Enhancing the estimation of compartmental model parameters for COVID-19 data with a high level of uncertainty

Gustavo B. Libotte[a,*], Lucas dos Anjos[a], Regina C. Almeida[a], Sandra M. C. Malta[a], Renato S. Silva[a]

[a]*National Laboratory for Scientific Computing, Petrópolis, Brazil*

Abstract

Research on predictions related to the spread of the novel coronavirus are crucial in decision-making to mitigate the disease. Computational simulations are often used as a basis for forecasting the dynamics of epidemics and, for this purpose, compartmental models have been widely used to assess the situation resulting from the spread of the disease in the population. Reliable data is essential to obtain adequate simulations. However, several political, economic, and social factors have caused inconsistencies in the reported data, which are reflected in the capacity for realistic simulations and predictions. Such uncertainties are mainly motivated by a large-scale underreporting of cases due to the reduced testing capacity in some locations. In order to mitigate the effects of noise in the data used to estimate parameters of compartmental models, we propose strategies capable of improving the ability to predict the spread of the disease. We show that the regularization of data by means of Gaussian Process Regression can reduce the variability of successive forecasts, thus 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

1. Introduction

Since the onset of the novel coronavirus (SARS-CoV-2) pandemic, at the end of 2019, a wealth of research have been carried out from across the globe, aiming to understand the dynamics of the disease and transmission patterns. Even after almost a year since the notification of the first case, the number of infected individuals keeps rising significantly. In the meantime, the confirmation of reinfections and identification of seasonal immunity [1] 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 are to predict the way the disease spreads in the population, considering continuously reported data [2]. 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 [3, 4, 5]. The heterogeneity of the population in relation to 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. [6, 7, 8, 9]).

Asymptomatic people also play a significant role in the ongoing pandemic. Oran and Topol [10] 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 is analyzed by Ioannidis [11], Li et al. [12], and Wu et al. [13].

*Corresponding author

Email addresses: glibotte@lncc.br (Gustavo B. Libotte), lanjos@lncc.br (Lucas dos Anjos), rcca@lncc.br (Regina C. Almeida), smcm@lncc.br (Sandra M. C. Malta), rssr@lncc.br (Renato S. Silva)

Turning attention to the impact of the pandemic in Brazil, a country with a continental dimension, such problems tend to become even more evident [14]. Socioeconomic inequalities [15, 16] and cultural factors [17, 18] 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. [19] also analyze presumptive inconsistencies in the data collected and made available by the Ministry of Health in Brazil. They report that, according to the methodology adopted in their analysis, there may be a difference of approximately 41% in the number of deaths caused by complications arising from COVID-19.

In an attempt to better describe the effects of COVID-19 on different population groups, several models have been proposed, aiming to provide more accurate forecasts by means of integrating typical features related to the disease, such as the quarantine period, lockdown, social distancing and hospitalization. Massonis et al. [20] 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 [21] and Roda et al. [3].

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, as shown in Fig. 1. The data do not have a 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 afterwards cause unexpected increases when data are reported in the begining of the following week.

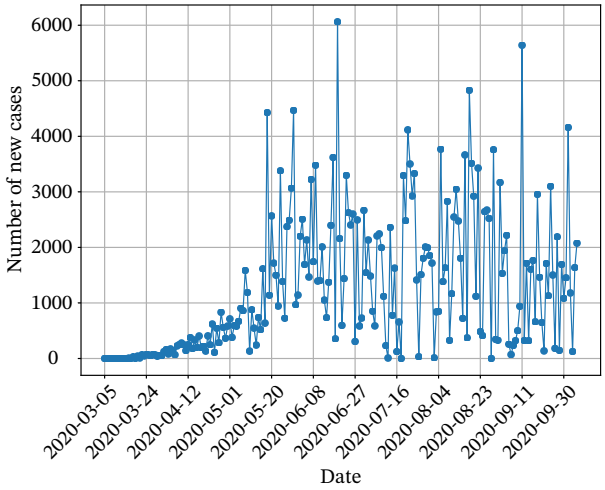


Figure 1: Number of new confirmed cases in the state of Rio de Janeiro by October 6, 2020.

In this context, the main objective of this work is to propose the use of strategies capable of reducing the variability of the estimation of compartmental model parameters, in order to expand the predictive capacity of such models. 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, 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 [22] proposed the Multi-Task GPR model, aiming to predict the COVID-19 outbreak worldwide. The authors compared the results obtained with other regression models, in order to analyze the effectiveness of the proposed method. Zhou and Ji [23] proposed a model for transmission dynamics of COVID-19 considering underreporting of cases (what they called undocumented infections) and estimated the time-varying disease rate of transmission using GPR and a Bayesian approach. Arias Velásquez and Mejía Lara [24] 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. [24] to report a long-term forecast for COVID-19 in the USA. In turn, Ribeiro et al. [25] 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 aware this is the first time that GPR is employed for regularization of data, which are subsequently analyzed using a compartmental model (although other methods for regularization have been applied [12]).

We also 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 the uncertainties related to calibration, as well as aspects such as non-identifiability.

2. Materials and methods

2.1. Compartmental model description

In order to understand the model response to noisy data, we adopt an extension of the compartmental SEIR model, termed SEIRPD-Q model, which encompasses concepts of the models proposed by Jia et al. [26] and Volpatto et al. [27]. Initially, consider a population susceptible to a viral outbreak, whose rate of transmission per contact is given by β. Infected individuals may experience a latent period, 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 $1/\tilde{\sigma}$. We divide this group 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 and, as a consequence of this policy, the diagnosis is made, primarily, in individuals who have severe symptoms or are hospitalized. The proportion 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 are those 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, that are removed at a rate of γ_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. 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. The mathematical formulation of the SEIRPD-Q model is given by Eq. (1), and its schematic description is presented in Fig. 2.

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

The model adopted presents some fundamental differences in relation to those on which we are based: first, Jia et al. [26] 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 due to the uncertainties regarding the reported data, and incorporate the dynamics related to positively diagnosed individuals, of which data are more reliable (as proposed by Volpatto et al. [27]); third, we consider only the mortality rate of positively diagnosed individuals, unlike Volpatto et al. [27]. Additionally, we also compute the cumulative number of infected and dead individuals, which are respectively given by

$$
C(t) = \int_{t_0}^{t} \tilde{\sigma}(1-\rho) E(t) \, \mathrm{d}t \,, \quad D(t) = \int_{t_0}^{t} d_P P(t) \, \mathrm{d}t \,.
$$

These quantities are used in the estimation of the model parameters and the term cumulative number of infected individuals is often used interchangeably with confirmed cases throughout the text.

Given the m-dimensional vector θ of model parameters, model responses at different times t_i, for $i = 1, \ldots, p$, are denoted by $y_i^{(j)} = y^{(j)}(t_i, \theta)$, for $j \in \{C, D\}$. More specifically, we denote $y_i^{(C)} = C(t_i)$ and $y_i^{(D)} = D(t_i)$. Likewise, let $\mathcal{D}_i^{(j)} = \mathcal{D}^{(j)}(t_i)$ denote the observable quantities so that $\boldsymbol{\mathcal{D}} = \left\{\mathcal{D}_1, \ldots, \mathcal{D}_p\right\}$ is a finite set of real-valued measurements collected from successive observations of cumulative number of infected and dead individuals at different times t_i.

2.2. *Deterministic approach for parameter estimation*

In mathematical modeling, the *direct approach* concerns the determination of system states, represented by the output of the model, which describes the relation of the variables taking into account a set of parameters. This approach is also known as *forward problem*. In general, the output of the system is subject to noise that cannot be measured directly and, therefore, the system has unknowns that must be determined. In turn, the so-called *inverse problem*, also

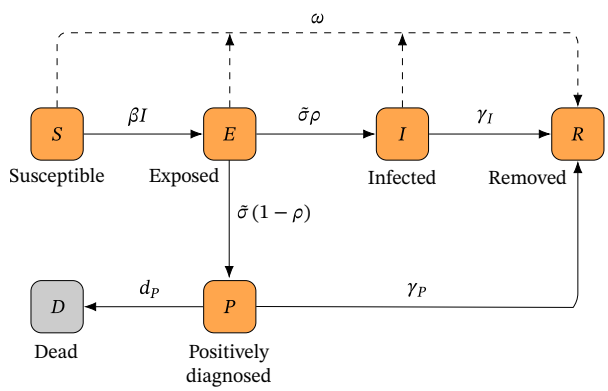


Figure 2: Schematic description of the SEIRPD-Q model.

known as *parameter calibration*, is mainly concerned with determining the model parameters that generate outputs as close as possible to the observable data.

The objective of the inverse problem is, therefore, to find the vector $\hat{\theta}$ (an estimate of θ) 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 $\mathbf{r}(\theta) = \hat{\mathbf{y}} - \mathcal{D}$. The solution to an inverse problem is, in other words, the data fitting whose objective is to calculate an estimate $\hat{\theta}$ that minimizes some error norm $\|\mathbf{r}\|$ of the residuals [28, 29]. The *least squares* fitting calculates the vector of optimal parameters by taking the mean squared error, given by

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

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

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

Equation (2) is usually called the *objective function* (or *cost function*). When $\mathcal{E} \to 0$, the estimate $\hat{\theta}$ generates an output vector $\hat{\mathbf{y}}$ that has a high level of agreement with the observed data $\mathcal{D}$, 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.

2.3. 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 by quantifying uncertainty, 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. [30, 31, 32].

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

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

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

A 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 data we have available, given a set of parameters. In other words, the likelihood function measures how good the data are being explained by the model. In this work we assume a Gaussian likelihood, which has the form

$$p_{\text{like}}(\mathcal{D} \,|\, \theta) = \prod_{j \in \{C,\, D\}} \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(- \sum_{i=1}^{n} \frac{\left(\mathcal{D}^{(j)}(t_i) - y^{(j)}(t_i, \theta)\right)^2}{2\sigma_j^2} \right),$$

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

Bayesian inference is conducted using the probabilistic programming library PyMC3 [33]. We adopt the Transitional Markov Chain Monte Carlo method proposed by Ching and Chen [34] for parameter estimation (which is not described here, for the sake of brevity).

2.4. 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 [35]. 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 [36]. 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 [37].

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

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

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

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

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 $k(t,\ t')$. 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 $\mathcal{D}^*$ is given by

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

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 [35]. Considering the Schur complement (for more details on Schur complements, refer to Puntanen and Styan [38]), the posterior predictive distribution is the multivariate Gaussian distribution

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

with mean

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

and covariance matrix

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

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 ($\mathcal{D}$). By aplying the GPR methodology to the observation set $\mathcal{D}^{(j)}$, for $j \in \{C,\ D\}$, we then obtain $p\left(\mathcal{D}^{(j)*} \,|\, \mathbf{t}^*,\ \mathbf{t},\ \mathcal{D}^{(j)}\right)$, 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 [39]. 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 [40]. 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 [41]. Different functions can be combined to produce kernels

with a variety of features, generally by both adding and multiplying kernels. Rasmussen and Williams [35] 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 following combined kernel:

$$k(t,\ t') = k_{\mathrm{c}}(t,\ t') \times k_{\mathrm{m}}(t,\ t') + k_{\mathrm{d}}(t,\ t')\ .$$

Its main component belongs to the *Matérn* class of covariance functions and is given by

$$k_{\mathrm{m}}(t,\ t') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{\ell} \right)\ ,$$

where $r = |t - t'|$, ν and ℓ are positive constants, and $\Gamma(\nu)$ and K_ν are the gamma and a modified Bessel functions, respectively. The two other components are the dot product and the constant kernels. The former is given by $k_{\mathrm{d}}(t,\ t') = \sigma_{\mathrm{d}}^2 + t \cdot t'$, and the latter is defined as $k_{\mathrm{c}}(t,\ t') = c$, where c is a real-valued constant. The hyperparameters ν controls the smoothness of the function, ℓ is the characteristic length scale, and σ_{d}^2 and c are multiplicative and additive scale factors, respectively.

Rasmussen and Williams [35] present an algorithm for GPR employing Cholesky factorization to solve the matrix inversion required by Eqs. (6) and (7). We use the Scikit-learn library [42] to implement the GPR. Hyperparameters are calculated using a computational routine internal to the library, which uses the L-BFGS-B algorithm [43] 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.

2.5. Data

The data used in this work are the cumulative number of infected individuals and the number of deaths caused by COVID-19 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. [44]. 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 this data to estimate the parameters of model since, due to the unseemly policy of testing the population, there is much uncertainty about these data, given that the number of infected individuals is not known with relative accuracy. In addition, we adopt the cumulative number of infected individuals, rather than the number of new daily cases, as an attempt to mitigate the effects of noise caused by inconsistent data (as can be seen in Fig. 1).

2.5.1. Parameter and numerical experiment setups

In models with more compartments and which, in general, have more parameters, finding a unique set of parameter 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 [45]. 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. The remaining parameters of the model are assumed to be known biological parameters. In this case, the incubation period is $\tilde{\sigma} = 1/5.8$ day^{-1} [46], the proportion of symptomatic infected individuals is $\rho = 0.6$ [10], and both the recovery rate of infected and positively diagnosed individuals are defined as $\gamma_I = \gamma_P = 1/16.7$ day^{-1} [47].

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 $N = 17264943$ individuals, according to the last demographic census conducted by the Brazilian Institute of Geography and Statistics [48]. 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 beginning of the epidemic. 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 model factor to be estimated and the initial condition for the number of susceptible individuals is given by $S(0) = N - (E(0) + I(0) + P(0) + R(0) + D(0))$. In addition, the initial condition for the number of confirmed cases is taken as $C(0) = I(0)$.

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 a real scenario. Furthermore, we only consider short-term predictions in our analyzes since, as the simulations are compared with original data. Therefore, we analyze two scenarios in which we adopt 7 and 14 data points in the test set.

Arbitrarily, we choose the minimum size of the training data set equal to 60. In our first analysis, we set up 7 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, a new data is added to the training set, so that the test set is composed of the next 7 values in the time series. As data for 210 days are available, 143 sets of parameters are estimated, using the deterministic approach described in Section 2.2. 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 root-mean-square error [49], considering both the number of confirmed cases and dead individuals.

All optimal parameters are calculated by combining the Differential Evolution [50] and the Nelder-Mead Simplex [51] 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, respecitively. 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 $0 \leq \beta \leq 10^{-6}$, $0 \leq \omega \leq 0.025$, $0 \leq d_P \leq 0.025$, and $0 \leq E(0) \leq 10^4$ in appropriate units (a. u.).

3. Results and Discussion

3.1. Influence of regularization of data

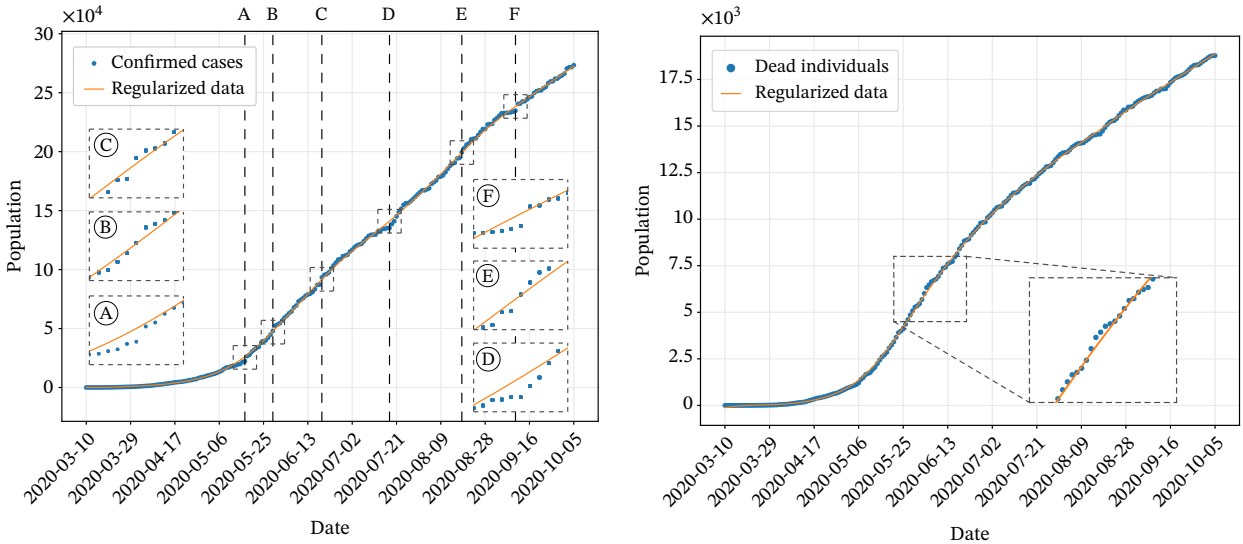
Initially, the data used in this analysis is regularized using GPR as described in Section 2.4. The values of the hyperparameters tunned for the used kernel are indicated in Table 1. Figure 3 shows an illustrative comparison between the original data for confirmed cases and dead individuals and the corresponding regularized data. In Fig. 3a, specific data are highlighted by the dashed vertical lines, identified by the letters A to F. Each of these lines has a dashed bounding box that represents the range of the zoom box with the corresponding letter. The behavior of the data in the highlighted points are discussed later. One may notice that regularized data show good agreement with the reported data, preserving the behavior of the curves over the entire interval, but smoothing out the noise—this statement is clear when looking at the zoom boxes in both figures.

Table 1: Tunned hyperparameters of the kernel used to regularize data of cumulative number of infected and dead individuals.

Data	k_c	k_m		k_d
	c	ν	ℓ	σ_d^2
Confirmed cases	313.29	1.5	132	2.09×10^{-5}
Dead individuals	222.01	1.5	104	7.44×10^{-5}

Next, the influence of estimating the parameters of the SEIRPD-Q model, presented in Section 2.1, using the regularized data set, in relation to the approach that adopts the original data is analyzed. Our focus is to analyze 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. The analysis is performed by evaluating the root-mean-squared error (RMSE) between model predictions and test data.

The results concerning the optimal parameters calculated by the Differential Evolution and the Nelder-Mead Simplex (Section 2.5.1) are shown in Fig. 4. The values referring to the analysis presented here are represented by the



(a) Confirmed cases.

(b) Dead individuals.

Figure 3: Results of data regularization using GPR. The blue dots are the original data and the orange curve represents the regularized data.

black (original data) and blue (regularized data) dots. The other results (orange dots) will be discussed below, but were depicted in the same axes to favor visual inspection and comparative analysis.

Analyzing Fig. 4, 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 72.03% of runs. On average, the RMSE obtained with original data is 24465.59, whereas for regularized data it is 22023.72. This represents an average improvement of approximately 9.98%.

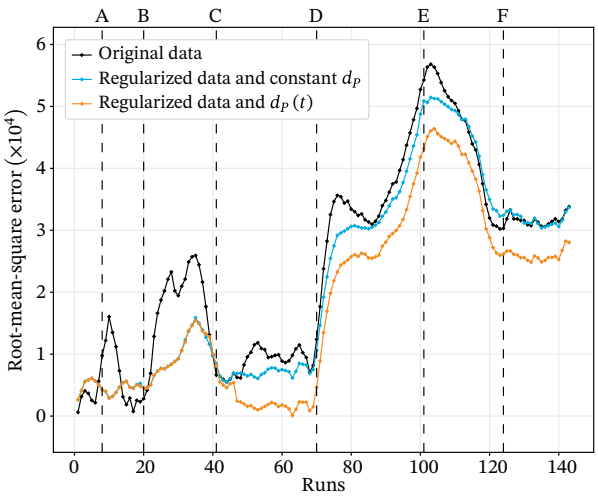
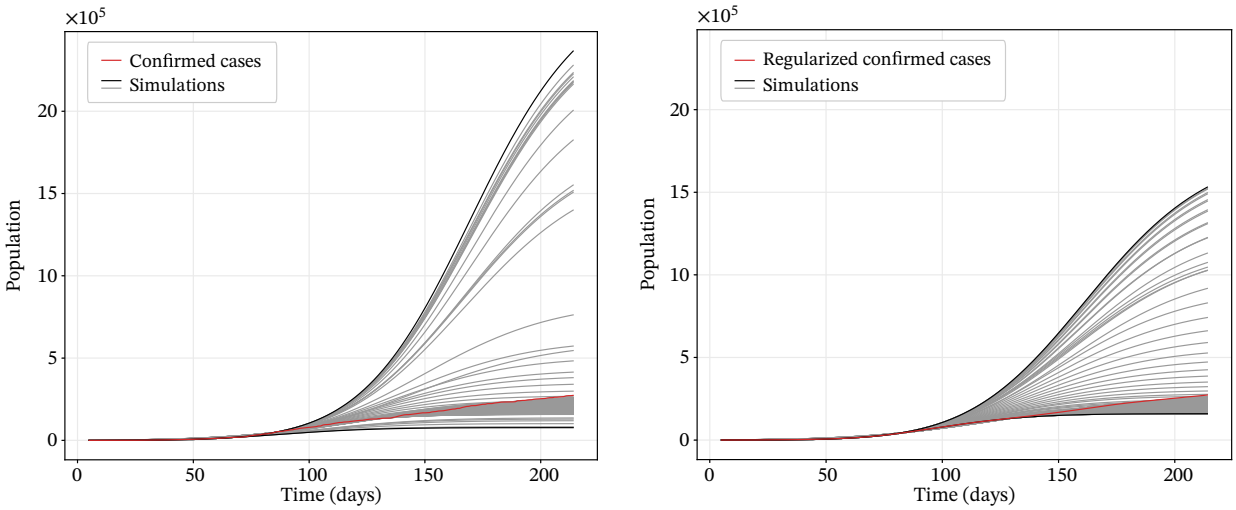


Figure 4: RMSE computed by comparing simulations of the SEIRPD-Q model, performed using parameters that best fit the training data, in relation to the test set (composed of 7 points). Each point represents the deviation relative to the test data for a given run, where the number of training data varies. The black dots show the results corresponding to the calibrations using original data; the blue dots refer to simulations with calibrated parameters using regularized data and constant mortality rate; and the orange dots represent the results for regularized training data set and time-varying d_P.

Consider now the RMSEs marked by the dashed vertical lines, identified in Fig. 4 by the letters A to F. These values are related to the corresponding letters in Fig. 3a, indicating the RMSE obtained when a certain amount of training data is used to estimate the parameters. As the test data is never regularized, it is possible to associate the variations of RMSEs with the effects of the noise on the original data, as can be seen in the zoom boxes of Fig. 3a. This relationship is more relevant in the data of confirmed cases because, comparing the order of magnitude of these data with that of dead individuals, the former has more influence on the parameter estimation.

The implication of using regularized data is even more straightforward when we analyze the variability of simulations resulting from the set of parameters corresponding to each point in Fig. 4. For this purpose, consider the results shown in Fig. 5, which presents the simulations of confirmed cases, whose parameters are estimated by varying the amount of training data, as previously described. The simulations are represented by the gray curves (those that resulted in the extreme values are shown in black), and the training data set is shown in red. It is worth mentioning that the set of training data, even presenting variable size for each parameter estimation, is completely shown in both cases, in order to favor the comparison with the simulated results. Overall, 143 simulations are shown in each of the Figs. 5a and 5b, which cannot be distinguished from each other due to the proximity of the curves and the scale of the results.



(a) Original data.

(b) Regularized data.

Figure 5: Simulations of confirmed cases in Rio de Janeiro with parameters estimated using the deterministic approach. For each calibration, a data point is added to the set of used data. The gray curves represent the resulting simulations, and the black curves are the extreme results. The corresponding data is illustrated by the red curve.

Comparing Figs. 5a and 5b, 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 (last day of the simulation), Rio de Janeiro had 273338 confirmed cases, the extreme values (corresponding to the final values of the black curves) obtained in Fig. 5a are 77121 and 2365715, whereas in Fig. 5b the extreme values are in a narrower range, between 159454 and 1531221. The dispersion of these results can be alternatively analyzed using a box-and-whisker diagram, as shown in Fig. 6. The boxes are bounded by the lower (first) and upper (third) quartiles of the full model prediction values, and the orange line 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. In turn, the blue dashed line represents the experimental observed value of the cumulative number of infected individuals at the corresponding time.

Using original data, the median of the simulated values is 203877, whereas for regularized data the median is 218528; the interquartile ranges (the distance between the upper and lower quartiles) in both cases are [167865, 252711] and [172209, 258302], respectively; whiskers range from the lowest to the highest number of confirmed cases, both

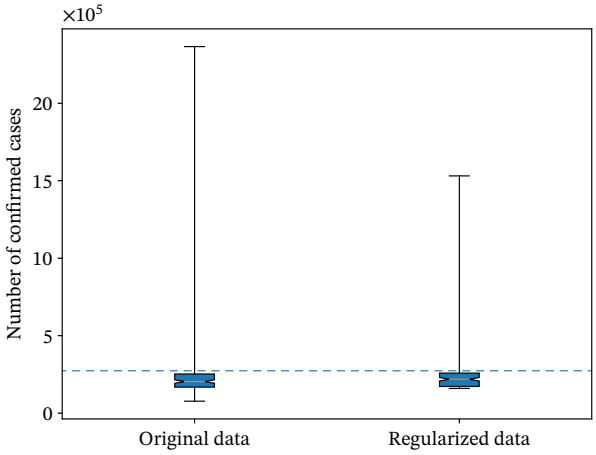


Figure 6: Box-and-whisker diagram of the simulated values of the number of confirmed cases obtained with the model calibrated using the original and regularized data. The blue dashed line stands for the true value of the cumulative number of infected individuals at the same time.

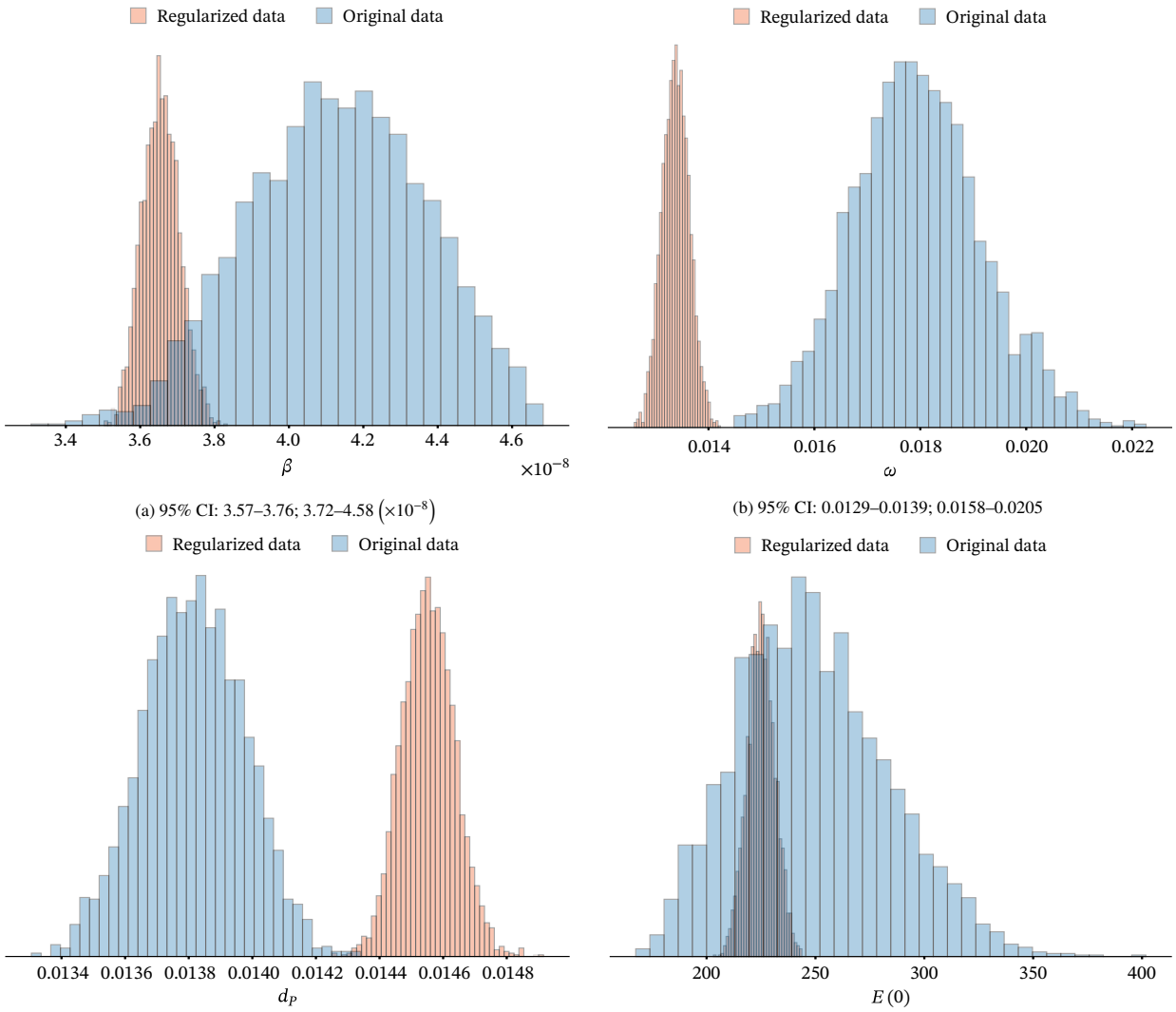
for original and regularized data, since no simulation can be considered an outlier. All values are shown as integers, given the nature of the data. It is clear that regularization using GPR preserves the behavior of the data, although substantially reducing the variability of the simulations.

Now consider a forecast of the epidemic in terms of the number of confirmed cases 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 Section 2.3. The training data sets for confirmed cases and dead individuals, both original and regularized, have 70 values. This choice is arbitrary and motivated by the result obtained in Fig. 4, regarding run 10 (which presents a significant difference between the corresponding RMSEs). In turn, the test set is made up of the next 7 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.

For the sake of fairness, the parameters to be estimated are the same as in the previous analysis, and are assumed to follow uniform probability density distributions. Table 2 shows the range of prior distributions for each parameter, as well as the respective *maximum a posterior* (MAP) estimates. Likewise, all other parameters of the SEIRPD-Q model are taken as biological parameters and have the same values previously reported. Figure 7 shows the posterior distributions of the estimated parameters of the SEIRPD-Q model, obtained when original (blue bins) and regularized (orange bins) data are employed for calibration. All of them have nice bell shapes. For each parameter, the corresponding 95% CI is also shown, where the first reading corresponds to the regularized data and the second reading

Table 2: Prior distributions and MAP values of the parameters estimated using Bayesian calibration (in a. u.).

	Prior distribution	Data type	
		Original	Regularized
		MAP	
β	$\mathcal{U}\left(0,\ 10^{-6}\right)$	4.0503×10^{-8}	3.6593×10^{-8}
ω	$\mathcal{U}\,(0,\ 0.025)$	0.017528	0.013398
d_P	$\mathcal{U}\,(0,\ 0.025)$	0.013804	0.014557
$E\,(0)$	$\mathcal{U}\left(0,\ 10^{4}\right)$	259.8910	225.3423



(a) 95% CI: 3.57–3.76; 3.72–4.58 $(\times 10^{-8})$

(b) 95% CI: 0.0129–0.0139; 0.0158–0.0205

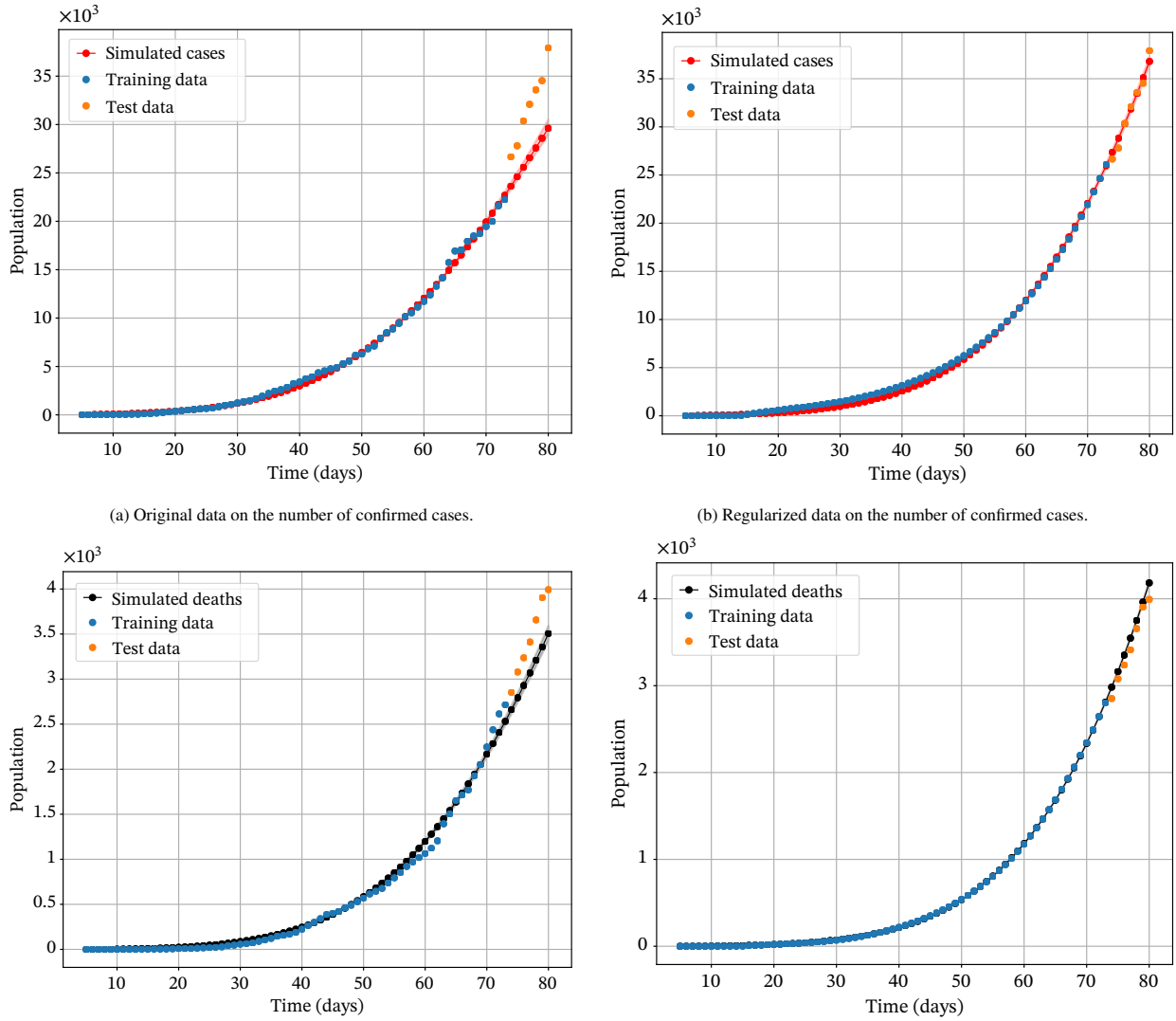
(c) 95% CI: 0.0144–0.0147; 0.0135–0.0141

(d) 95% CI: 213.63–237.37; 190.66–317.14

Figure 7: Posterior distributions of the SEIRPD-Q model parameters for calibrations using original and regularized data. The credible intervals for each parameter are shown in the corresponding frame, where the first reading refers to the regularized data and the second reading refers to the original data.

corresponds to the original data. Of note, the former is much narrower than the latter. This fact is reflected in the stochastic simulation of the SEIRPD-Q model shown in Fig. 8, along with the training data sets (original and regularized) and the test data set (note that the latter is never regularized). The shading around the curves represents the 95% credible interval (CI) [32].

Figures 8a and 8c show that with a relatively small set of training data compared to the amount of data available, the model calibrated with the original data faces difficulties in predicting the short-term behavior of the test data. The noise of the data, in this case, seems to represent a handicap in the predictive capacity of the model, causing the deviation in the short-term prediction to be reflected in unrealistic values in the long-term scenario. On the other hand, 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, as can be seen in Figs. 8b and 8d. 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. Figure 9 shows how these



(a) Original data on the number of confirmed cases.

(b) Regularized data on the number of confirmed cases.

(c) Original data on the number of dead individuals.

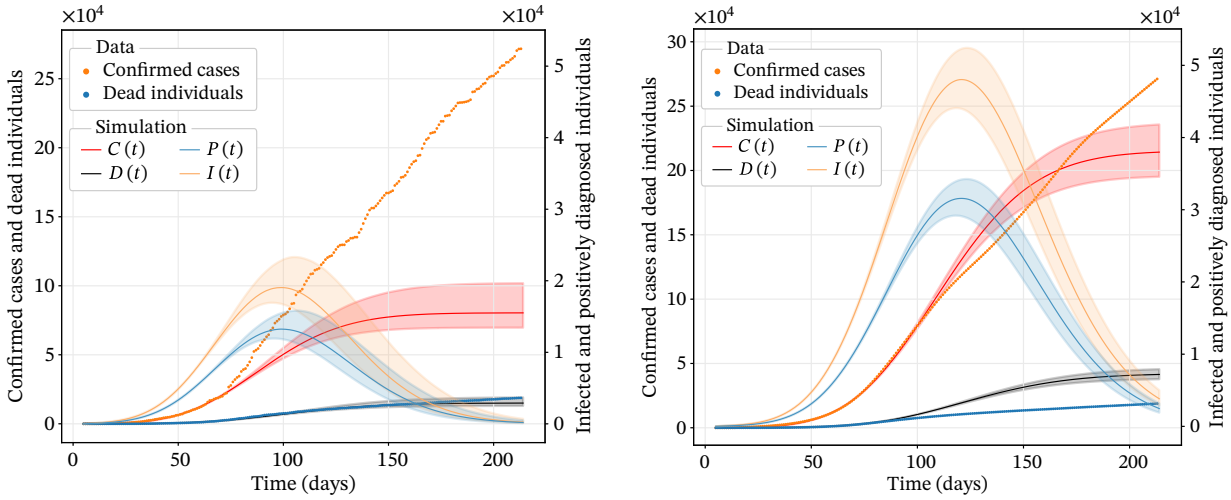
(d) Regularized data on the number of dead individuals.

Figure 8: Forecast of the number of confirmed cases and dead individuals using the Bayesian approach for original and regularized data with a constant d_P. All training sets have 70 values and test is analyzed for the next 7 days.

uncertainties impact long-term predictions. This figure was obtained by runing the stochastic SEIRPD-Q model with the same parameters used in Fig. 8 until day 210. The short-term deviation from the test data shown in Figs. 8a and 8c is amplified in the long-term simulations for the original data, as shown in Fig. 9a. Although there is still a significant deviation in the simulations shown in Fig. 9b, related to the regularized data, since only 70 values were used in each training set to estimate the parameters, the results are more consistent with the behavior of the data over the long run. In addition, the use of regularized data also leads to a delay in the occurrence of the peak of the curve of infected individuals in approximately 20 days.

3.2. Influence of time-varying parameters

A further feature that tends to improve the predictive capacity of compartmental models is to adopt time-varying parameters. In an effort to analyze this scenario, consider the 143 sets of optimal parameters, obtained using the deterministic approach that led to the results shown in Fig. 4. In Fig. 10 we show the values of the calibrated parameters



(a) Original data.

(b) Regularized data.

Figure 9: Long-term simulations of the stochastic SEIRPD-Q model, compared to the number of confirmed cases and dead individuals. Readings for the number of confirmed cases and dead individuals are on the left, and readings for infected and positively diagnosed individuals are on the right.

for each corresponding run, both using original and regularized data. In fact, the behavior of the optimal parameters obtained with original and regularized data has some similarity. However, the influence of noise on the results referring to the original data is evident, in comparison with those obtained with regularized data. Such influence is more effective on the optimal parameters obtained in the first runs for which the training data set is relatively smaller.

The optimal parameters in Fig. 10 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). This behavior may be occuring because quarantine measures tend to become more restrictive when the number of infections increases. 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.

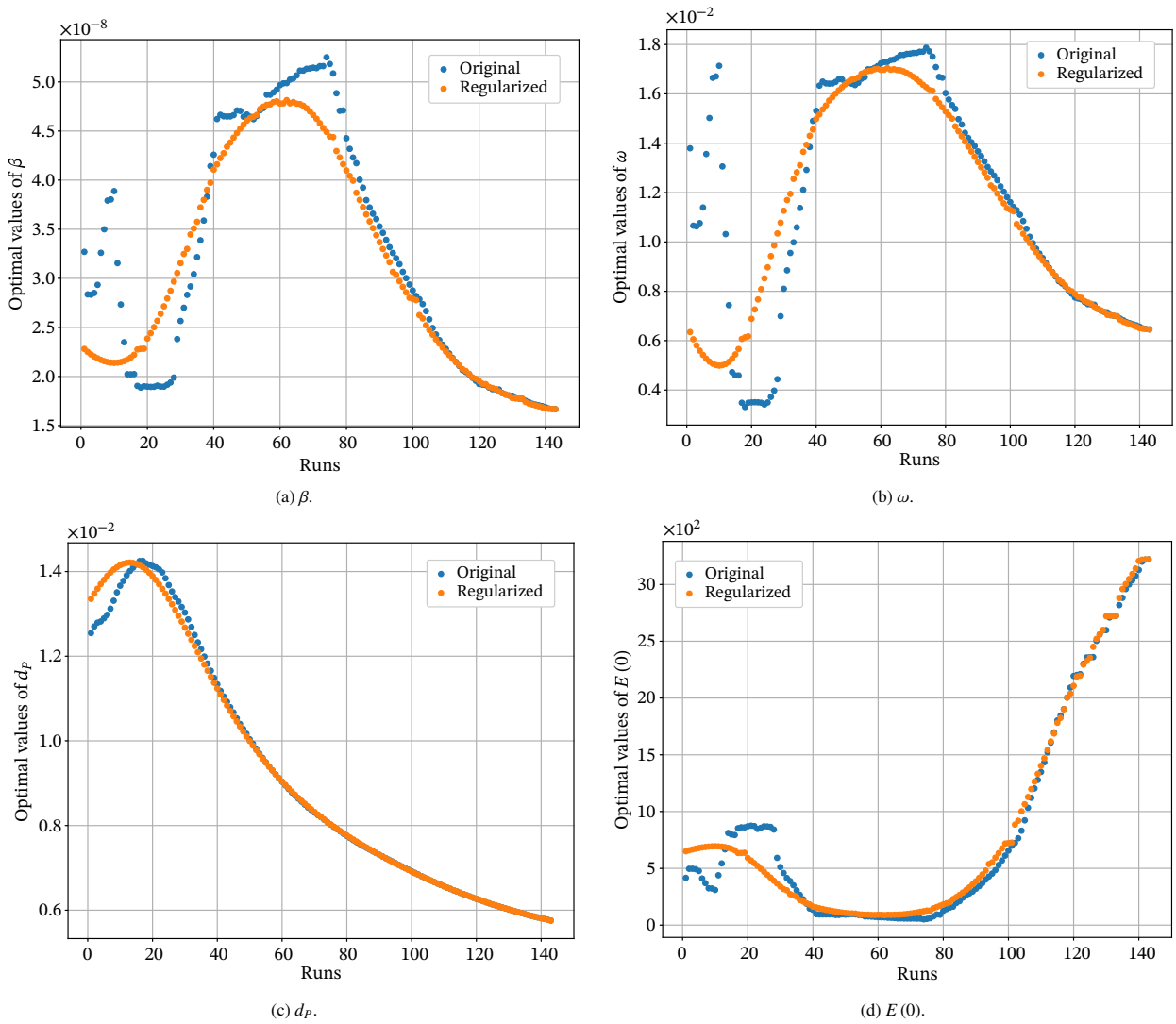
The inspection of Fig. 10c indicates that the mortality rate of positively diagnosed individuals basically only decreases after a certain run. This behavior suggests that the parameter can be approximated by a function. In this case, we propose to represent d_P as a function of the form

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

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 $\theta = (\beta,\ \omega,\ d_0,\ d_1,\ E(0))$.

Following similar reasoning, one could argue that it would be possible to improve the descriptions of the other parameters. We chose to model only d_P because it is the only parameter whose behavior showed a single inflection point, that can be approximated by a simple function in the analyzed interval. The description of the other parameters are likely to require higher order functions which would imply in estimating a greater number of parameters and even leading to non-identifiability issues.

Consider the methodology previously adopted, where the set of training data is gradually expanded with each parameter estimation, but now taking into account the proposed approximation for $d_P(t)$ defined in Eq. (8). The search intervals for the new parameters d_0 and d_1 are $0 \leq d_0 \leq 0.5$ and $0 \leq d_1 \leq 0.5$ (all other parameters follow the same specifications defined before). Figure 4 shows that this strategy is reflected in the reduction of the RMSE



(a) β.

(b) ω.

(c) d_P.

(d) $E(0)$.

Figure 10: Estimated parameters in each run of the deterministic approach using original and regularized data.

considering the test data set with 7 values, in most simulated forecasts. The RMSEs considering regularized data and time-varying d_P (orange dots) represent a better approximation of the test data set in 88.81% of the analyzed runs, in relation to the results considering original data (black dots). On average, the root-mean-square error obtained with regularized data and $d_P(t)$ is 18186.52, which represents a reduction of approximately 25.66% in relation to the results with original data (which is 24465.59, on average, as previously reported).

It is relevant to assess the influence of the proposed approaches for a larger test set. For that matter, consider the same methodology adopted to obtain the results of Fig. 4, but now with the test set composed of the next 14 data following the last data of the training set. Thus, 136 sets of optimal parameters are calculated, whose respective simulations are compared to the test set of the corresponding run, and the deviation is computed by the RMSE. Figure 11 shows the results for original and regularized data, also considering constant and time-varying d_P.

Comparing the results in Fig. 11, the RMSE calculated with regularized data and constant d_P (blue dots) is lower than the deviation computed using original data (black dots) in 72.06% of the tests, whereas for time-varying d_P (orange dots) this value rises to 89.71%. On average, the root-mean-square error obtained with original data is 45526.23. In the case of using regularized data, the error is 39794.67 when adopting a constant d_P, and 34300.77

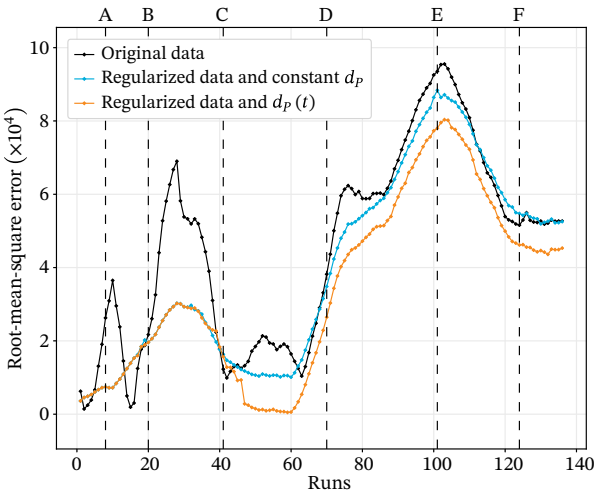


Figure 11: RMSE computed by comparing simulations of the SEIRPD-Q model, performed using parameters that best fit the training data, in relation to the test set (composed of 14 points). Each point represents the deviation relative to the test data for a given run, where the number of training data varies. The black dots show the results corresponding to the calibrations using original data; the blue dots refer to simulations with calibrated parameters using regularized data and constant mortality rate; and the orange dots represent the results for regularized training data set and time-varying d_P.

when d_P varies over time. Such values represent an average gain of approximately 12.59% for the former case, and this value goes to 24.66% for the latter case. The values marked by the dashed vertical lines, labeled A to F, are analogous to those in Fig. 4 and demonstrate that the behavior of the RMSE is fairly affected by the noise of the test set, even when the set is larger. Table 3 summarizes the results of Figs. 4 and 11.

Table 3: Average RMSE calculated with original and regularized data. For the latter, the values obtained with constant and variable d_P are also considered. In all cases, the simulations are compared with test data containing 7 and 14 records.

Test	Data type		
	Original	Regularized	
		Constant d_P	Time-varying d_P
7 days	24465.59	22023.72	18186.52
14 days	45526.23	39794.67	34300.77

As in Fig. 10, we are interested in understanding the behavior of the parameters inherent to the function when d_P varies over time. Figure 12 shows the parameters d_0 and d_1, calculated in each parameter estimation procedure related to the orange points in Fig. 4. In addition, Fig. 13 shows the behavior of Eq. (8), taking the corresponding values of d_0 and d_1 in Fig. 12, as well as the values of d_P shown in Fig. 10c (obtained with regularized data). Therefore, 143 curves in different shades of red are shown. The colors correspond to the amount of training data used in the calibration, varying according to the color map shown next to the axes. 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. Note that in Fig. 13, the curves cannot be distinguished from each other due to their proximity. In addition, it is worth mentioning that the other parameters are not shown for this case as they are similar to those in Fig. 10.

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 Fig. 12b). In this case, $d_P(t) \approx d_0$, the darkest curves in Fig. 13. These curves practically coincide with the first points shown in Fig. 13, which means that, in fact, the calibrations are quite similar, and this hypothesis is confirmed when we compare the first values of the RMSE in Fig. 4 (blue and orange dots), which are practically overlapping. This behavior occurs because, given the training set for these runs, the model identifies that the mortality rate is not

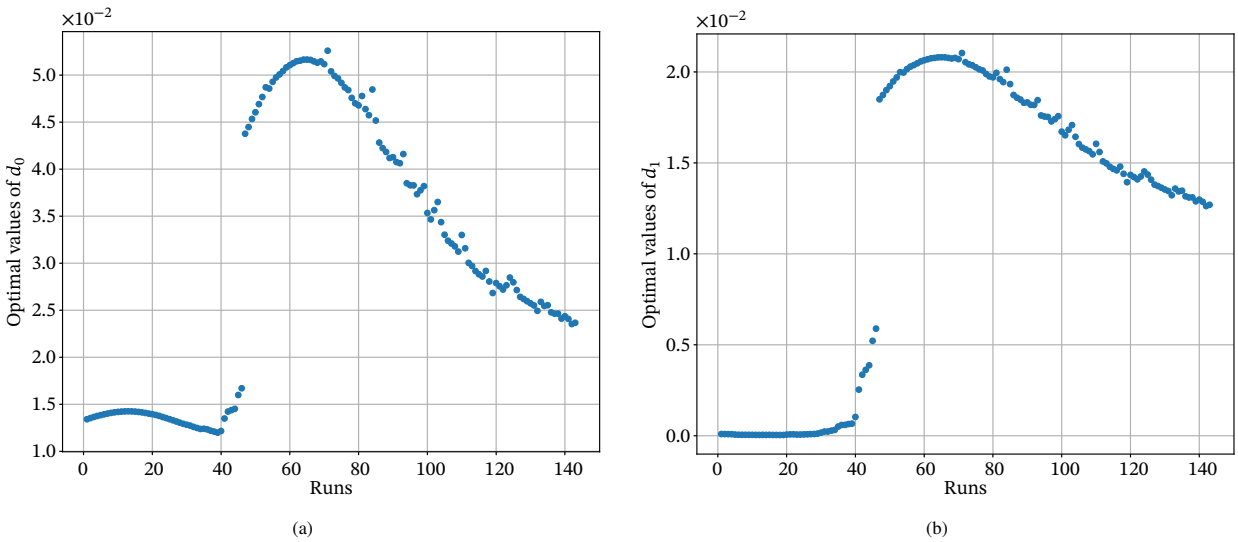


Figure 12: Parameters of the function for time-varying mortality rate estimated in each run of the deterministic approach.

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 scenario, the first calibrations provide functions relatively distant from the corresponding points in Fig. 13, 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. 4, where the orange dots always represent the best approximation in relation to the test data.

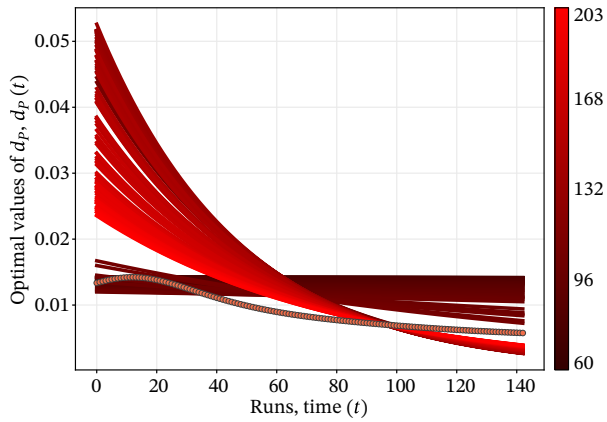


Figure 13: Curves of Eq. (8) considering each pair of optimal parameters d_0 and d_1 in Fig. 12, in comparison to the results for constant d_P (represented by the points) in Fig. 10c. The inserted color map associates the color of a curve with to the amount of data present in the training set used for calibration.

4. Conclusions

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 towards 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 can represent an alternative to mitigate the effect of noise in a given parameter calibration. 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. 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.

Declaration of Competing Interest

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

Code availability

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

Acknowledgments

The authors would like to thank the Ministry of Science, Technology, Innovation and Communication (MCTIC) of Brazil. Gustavo Libotte is 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.

References

[1] A. W. D. Edridge, J. Kaczorowska, A. C. R. Hoste, M. Bakker, M. Klein, K. Loens, M. F. Jebbink, A. Matser, C. M. Kinsella, P. Rueda, M. Ieven, H. Goossens, M. Prins, P. Sastre, M. Deijs, L. van der Hoek, Seasonal coronavirus protective immunity is short-lasting, Nature Medicine (2020). doi:10.1038/s41591-020-1083-1.

[2] C. S. Currie, J. W. Fowler, K. Kotiadis, T. Monks, B. S. Onggo, D. A. Robertson, A. A. Tako, How simulation modelling can help reduce the impact of COVID-19, Journal of Simulation 14 (2) (2020) 83–97. doi:10.1080/17477778.2020.1751570.

[3] W. C. Roda, M. B. Varughese, D. Han, M. Y. Li, Why is it difficult to accurately predict the COVID-19 epidemic?, Infectious Disease Modelling 5 (2020) 271–281. doi:10.1016/j.idm.2020.03.001.

[4] I. Holmdahl, C. Buckee, Wrong but useful—what COVID-19 epidemiologic models can and cannot tell us, New England Journal of Medicine 383 (4) (2020) 303–305. doi:10.1056/NEJMp2016822.

[5] C. E. Overton, H. B. Stage, S. Ahmad, J. Curran-Sebastian, P. Dark, R. Das, E. Fearon, T. Felton, M. Fyles, N. Gent, I. Hall, T. House, H. Lewkowicz, X. Pang, L. Pellis, R. Sawko, A. Ustianowski, B. Vekaria, L. Webb, Using statistics and mathematical modelling to understand infectious disease outbreaks: COVID-19 as an example, Infectious Disease Modelling 5 (2020) 409–441. doi:10.1016/j.idm.2020.06.008.

[6] S. S. Bhopal, R. Bhopal, Sex differential in COVID-19 mortality varies markedly by age, The Lancet 396 (10250) (2020) 532–533. doi:10.1016/S0140-6736(20)31748-7.

[7] D. Calvetti, A. P. Hoover, J. Rose, E. Somersalo, Metapopulation network models for understanding, predicting, and managing the coronavirus disease COVID-19, Frontiers in Physics 8 (2020). doi:10.3389/fphy.2020.00261.

[8] A. Clark, M. Jit, C. Warren-Gash, B. Guthrie, H. H. X. Wang, S. W. Mercer, C. Sanderson, M. McKee, C. Troeger, K. L. Ong, F. Checchi, P. Perel, S. Joseph, H. P. Gibbs, A. Banerjee, R. M. Eggo, E. S. Nightingale, K. O'Reilly, T. Jombart, W. J. Edmunds, A. Rosello, F. Y. Sun, K. E. Atkins, N. I. Bosse, S. Clifford, T. W. Russell, A. K. Deol, Y. Liu, S. R. Procter, Q. J. Leclerc, G. Medley, G. Knight, J. D. Munday, A. J. Kucharski, C. A. B. Pearson, P. Klepac, K. Prem, R. M. G. J. Houben, A. Endo, S. Flasche, N. G. Davies, C. Diamond, K. van Zandvoort, S. Funk, M. Auzenbergs, E. M. Rees, D. C. Tully, J. C. Emery, B. J. Quilty, S. Abbott, C. J. Villabona-Arenas, S. Hué, J. Hellewell, A. Gimma, C. I. Jarvis, 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, The Lancet Global Health 8 (8) (2020) e1003–e1017. doi:10.1016/S2214-109X(20)30264-3.

[9] N. G. Davies, P. Klepac, Y. Liu, K. Prem, M. Jit, R. M. Eggo, Age-dependent effects in the transmission and control of COVID-19 epidemics, Nature Medicine 26 (8) (2020) 1205–1211. doi:10.1038/s41591-020-0962-9.

[10] D. P. Oran, E. J. Topol, Prevalence of asymptomatic SARS-CoV-2 Infection, Annals of Internal Medicine 173 (5) (2020) 362–367. doi:10.7326/M20-3012.

[11] J. P. A. Ioannidis, Coronavirus disease 2019: The harms of exaggerated information and non-evidence-based measures, European Journal of Clinical Investigation 50 (4) (2020). doi:10.1111/eci.13222.

[12] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), Science 368 (6490) (2020) 489–493. doi:10.1126/science.abb3221.

[13] S. L. Wu, A. N. Mertens, Y. S. Crider, A. Nguyen, N. N. Pokpongkiat, S. Djajadi, A. Seth, M. S. Hsiang, J. M. Colford, A. Reingold, B. F. Arnold, A. Hubbard, J. Benjamin-Chung, Substantial underestimation of SARS-CoV-2 infection in the United States, Nature Communications 11 (1) (2020) 4507. doi:10.1038/s41467-020-18272-4.

[14] F. A. L. Marson, COVID-19 – 6 million cases worldwide and an overview of the diagnosis in Brazil: a tragedy to be announced, Diagnostic Microbiology and Infectious Disease 98 (2) (2020) 115113. doi:10.1016/j.diagmicrobio.2020.115113.

[15] F. Ribeiro, A. Leist, Who is going to pay the price of COVID-19? Reflections about an unequal Brazil, International Journal for Equity in Health 19 (1) (2020) 91. doi:10.1186/s12939-020-01207-2.

[16] T. S. Torres, B. Hoagland, D. R. B. Bezerra, A. Garner, E. M. Jalil, L. E. Coelho, M. Benedetti, C. Pimenta, B. Grinsztejn, V. G. Veloso, 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 and Behavior (2020). doi:10.1007/s10461-020-02984-1.

[17] G. A. Cupertino, M. d. C. Cupertino, A. P. Gomes, L. M. Braga, R. Siqueira-Batista, COVID-19 and brazilian indigenous populations, The American Journal of Tropical Medicine and Hygiene 103 (2) (2020) 609–612. doi:10.4269/ajtmh.20-0563.

[18] M. Polidoro, F. de Assis Mendonça, S. N. Meneghel, A. Alves-Brito, M. Gonçalves, F. Bairros, D. Canavese, Territories under siege: risks of the decimation of indigenous and quilombolas peoples in the context of COVID-19 in south Brazil, Journal of Racial and Ethnic Health Disparities (2020). doi:10.1007/s40615-020-00868-7.

[19] L. Veiga e Silva, M. D. P. de Andrade Abi Harb, A. M. Teixeira Barbosa dos Santos, C. A. de Mattos Teixeira, V. H. Macedo Gomes, E. H. Silva Cardoso, M. S da Silva, N. L. Vijaykumar, S. Venâncio Carvalho, A. Ponce de Leon Ferreira de Carvalho, C. R. Lisboa Frances, COVID-19 mortality underreporting in Brazil: analysis of data from government internet portals, Journal of Medical Internet Research 22 (8) (2020) e21413. doi:10.2196/21413.

[20] G. Massonis, J. R. Banga, A. F. Villaverde, Structural identifiability and observability of compartmental models of the COVID-19 pandemic (2020). arXiv:2006.14295.

[21] T. Alberti, D. Faranda, On the uncertainty of real-time predictions of epidemic growths: a COVID-19 case study for China and Italy, Communications in Nonlinear Science and Numerical Simulation 90 (2020) 105372. doi:10.1016/j.cnsns.2020.105372.

[22] S. Ketu, P. K. Mishra, Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection, Applied Intelligence (2020). doi:10.1007/s10489-020-01889-9.

[23] T. Zhou, Y. Ji, Semiparametric Bayesian inference for the transmission dynamics of COVID-19 with a state-space model, Contemporary Clinical Trials 97 (2020) 106146. doi:10.1016/j.cct.2020.106146.

[24] R. M. Arias Velásquez, J. V. Mejía Lara, Gaussian approach for probability and correlation between the number of COVID-19 cases and the air pollution in Lima, Urban Climate 33 (2020) 100664. doi:10.1016/j.uclim.2020.100664.

[25] M. H. D. M. Ribeiro, R. G. Silva, V. C. Mariani, L. d. S. Coelho, Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil, Chaos, Solitons & Fractals 135 (2020) 109853. doi:10.1016/j.chaos.2020.109853.

[26] J. Jia, J. Ding, S. Liu, G. Liao, J. Lin, B. Duan, G. Wang, R. Zhang, Modeling the control of COVID-19: impact of policy interventions and meteorological factors, Electronic Journal of Differential Equations 2020 (23) (2020) 1–24.

[27] D. T. Volpatto, A. C. M. Resende, L. Anjos, J. V. O. Silva, C. M. Dias, R. C. Almeida, S. M. C. Malta, A generalized 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, medRxiv (2020). doi:10.1101/2020.05.30.20117283.

[28] N.-Z. Sun, A. Sun, The Classical Inverse Problem, Springer New York, New York, 2015, Ch. 2, pp. 25–67. doi:10.1007/978-1-4939-2323-6_2.

[29] D. de Ridder, D. M. J. Tax, B. Lei, G. Xu, M. Feng, Y. Zou, F. van der Heijden, Parameter Estimation, John Wiley & Sons, Ltd, 2017, Ch. 4, pp. 77–113. doi:10.1002/9781119152484.ch4.

[30] A. Tarantola, Inverse Problem Theory and Methods for Model Parameter Estimation, Society for Industrial and Applied Mathematics, Philadelphia, 2005. doi:10.1137/1.9780898717921.

[31] W. A. Link, R. J. Barker, Bayesian Inference, Academic Press, London, 2010. doi:10.1016/B978-0-12-374854-6.00004-1.

[32] C. A. L. Bailer-Jones, Practical Bayesian Inference, Cambridge University Press, Cambridge, 2017. doi:10.1017/9781108123891.

[33] J. Salvatier, T. V. Wiecki, C. Fonnesbeck, Probabilistic programming in Python using PyMC3, PeerJ Computer Science 2 (2016) e55. arXiv:1507.08050, doi:10.7717/peerj-cs.55.

[34] J. Ching, Y.-C. Chen, Transitional Markov Chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, Journal of Engineering Mechanics 133 (7) (2007) 816–832. doi:10.1061/(ASCE)0733-9399(2007)133:7(816).

[35] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning, MIT Press,

Cambridge, 2006.
[36] K. P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, Cambridge, 2012.
[37] J. Q. Shi, T. Choi, Gaussian Process Regression Analysis for Functional Data, 1st Edition, Chapman and Hall/CRC, New York, 2011. doi:10.1201/b11038.
[38] S. Puntanen, G. P. H. Styan, Schur complements in statistics and probability, in: F. Zhang (Ed.), The Schur Complement and Its Applications, Springer US, Boston, 2005, pp. 163–226. doi:10.1007/0-387-24273-2_7.
[39] J. Kocijan, Modelling and Control of Dynamic Systems Using Gaussian Process Models, Springer International Publishing, Cham, 2016. doi:10.1007/978-3-319-21021-6_2.
[40] E. Schulz, M. Speekenbrink, A. Krause, A tutorial on Gaussian Process regression: Modelling, exploring, and exploiting functions, Journal of Mathematical Psychology 85 (2018) 1–16. doi:10.1016/j.jmp.2018.03.001.
[41] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, Berlin, Heidelberg, 2006.
[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[43] J. Nocedal, S. Wright, Numerical Optimization, Springer New York, New York, 2006. doi:10.1007/978-0-387-40065-5.
[44] W. Cota, Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level, SciELOPreprints:362 (2020). doi:10.1590/scielopreprints.362.
URL https://doi.org/10.1590/scielopreprints.362
[45] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, J. Timmer, Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, Bioinformatics 25 (15) (2009) 1923–1929. doi:10.1093/bioinformatics/btp358.
[46] C. McAloon, Á. Collins, K. Hunt, A. Barber, A. W. Byrne, F. Butler, M. Casey, J. Griffin, E. Lane, D. McEvoy, P. Wall, M. Green, L. O'Grady, S. J. More, Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research, BMJ Open 10 (8) (2020) e039652. doi:10.1136/bmjopen-2020-039652.
[47] L. Taghizadeh, A. Karimi, C. Heitzinger, Uncertainty quantification in epidemiological models for the COVID-19 pandemic, Computers in Biology and Medicine 125 (2020) 104011. doi:10.1016/j.compbiomed.2020.104011.
[48] Brazilian Institute of Geography and Statistics, Demographic Census, accessed November 20, 2020 (2020).
URL https://www.ibge.gov.br/cidades-e-estados/rj/rio-de-janeiro.html
[49] D. N. Gujarati, D. C. Porter, Basic Econometrics, 5th Edition, McGraw-Hill Irwin, New York, 2008.
[50] R. Storn, K. Price, Differential Evolution—a simple and efficient heuristic for global optimization over continuous spaces, Journal of Global Optimization 11 (4) (1997) 341–359. doi:10.1023/A:1008202821328.
[51] J. A. Nelder, R. Mead, A simplex method for function minimization, The Computer Journal 7 (4) (1965) 308–313. doi:10.1093/comjnl/7.4.308.