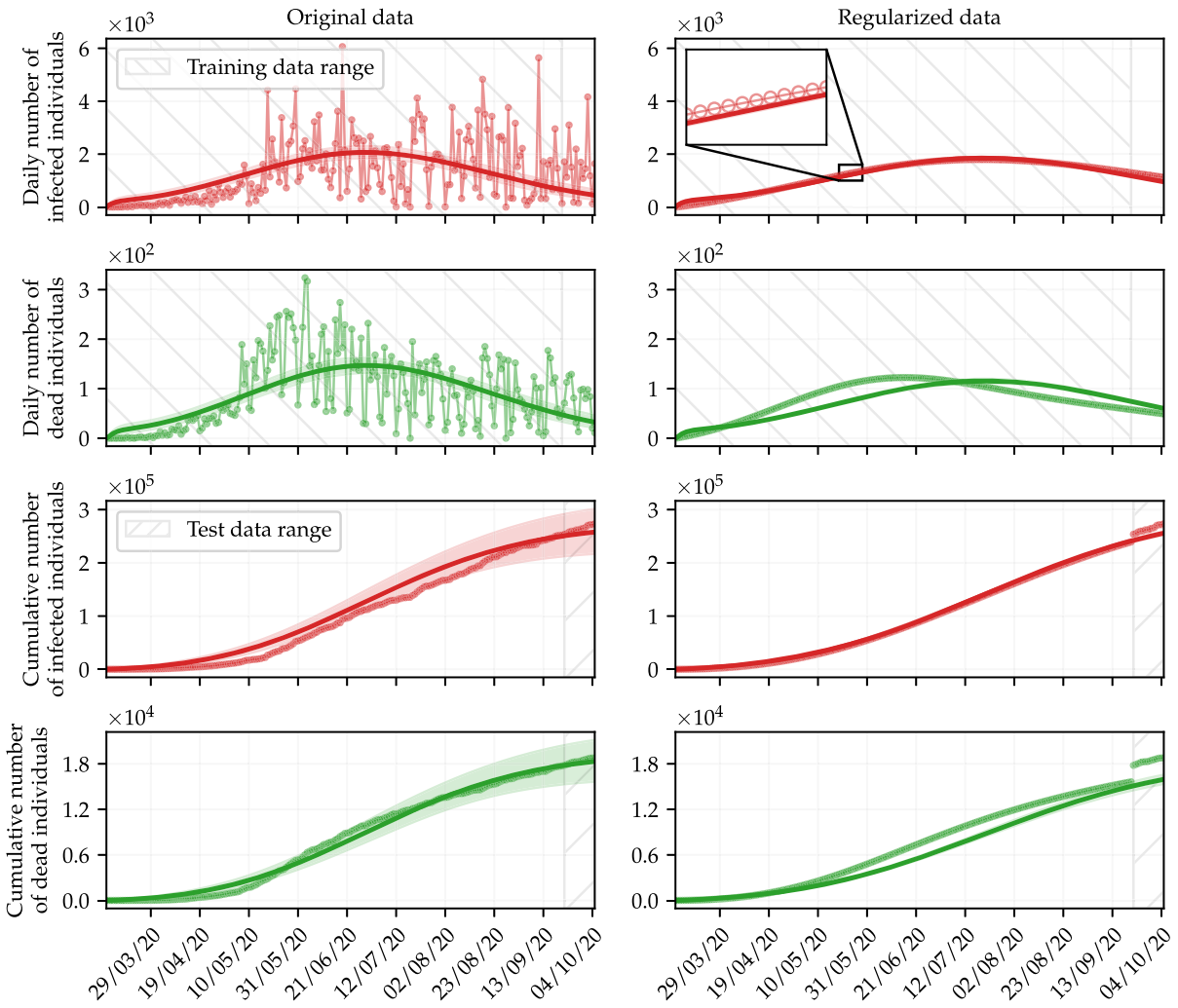

Fig. 6 Simulations with parameters obtained by fitting the daily number of infected (in red) and dead (in green) individuals using the Bayesian approach. The shaded areas refer to the 95% CI (see Table 2). Hatched areas bound the range of training and test data. Training data is never regularized. (Color figure online)

$d_P(t)$ defined in Eq. (8) on the model outcomes. Specifically, we sequentially estimated the new 136 sets of parameters $\boldsymbol{\theta}$ as additional data, original and regularized, were gradually included in the analysis. The search intervals for the new parameters are $0 \leq d_0 \leq 1$ and $0 \leq d_1 \leq 1$ and all other parameters follow the same specifications defined before. Figure 8 shows that this strategy is reflected in the reduction of the NRMSE considering the test data set with 14 values, in most simulated forecasts. The NRMSEs considering regularized data and time-varying d_P (orange dots) represent a better approximation of the test data set in 73.53% of the analyzed runs, in relation to the results considering original data (blue dots). The area under the curve obtained with regularized data and $d_P(t)$ is 733.06 u. a., whereas for original data the area is equal to 1427.28 u. a., which represents a reduction of approximately 48.64%.

As in Fig. 7, we are interested in understanding the behavior of the parameters inherent to the function when d_P varies over time. Figure 9 shows the parameters d_0 and d_1, as well as the other calibrated parameters of the model, calculated in each parameter estimation procedure using both original and regularized data. The color scheme follows what has been adopted, blue for original data and orange for regularized data.

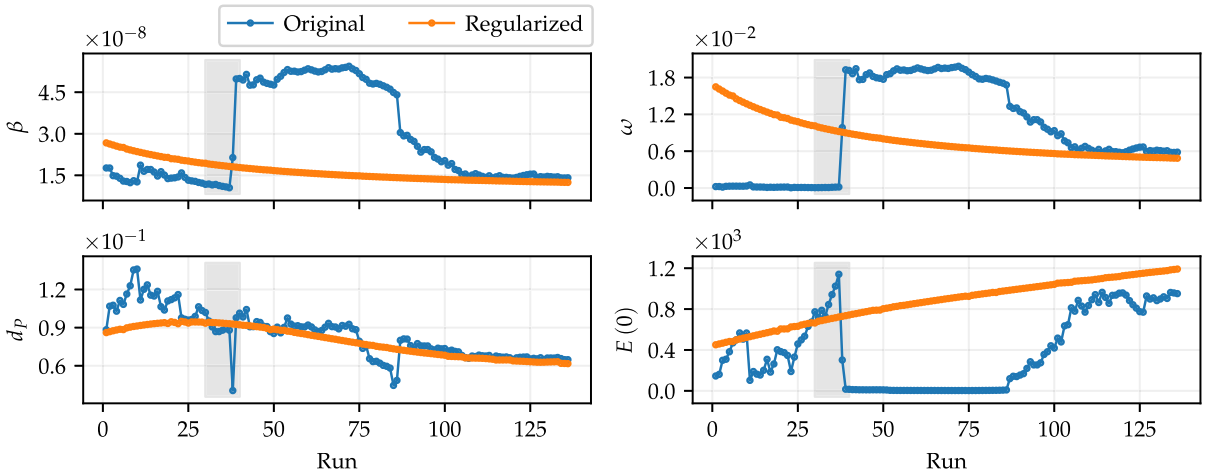

Fig. 7 Optimal parameters obtained by fitting the daily data of infected and dead individuals, for $\boldsymbol{\theta} = (\beta,\ \omega,\ d_P,\ E(0))$. Each run is associated with a training set with a specific size, from 60 to 196 data in each set. The test data refer to the 14 subsequent data from the corresponding run. Parameters obtained by fitting original data are shown in blue, and regularized data are shown in orange. (Color figure online)

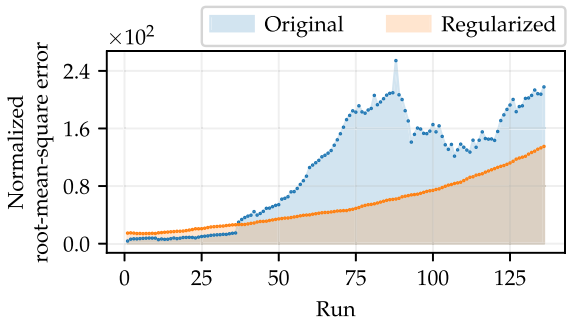

Fig. 8 NRMSEs computed by comparing simulations of the SEIRPD-Q model, performed using parameters that best fit the training data, where $\boldsymbol{\theta} = (\beta,\ \omega,\ d_0,\ d_1,\ E(0))$, in relation to the test set (composed of 14 points). The area under the curve represents the total deviation relative to the test data in all runs, where the number of training data varies

In addition, Fig. 9 shows the range of Eq. (8), taking the optimal values of d_0 and d_1, alongside the values of d_P (shown in Fig. 7) for both original and regularized data. In this regard, we are interested in analyzing the behavior of $d_P(t)$ in terms of the values obtained for the case where d_P is constant.

In the first runs, $d_P(t)$ is expressed by a nearly constant function, since the values of d_1 are very close to zero (see the behavior of d_1 in Fig. 9). In this case, $d_P(t) \approx d_0$. These curves practically coincide with the first points shown in the frame corresponding to $d_P(t)$ obtained with regularized data in Fig. 9, which means that, in fact, the calibrations are quite similar. This behavior occurs because, given the training set for such runs, the model identifies that the mortality rate is not decreasing and, therefore, the calibration procedure estimates the most suitable function for such data (by means of d_0 and d_1), which in this case is nearly constant, without loss of generality.

As the training set gets larger, $d_P(t)$ starts to behave as expected, exponentially decreasing. In this case, the first calibrations provide functions relatively distant from the corresponding points in Fig. 9, especially in the early times. In the last runs, the functions show good agreement with the compared points. However, it is important to note that $d_P(t)$ is not expected to represent the exact behavior of such data in the long run. In general, when the mortality rate varies over time, the compartmental model may be more capable of capturing the dynamics of the data, allowing for more accurate predictions. This hypothesis is supported by the last results obtained in Fig. 8, where the orange dots always represent the best approximation in relation to the test data.

4 Conclusions

In this work, the focus was to provide methodologies capable of improving the predictive capacity of compartmental models for intrinsically noisy data such as the COVID-19 pandemic. Our contribution is divided

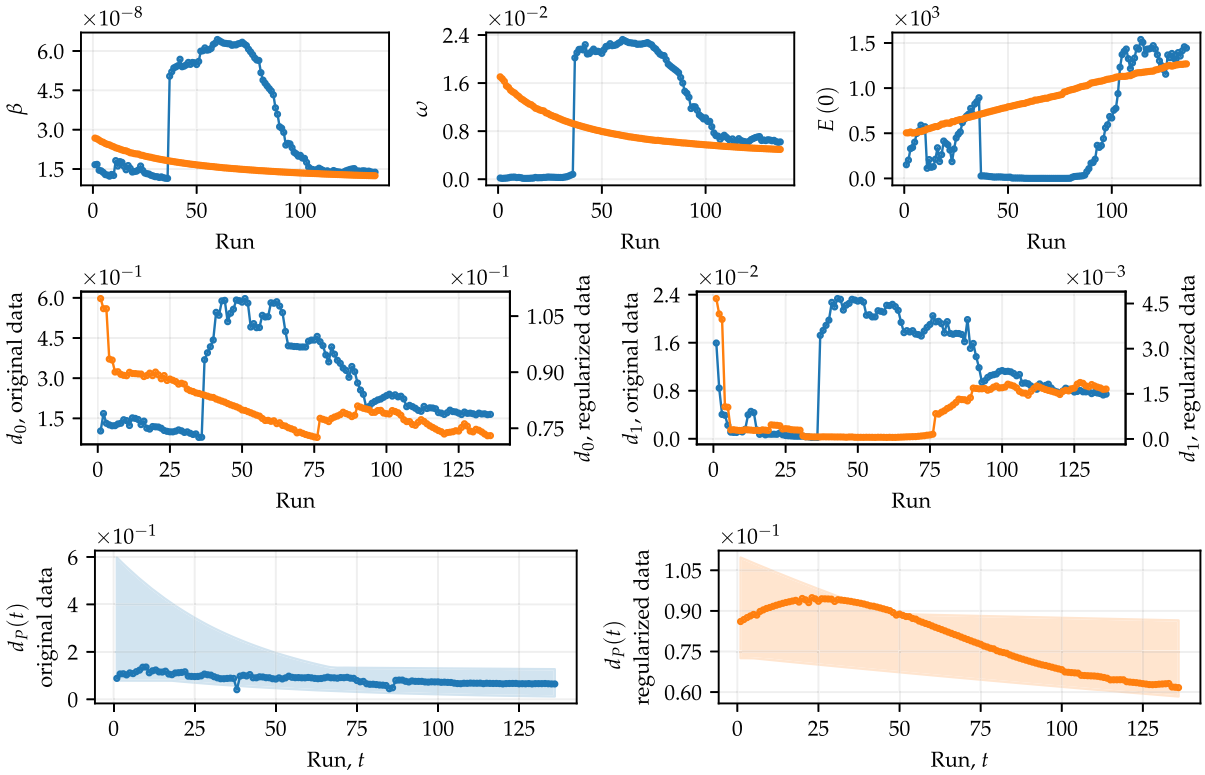

Fig. 9 Optimal parameters obtained in a procedure similar to that of Fig. 7, for $\boldsymbol{\theta} = (\beta,\ \omega,\ d_0,\ d_1,\ E(0))$. We also show the interval that includes the curves of $d_P(t)$, for each optimal value of d_0 and d_1 associated with the same run. In addition, we also show the optimal values of d_P presented in Fig. 7, for purposes of comparison with the curves of $d_P(t)$

into two parts: first, we proposed the use of GPR to regularize the training data. Second, we proposed the use of a time-varying parameter, rather than being constant as usual, in order to increase the descriptiveness of the model. In summary, we conducted successive predictions, increasing the amount of data in the training set, so that the variability of the model outcomes was clearly reduced when such strategies were adopted for estimating the parameters of the model, leading to more accurate predictions.

Our study provides a framework that aims to increase the predictive capacity of compartmental models. This is relevant from the point of view that the proposed strategies can be extended to other data sets and compartmental models. Especially in the context of the epidemiological modeling of COVID-19, approaches of this type can be useful, taking into account the wide range of existing compartmental models and the fact that the data analyzed here are similar to others regarding noise.

This study has gone some way toward enhancing our understanding of the influence of noise on the estimation of parameters of compartmental models. The work has revealed that the regularization of data by means of GPR is effective to mitigate the effect of noise in a given parameter calibration. We analyzed the gain of data regularization in parameter estimation procedures in the scope of compartmental models. Therefore, we estimated 136 sets of parameters, using both original and regularized data, increasing the amount of training data by one. We have shown that simulations with estimated parameters using regularized data have more agreement with test data in general. Furthermore, NRMSE values obtained with simulations computed using parameters estimated with regularized data were below the corresponding results with original data in most cases. Another relevant contribution is the smaller variability of the set of optimal parameters obtained in successive calibrations when regularized data are used. Since this procedure must not be repeatedly applied, in

the context to which it is proposed, the computational cost of the GPR can be considered irrelevant.

Our research also suggests that it may be useful to use time-varying parameters, to the detriment of the usual approach that adopts constant parameters. We have defined a functional form for the mortality rate of the compartmental model based on the set of optimal parameters obtained through successive calibrations considering the parameter in each calibration constant. Through the analysis of the NRMSE values, we have noticed that the descriptive capacity of the model has increased when the time-varying parameter was adopted, causing the variability of the simulations to be reduced and, therefore, increasing the agreement between the simulations and the test data. It incorporates additional degrees of freedom into the model, in order to provide more flexibility to describe the behavior of the analyzed data. The choice of such a function depends on several factors, as for example the physical meaning of the parameter and the additional parameters inherent to the function. This analysis can be conducted for any model and, clearly, the gain is related to the appropriate choice of the function, especially if many parameters to be estimated are included.

The proposed GPR regularization can be used as an accessory approach in problems of engineering and sciences, whenever noisy data can compromise or preclude the desired analysis. We are currently using data regularization through GP, together with a compartmental model with a time-varying parameter, to analyze the effect of vaccination on the population of the state of Rio de Janeiro, under various aspects, such as any delays in the start of vaccination, people who refuse to get vaccinated (or refuse to receive the second dose), and the burden of slow vaccination. In other research, we use GPs to regularize spatial data from susceptible, exposed, and infected individuals, in order to take such values as initial conditions of models with temporal and spatial dependence.

Funding The authors would like to thank the Ministry of Science, Technology, Innovation, and Communication (MCTIC) of Brazil. Lucas dos Anjos was supported by a postdoctoral fellowship from the Institutional Training Program (PCI) of the Brazilian National Council for Scientific and Technological Development (CNPq), Grant Number 301327/2020-3. Gustavo Barbosa Libotte was supported by a postdoctoral fellowship from the Institutional Training Program (PCI) of the Brazilian National Council for Scientific and Technological Development (CNPq), Grant Number 303185/2020-1, and is currently supported by a postdoctoral fellowship from the Carlos Chagas Filho Foundation for Supporting Research in the State of Rio de Janeiro (FAPERJ), Grant Number E-26/200.347/2021.

Availability of data and material The data used in this work are publicly available according to Ref. [13].

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Code availability The source code used to generate the results is publicly available at https://github.com/gustavolibotte/enhancing-forecast-COVID-19github.com/gustavolibotte/enhancing-forecast-COVID-19.

References

1. Ahmad, A., Garhwal, S., Ray, S.K., Kumar, G., Malebary, S.J., Barukab, O.M.: The number of confirmed cases of Covid-19 by using machine learning: methods and challenges. Arch. Comput. Methods Eng. **28**(4), 2645–2653 (2021). https://doi.org/10.1007/s11831-020-09472-8
2. Alberti, T., Faranda, D.: On the uncertainty of real-time predictions of epidemic growths: a COVID-19 case study for China and Italy. Commun. Nonlinear Sci. Numer. Simul. **90**, 105372 (2020). https://doi.org/10.1016/j.cnsns.2020.105372
3. Arias Velásquez, R.M., Mejía Lara, J.V.: Gaussian approach for probability and correlation between the number of COVID-19 cases and the air pollution in Lima. Urban Clim. **33**, 100664 (2020). https://doi.org/10.1016/j.uclim.2020.100664
4. Arroyo-Marioli, F., Bullano, F., Kucinskas, S., Rondón-Moreno, C.: Tracking R of COVID-19: A new real-time estimation using the Kalman filter. PLoS ONE **16**(1), 0244474 (2021). https://doi.org/10.1371/journal.pone.0244474
5. Bailer-Jones, C.A.L.: Practical Bayesian Inference. Cambridge University Press, Cambridge (2017). https://doi.org/10.1017/9781108123891
6. Banerjee, A., Dunson, D.B., Tokdar, S.T.: Efficient Gaussian process regression for large datasets. Biometrika **100**(1), 75–89 (2013). https://doi.org/10.1093/biomet/ass068
7. Bhopal, S.S., Bhopal, R.: Sex differential in COVID-19 mortality varies markedly by age. Lancet **396**(10250), 532–533 (2020). https://doi.org/10.1016/S0140-6736(20)31748-7
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)
9. Brazilian Institute of Geography and Statistics: Demographic Census (2020). https://www.ibge.gov.br/cidades-e-estados/rj/rio-de-janeiro.html. Accessed 20 Nov 2020
10. Calvetti, D., Hoover, A.P., Rose, J., Somersalo, E.: Metapopulation network models for understanding, predicting, and

managing the coronavirus disease COVID-19. Front. Phys. **8**, 261 (2020). https://doi.org/10.3389/fphy.2020.00261
11. Ching, J., Chen, Y.C.: Transitional Markov Chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. J. Eng. Mech. **133**(7), 816–832 (2007). https://doi.org/10.1061/(ASCE)0733-9399(2007)133:7(816)
12. Clark, A., Jit, M., Warren-Gash, C., Guthrie, B., Wang, H.H.X., Mercer, S.W., Sanderson, C., McKee, M., Troeger, C., Ong, K.L., Checchi, F., Perel, P., Joseph, S., Gibbs, H.P., Banerjee, A., Eggo, R.M., Nightingale, E.S., O'Reilly, K., Jombart, T., Edmunds, W.J., Rosello, A., Sun, F.Y., Atkins, K.E., Bosse, N.I., Clifford, S., Russell, T.W., Deol, A.K., Liu, Y., Procter, S.R., Leclerc, Q.J., Medley, G., Knight, G., Munday, J.D., Kucharski, A.J., Pearson, C.A.B., Klepac, P., Prem, K., Houben, R.M.G.J., Endo, A., Flasche, S., Davies, N.G., Diamond, C., van Zandvoort, K., Funk, S., Auzenbergs, M., Rees, E.M., Tully, D.C., Emery, J.C., Quilty, B.J., Abbott, S., Villabona-Arenas, C.J., Hué, S., Hellewell, J., Gimma, A., Jarvis, C.I.: Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. Lancet Glob. Health **8**(8), e1003–e1017 (2020). https://doi.org/10.1016/S2214-109X(20)30264-3
13. Cota, W.: Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level. SciELOPreprints:362 (2020). https://doi.org/10.1590/scielopreprints.362
14. Cupertino, G.A., Cupertino, M.D.C., Gomes, A.P., Braga, L.M., Siqueira-Batista, R.: COVID-19 and Brazilian indigenous populations. Am. J. Trop. Med. Hyg. **103**(2), 609–612 (2020). https://doi.org/10.4269/ajtmh.20-0563
15. Currie, C.S., Fowler, J.W., Kotiadis, K., Monks, T., Onggo, B.S., Robertson, D.A., Tako, A.A.: How simulation modelling can help reduce the impact of COVID-19. J. Simul. **14**(2), 83–97 (2020). https://doi.org/10.1080/17477778.2020.1751570
16. Davies, N.G., Klepac, P., Liu, Y., Prem, K., Jit, M., Eggo, R.M.: Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat. Med. **26**(8), 1205–1211 (2020). https://doi.org/10.1038/s41591-020-0962-9
17. Edridge, A.W.D., Kaczorowska, J., Hoste, A.C.R., Bakker, M., Klein, M., Loens, K., Jebbink, M.F., Matser, A., Kinsella, C.M., Rueda, P., Ieven, M., Goossens, H., Prins, M., Sastre, P., Deijs, M., van der Hoek, L.: Seasonal coronavirus protective immunity is short-lasting. Nat. Med. (2020). https://doi.org/10.1038/s41591-020-1083-1
18. Foster, L., Waagen, A., Aijaz, N., Hurley, M., Luis, A., Rinsky, J., Satyavolu, C., Way, M.J., Gazis, P., Srivastava, A.: Stable and efficient Gaussian Process calculations. J. Mach. Learn. Res. **10**, 857–882 (2009)
19. Gujarati, D.N., Porter, D.C.: Basic Econometrics, 5th edn. McGraw-Hill Irwin, New York (2008)
20. Hartikainen, J., Särkkä, S.: Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In: 2010 IEEE International Workshop on Machine Learning for Signal Processing, pp. 379–384. IEEE (2010). https://doi.org/10.1109/MLSP.2010.5589113. http://ieeexplore.ieee.org/document/5589113/
21. Holmdahl, I., Buckee, C.: Wrong but useful—what COVID-19 epidemiologic models can and cannot tell us. N. Engl. J. Med. **383**(4), 303–305 (2020). https://doi.org/10.1056/NEJMp2016822
22. Ioannidis, J.P.A.: Coronavirus disease 2019: The harms of exaggerated information and non-evidence-based measures. Eur. J. Clin. Invest. **50**(4), 13222 (2020). https://doi.org/10.1111/eci.13222
23. Jia, J., Ding, J., Liu, S., Liao, G., Lin, J., Duan, B., Wang, G., Zhang, R.: Modeling the control of COVID-19: impact of policy interventions and meteorological factors. Electron. J. Differ. Equ. **2020**(23), 1–24 (2020)
24. Ketu, S., Mishra, P.K.: Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection. Appl. Intell. (2020). https://doi.org/10.1007/s10489-020-01889-9
25. Kocijan, J.: Modelling and Control of Dynamic Systems Using Gaussian Process Models. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-21021-6_2
26. Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., Shaman, J.: Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science **368**(6490), 489–493 (2020). https://doi.org/10.1126/science.abb3221
27. Link, W.A., Barker, R.J.: Bayesian Inference. Academic Press, London (2010). https://doi.org/10.1016/B978-0-12-374854-6.00004-1
28. Marson, F.A.L.: COVID-19—6 million cases worldwide and an overview of the diagnosis in Brazil: a tragedy to be announced. Diagn. Microbiol. Infect. Dis. **98**(2), 115113 (2020). https://doi.org/10.1016/j.diagmicrobio.2020.115113
29. Massonis, G., Banga, J.R., Villaverde, A.F.: Structural identifiability and observability of compartmental models of the COVID-19 pandemic (2020)
30. McAloon, C., Collins, Á., Hunt, K., Barber, A., Byrne, A.W., Butler, F., Casey, M., Griffin, J., Lane, E., McEvoy, D., Wall, P., Green, M., O'Grady, L., More, S.J.: Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. BMJ Open **10**(8), e039652 (2020). https://doi.org/10.1136/bmjopen-2020-039652
31. Mohammed, R.O., Cawley, G.C.: Over-fitting in model selection with Gaussian Process Regression. In: Perner, P. (ed.) Machine Learning and Data Mining in Pattern Recognition, pp. 192–205. Springer, Cham (2017)
32. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge (2012)
33. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**(4), 308–313 (1965). https://doi.org/10.1093/comjnl/7.4.308
34. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (2006). https://doi.org/10.1007/978-0-387-40065-5
35. Oran, D.P., Topol, E.J.: Prevalence of asymptomatic SARS-CoV-2 Infection. Ann. Intern. Med. **173**(5), 362–367 (2020). https://doi.org/10.7326/M20-3012
36. Overton, C.E., Stage, H.B., Ahmad, S., Curran-Sebastian, J., Dark, P., Das, R., Fearon, E., Felton, T., Fyles, M., Gent, N., Hall, I., House, T., Lewkowicz, H., Pang, X., Pellis, L., Sawko, R., Ustianowski, A., Vekaria, B., Webb, L.: Using statistics and mathematical modelling to understand infectious disease outbreaks: COVID-19 as an example. Infect.

Dis. Model. **5**, 409–441 (2020). https://doi.org/10.1016/j.idm.2020.06.008
37. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
38. Polidoro, M., de Assis Mendonça, F., Meneghel, S.N., Alves-Brito, A., Gonçalves, M., Bairros, F., Canavese, D.: Territories under siege: risks of the decimation of indigenous and quilombolas peoples in the context of COVID-19 in south Brazil. J Rac. Ethnic Health Disparit. (2020). https://doi.org/10.1007/s40615-020-00868-7
39. Puntanen, S., Styan, G.P.H.: Schur complements in statistics and probability. In: Zhang, F. (ed.) The Schur Complement and Its Applications, pp. 163–226. Springer, US, Boston (2005). https://doi.org/10.1007/0-387-24273-2_7
40. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge (2006)
41. Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., Timmer, J.: Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics **25**(15), 1923–1929 (2009). https://doi.org/10.1093/bioinformatics/btp358
42. Reece, S., Roberts, S.: An introduction to Gaussian processes for the Kalman filter expert. In: 2010 13th International Conference on Information Fusion, pp. 1–9. IEEE (2010). https://doi.org/10.1109/ICIF.2010.5711863. http://ieeexplore.ieee.org/document/5711863/
43. Ribeiro, F., Leist, A.: Who is going to pay the price of COVID-19? Reflections about an unequal Brazil. Int. J. Equity Health **19**(1), 91 (2020). https://doi.org/10.1186/s12939-020-01207-2
44. Ribeiro, M.H.D.M., Silva, R.G., Mariani, V.C., Coelho, L.D.S.: Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Solitons Fractals **135**, 109853 (2020). https://doi.org/10.1016/j.chaos.2020.109853
45. de Ridder, D., Tax, D.M.J., Lei, B., Xu, G., Feng, M., Zou, Y., van der Heijden, F.: Parameter Estimation, Chap. 4, pp. 77–113. Wiley, New York (2017). https://doi.org/10.1002/9781119152484.ch4
46. Roda, W.C., Varughese, M.B., Han, D., Li, M.Y.: Why is it difficult to accurately predict the COVID-19 epidemic? Infect. Dis. Model. **5**, 271–281 (2020). https://doi.org/10.1016/j.idm.2020.03.001
47. Salvatier, J., Wiecki, T.V., Fonnesbeck, C.: Probabilistic programming in Python using PyMC3. PeerJ Comput. Sci. **2**, e55 (2016). https://doi.org/10.7717/peerj-cs.55
48. Schulz, E., Speekenbrink, M., Krause, A.: A tutorial on Gaussian Process regression: Modelling, exploring, and exploiting functions. J. Math. Psychol. **85**, 1–16 (2018). https://doi.org/10.1016/j.jmp.2018.03.001
49. Shi, J.Q., Choi, T.: Gaussian Process Regression Analysis for Functional Data, 1st edn. Chapman and Hall/CRC, New York (2011). https://doi.org/10.1201/b11038
50. Singh, S., Parmar, K.S., Makkhan, S.J.S., Kaur, J., Peshoria, S., Kumar, J.: Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. Chaos Solitons Fractals **139**, 110086 (2020). https://doi.org/10.1016/j.chaos.2020.110086
51. Storn, R., Price, K.: Differential Evolution–a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**(4), 341–359 (1997). https://doi.org/10.1023/A:1008202821328
52. Sun, N.Z., Sun, A.: The Classical Inverse Problem, Chap. 2, pp. 25–67. Springer, New York (2015). https://doi.org/10.1007/978-1-4939-2323-6_2
53. Taghizadeh, L., Karimi, A., Heitzinger, C.: Uncertainty quantification in epidemiological models for the COVID-19 pandemic. Comput. Biol. Med. **125**, 104011 (2020). https://doi.org/10.1016/j.compbiomed.2020.104011
54. Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. Society for Industrial and Applied Mathematics, Philadelphia (2005). https://doi.org/10.1137/1.9780898717921
55. Torres, T.S., Hoagland, B., Bezerra, D.R.B., Garner, A., Jalil, E.M., Coelho, L.E., Benedetti, M., Pimenta, C., Grinsztejn, B., Veloso, V.G.: Impact of COVID-19 pandemic on sexual minority populations in Brazil: an analysis of social/racial disparities in maintaining social distancing and a description of sexual behavior. AIDS Behav. (2020). https://doi.org/10.1007/s10461-020-02984-1
56. Veiga e Silva, L., de Andrade Abi Harb, M.D.P., Teixeira Barbosa dos Santos, A.M., de Mattos Teixeira, C.A., Macedo Gomes, V.H., Silva Cardoso, E.H., S da Silva, M., Vijaykumar, N.L., Venâncio Carvalho, S., Ponce de Leon Ferreira de Carvalho, A., Lisboa Frances, C.R.: COVID-19 mortality underreporting in Brazil: analysis of data from government internet portals. J. Med. Internet Res. **22**(8), e21413 (2020). https://doi.org/10.2196/21413
57. Volpatto, D.T., Resende, A.C.M., dos Anjos, L., Silva, J.V.O., Dias, C.M., Almeida, R.C., Malta, S.M.C.: A generalised SEIRD model with implicit social distancing mechanism: a Bayesian approach for the identification of the spread of COVID-19 with applications in Brazil and Rio de Janeiro state. J. Simul. (2021). https://doi.org/10.1080/17477778.2021.1977731
58. Wu, S.L., Mertens, A.N., Crider, Y.S., Nguyen, A., Pokpongkiat, N.N., Djajadi, S., Seth, A., Hsiang, M.S., Colford, J.M., Reingold, A., Arnold, B.F., Hubbard, A., Benjamin-Chung, J.: Substantial underestimation of SARS-CoV-2 infection in the United States. Nat. Commun. **11**(1), 4507 (2020). https://doi.org/10.1038/s41467-020-18272-4
59. Zeng, X., Ghanem, R.: Dynamics identification and forecasting of COVID-19 by switching Kalman filters. Comput. Mech. **66**(5), 1179–1193 (2020). https://doi.org/10.1007/s00466-020-01911-4
60. Zhou, T., Ji, Y.: Semiparametric Bayesian inference for the transmission dynamics of COVID-19 with a state-space model. Contemp. Clin. Trials **97**, 106146 (2020). https://doi.org/10.1016/j.cct.2020.106146