Inversion of a SIR-based model: a critical analysis about the application to COVID-19 epidemic

M. Giudici[◇], A. Comunian[◇], R. Gaburro[⋆]
[◇]Università degli Studi di Milano, Italy
[⋆]University of Limerick, Ireland

April 17, 2020

Abstract

Calibration of a SIR (Susceptibles-Infected-Recovered) model with official data at international level for the COVID-19 pandemics provides a good example of the difficulties inherent the solution of inverse problems. Inverse modeling is set up in a framework of discrete inverse problems, which explicitly considers the role and the relevance of data. Together with a physical vision of the model, this is very useful to discuss the uncertainties on the data and how they influence the reliability of calibrated model parameters and, ultimately, of model predictions.

Keywords: Inverse problems; Mathematical modelling; Model calibration; Epidemic modeling

1 Introduction

Epidemic modeling is usually performed with compartmental models, often called SIR models, which are claimed to go back to the works by Ronald Ross and Hilda P. Hudson more than one century ago [13, 14] and by Anderson Gray McKendrick and William Ogilvy Kermack ten years later [9, 10]. This class of models shares several characteristics with models of population dynamics and with conceptual lumped models in hydrology. These models simulate the temporal evolution of some compartments of the population, which is normally subdivided among Susceptibles (i.e., those persons who have not yet been affected by the virus and which could be subject to infection), Infected (i.e., those persons who have been infected by the virus) and Recovered (i.e., those persons who have recovered, after having been infected). For this reason these models are usually referred to as SIR models. They are based on simple laws to describe the transfer of individuals from one class to the others.

These models have found wide application both in life sciences, mostly in epidemiology, and in the field of economic, political and social sciences, e.g., to assess the costs of different policies to block epidemics and the diffusion of viruses. Most papers consider academic issues and are rarely calibrated against real data.

Model calibration is a common problem in geophysical and environmental modeling. A general framework to handle discrete inverse problems for model calibration is proposed in [7] and can be useful to discuss some characteristics of SIR models and the role of data, also following the discussion in [6].

The wide number of data which are collected during the COVID-19 pandemic due to the diffusion of the SARS-CoV-2 virus (also called "coronavirus") provides an exceptional basis to perform some modeling exercises and to test different calibration methods.

The objectives of this paper are to fix some concepts about SIR models and their calibration and to discuss the relevance of data for reliability of model outcomes.

The paper is designed to advance the knowledge on the functioning, potentialities and limitations of epidemic models. Also it is expected to provide further insights in epidemic model calibration. Instead, it is not designed to provide forecasts of the pandemic evolution. In authors' opinion, the quality of available data does not permit to perform reliable forecast and model outcomes should be used with high prudence.

The paper is organized as follows. Next section is devoted to the description of the model, in the continuous and discrete case, of the methods used for the calibration of the numerical model, and of the data for the application to the COVID-19 pandemic; in particular, inverse modeling, i.e., model calibration, will be set up and discussed within the framework proposed by [7]. Some results of the model with reference to COVID-19 pandemic will be shown in the third section, whereas the fourth section will be devoted to a discussion of several topics: the assumptions at the basis of the SIR model; some remarks about model calibration; some remarks about data uncertainty. The concluding section will also include some hints for future developments of this work.

2 Methods and materials

2.1 The continuous model

This section is devoted to the description of the SIR model considered in this paper.

First of all, $S(t)$, $I(t)$, $R(t)$ and $D(t)$ represent the number of, respectively, susceptible, infected, recovered and dead individuals of the population under study as a function of time. Notice that D includes only those persons who died while being infected. $P = S + I + R$ denotes the total number of population individuals.

The coefficients β and δ denote the birth and death rate, respectively, under normal conditions, i.e., without considering deaths caused by the epidemic. These coefficients are rarely considered in epidemic modeling, as the variation due to the normal evolution of the population is negligible or at least smoother than the variations caused by epidemics.

The following equations, based on historical papers [13, 14, 9, 10], are used to describe the time evolution of S, I, D and R:

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \beta S - \gamma \frac{IS}{P} - \delta S, \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \gamma \frac{IS}{P} - \rho I - \phi I + \beta I, \\
\frac{\mathrm{d}D}{\mathrm{d}t} &= \phi I, \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \beta R + \rho I - \delta R.
\end{aligned} \tag{1}$$

The second term on the right hand side of the first equation of (1) represents the number of individuals who are infected per unit time. It is based on the assumption that each infected person has contacts with a given number of persons in a certain time interval and that the fraction of them who are susceptible is given by S/P, whereas $(I+R)/P$ is the fraction of those persons who cannot be infected, if it is assumed that recovered people are immunized. The coefficient γ is the infection coefficient, i.e., the rate of potential infection.

The coefficients ρ and ϕ represent the recovery and fatality rate, respectively. The coefficients β, γ, δ, ρ and ϕ are assumed to be constant and their dimension is [time^{-1}]. The assumptions behind this model are discussed thoroughly in section 4.

The initial conditions of the model are given by $S(0) = P(0)$, $I(0) = 1$, $R(0) = 0$ and $D(0) = 0$. This means that $t = 0$ corresponds to the time at which the first individual is infected.

Notice that from equations (1), it follows that

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \beta P - \delta P - \phi I. \tag{2}$$

2.2 The discrete model

The discrete model is a simple forward-time finite-differences discretization of equations (1). Therefore, if $t' \in \mathbb{N}$ is the index used to denote the discrete time steps and a uniform spacing Δt is considered, then the following explicit iterative equations are obtained:

$$\begin{aligned} S(t'+1) &= \left\{1+\left[\beta-\gamma\frac{I(t')}{P(t')}-\delta\right]\Delta t\right\}S(t'), \\ I(t'+1) &= \left\{1+\left[\gamma\frac{S(t')}{P(t')}-\rho-\phi+\beta\right]\Delta t\right\}I(t'), \\ D(t'+1) &= D(t')+\phi I(t')\Delta t, \\ R(t'+1) &= \left[1+(\beta-\delta)\Delta t\right]R(t')+\rho I(t')\Delta t. \end{aligned} \tag{3}$$

The initial conditions of the discrete model are given by:

$$S(0)=P_0-1, \qquad I(0)=1, \qquad D(0)=0, \qquad R(0)=0, \tag{4}$$

where P_0 is the population at $t'=0$.

The results presented in this paper consider the application of the model to a nation. In other words, the population of the whole nation is considered, without any further subdivision in provinces, regions or states. Moreover, the time spacing is 1 day, in agreement with the sampling of the available data set on COVID-19 pandemic (see section 2.4).

2.3 The inverse problem

As stated in the introduction, the inverse problem is defined by making use of the conceptual framework and the notation of [7].

The time-varying state of the system is included in an array $\mathbf{s}$:

$$\mathbf{s}=\left\{S(t'),I(t'),R(t'),D(t'),\, t'=0,\dots,N_{\mathrm{mod}}-1\right\}, \tag{5}$$

where N_{mod} is the number of modeled time steps.

The available data are collected in an array $\mathbf{d}$ that, in the specific case considered here, includes the number of infected, recovered and dead persons, released by sanitary official organizations:

$$\{I_{\mathrm{obs}}(t),R_{\mathrm{obs}}(t),D_{\mathrm{obs}}(t),\, t=0,\dots,N_{\mathrm{obs}}-1\}\subset\mathbf{d}. \tag{6}$$

N_{obs} denotes the number of time steps for which data are available. Notice that in the discrete case, $t\in\mathbb{Z}$ is used as the time-index which denotes days, starting from the reference date, taken as $t=0$, which corresponds to the first day for which epidemic data are available.

The model parameters are included in an array $\mathbf{p}$:

$$\mathbf{p}=\{\beta,\delta,\Delta t,\rho,\phi,\gamma,\, t_0,P_0\}, \tag{7}$$

where t_0 represents the day, at which the first individual of the population is infected and the total population is P_0.

The discrete model given by equations (3) and (4), can be written as the following system of equations

$$\mathbf{f}(\mathbf{p},\mathbf{s}) = 0. \tag{8}$$

If the numbers of model parameters and state parameters are $N_{(\mathrm{p})}$ and $N_{(\mathrm{s})}$, respectively, then $\mathbf{p} \in \mathcal{P} \subseteq \mathbb{R}^{N_{(\mathrm{p})}}$ and $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^{N_{(\mathrm{s})}}$, where the subset $\mathcal{P}$ takes into account some conditions on the parameters, e.g., that the parameters ρ, ϕ, γ and P_0 must satisfy the following constraints:

$$0 \le \rho \le 1-\phi \le 1, \quad 0 \le \phi \le 1-\rho \le 1, \quad 0 \le \gamma, \quad 0 \le P_0. \tag{9}$$

The unique solution of the forward problem, i.e., of (8), can be expressed as

$$\mathbf{s} = \mathbf{g}(\mathbf{p}). \tag{10}$$

The array $\mathbf{p}$ can be subdivided in the two sub-arrays

$$\mathbf{p}^{(\mathrm{fix})} = \{\beta, \delta, \Delta t\}, \tag{11}$$

which includes the model parameters, whose values are fixed before the simulation, and

$$\mathbf{p}^{(\mathrm{cal})} = \{\rho, \phi, \gamma, t_0, P_0\}, \tag{12}$$

which includes the model parameters, whose values are obtained from the solution of an inverse problem. Therefore

$$\mathbf{p} = \left(\mathbf{p}^{(\mathrm{fix})^t}, \mathbf{p}^{(\mathrm{cal})^t}\right)^t. \tag{13}$$

The array of fixed parameter could be a function of $\mathbf{d}$: $\mathbf{p}^{(\mathrm{fix})}(\mathbf{d})$.

The model outcome, i.e., the state of the system, is used to forecast the number of infected, recovered and dead individuals at the times for which some estimates are available from epidemiological data. Therefore, the model forecast is expressed as an array $\mathbf{y}$, which is function of $\mathbf{s}$, $\mathbf{p}$ and $\mathbf{d}$: $\mathbf{y}\left(\mathbf{d},\mathbf{s},\mathbf{p}\right)$.

Then, model calibration requires that the model forecast be close to a calibration target, an array $\mathbf{t}$ that collects the values which should be attained by the model forecast, if the model were physically "correct" and the model parameters were "optimal". Recall that $\mathbf{t}$ may depend on $\mathbf{d}$ and $\mathbf{p}^{(\mathrm{fix})}$, but should be independent of $\mathbf{p}^{(\mathrm{cal})}$: $\mathbf{t} = \mathbf{t}\left(\mathbf{d}, \mathbf{p}^{(\mathrm{fix})}\right)$.

The correspondence between model forecast and data must take into account that $t' = t - t_0$, so that $t' = 0$ when $t = t_0$, i.e., the time at which the first individual is infected. Moreover, the objective of model calibration will be to find the parameter values which best fit the observations in a given time interval, $t_1 \le t < t_2$. Since data and model outcomes refer to different time intervals, the following inequalities must be verified:

$$0 \le t - t_0 < N_{\mathrm{mod}}, \quad 0 \le t_1 \le t < t_2 \le N_{\mathrm{obs}}. \tag{14}$$

Equation (14) are satisfied if one picks

$$t_1 \geq t_0 \quad \text{and} \quad t_2 \leq \min\left(N_{\mathrm{mod}} + t_0, N_{\mathrm{obs}}\right). \tag{15}$$

Then, the arrays $\mathbf{y}$ and $\mathbf{t}$ are given by:

$$\begin{aligned} \mathbf{y} &= \left\{I(t'), R(t'), D(t'),\, t_1 - t_0 \leq t' < t_2 - t_0\right\}, \\ \mathbf{t} &= \left\{I_{\mathrm{obs}}(t), R_{\mathrm{obs}}(t), D_{\mathrm{obs}}(t),\, t_1 \leq t < t_2\right\}. \end{aligned} \tag{16}$$

The misfit between model predictions and the target values is computed by means of the following objective function:

$$\begin{aligned} \mathsf{O}\left(\mathbf{p}^{(\mathrm{cal})}\right) &= \left\|\mathbf{y}\left(\mathbf{d}, \mathbf{s}, \mathbf{p}\right) - \mathbf{t}\left(\mathbf{d}, \mathbf{p}^{(\mathrm{fix})}\right)\right\| \\ &= \mathsf{O}_I\left(\mathbf{p}^{(\mathrm{cal})}\right) + \mathsf{O}_R\left(\mathbf{p}^{(\mathrm{cal})}\right) + \mathsf{O}_D\left(\mathbf{p}^{(\mathrm{cal})}\right), \end{aligned} \tag{17}$$

where each of O_I, O_R and O_D is defined by

$$\mathsf{O}_X\left(\mathbf{p}^{(\mathrm{cal})}\right) = \left\{\frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2-1} \left[\frac{X(t-t_0) - X_{\mathrm{obs}}(t)}{\max(\xi, X_{\mathrm{obs}}(t))}\right]^2\right\}^{1/2}, \tag{18}$$

with $X \in \{I, R, D\}$, X_{obs} being the corresponding element of $\{I_{\mathrm{obs}}, R_{\mathrm{obs}}, D_{\mathrm{obs}}\}$, and $\xi \geq 1$ is a threshold. In other words, O is the sum of three functions, each of which considers one of the three observed quantities, separately.

The model calibration is performed by solution of the following inverse problem: given $\mathbf{p}^{(\mathrm{fix})}$ and $\mathbf{d}$, given the solution $\mathbf{s} = \mathbf{g}\left(\mathbf{p}\right)$ to (8), given the functions $\mathbf{y}\left(\mathbf{d}, \mathbf{g}\left(\mathbf{p}\right), \mathbf{p}\right)$ and $\mathbf{t}$, find $\mathbf{p}^{(\mathrm{cal})\star} \in \mathcal{P}^{(\mathrm{cal})}$, such that

$$\begin{aligned} &\mathbf{p}^{(\mathrm{cal})\star} = \arg \min_{\mathbf{p}^{(\mathrm{cal})} \in \mathcal{P}^{(\mathrm{cal})}} \mathsf{O}\left(\mathbf{p}^{(\mathrm{cal})}\right), \\ &\text{i.e., } \mathsf{O}\left(\mathbf{p}^{(\mathrm{cal})\star}\right) \leq \mathsf{O}\left(\mathbf{p}^{(\mathrm{cal})}\right),\ \forall \mathbf{p}^{(\mathrm{cal})} \in \mathcal{P}^{(\mathrm{cal})}, \end{aligned} \tag{19}$$

where $\mathcal{P}^{(\mathrm{cal})} = \left\{\mathbf{p}^{(\mathrm{cal})} : \left({\mathbf{p}^{(\mathrm{fix})}}^t, {\mathbf{p}^{(\mathrm{cal})}}^t\right)^t \in \mathcal{P}\right\}$.

The threshold $\xi \in \mathbb{R}$ plays a double role. First of all, it keeps positive the denominator of the fraction appearing in (18). Furthermore, it controls some characteristics of the objective function. For $\xi = 1$, O_X is nothing but the root-mean-squared relative difference between observed and modeled values of X. If a large value of ξ is used, then relative errors corresponding to large values of X_{obs} will be dominant; from the practical point of view, this means that early time behavior has a minor relevance for the model fitting. In particular, if $\xi > \max\{X_{\mathrm{obs}}(t),\, t_1 \leq t < t_2\}$, then O_X reduces to the standard root-mean-squared error.

2.4 Data and computer implementation for COVID-19

The application of the model introduced in section 2.2 and of the model calibration introduced in section 2.3 can be attempted thanks to publicly available data on COVID-19 pandemic. The application will be performed at national level, i.e., the considered population will be the whole population of some countries. For each country, the array $\mathbf{d}$ is populated with data coming from two basic sources.

Data on COVID-19 pandemic are available from the GitHub repository managed by the John Hopkins University [3]. This is a collection of publicly available data from multiple sources, which are processed and delivered by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Notice that the data are provided to the public strictly for educational and academic research purposes. The data are updated daily and the files used in this paper have been downloaded from `https://github.com/CSSEGISandData/COVID-19` on April 16, 2020. From those files, the array $\mathbf{t}$ is easily built.

A tailored code has been developed under Python 3.7.6 to download data from the Github repository, perform forward modeling and calibrate the model by solution of the inverse problem. In particular, inversion is based on the functions of the `optimize` module from SciPy v1.4.1.

Figure 1 shows the trend of confirmed cases, recovered and deceased people for some countries, among those that have been considered as the most relevant for the analysis of COVID-19 pandemic not only by the scientific community, but also by mass media. These plots show different trends for different countries and for the different quantities.

Aside from China, for which the starting phase is not reported, because the virus diffusion started earlier than the first date for which data are available in the data set, the number of confirmed cases (top plot in Figure 1) shows a first slow increase, followed by an exponential increase and possibly a slowdown after few weeks. It is highly questionable whether this behavior is related to the number of tests performed to confirm virus infection.

The most regular trends are clearly the ones describing the number of deceased people (bottom plot in Figure 1), after about one week since the first reported case in each country considered in this study. Doubts about comprehensiveness of official data on deaths caused by coronavirus have been raised by several sources of information and by some commentators. Nevertheless, it seems safe to state that the number of deaths represents the smoothest time series and possibly the less affected by uncertainties.

Notice that the daily sampling rate of these data induces to choose $\Delta t = 1\,\text{day}$, and the coefficients β, γ, δ, ϕ and ρ share the same measurement units, namely day^{-1}.

The second data source is the most updated version of the UN Demographic Yearbook [12]. Demographic data have been extracted from this

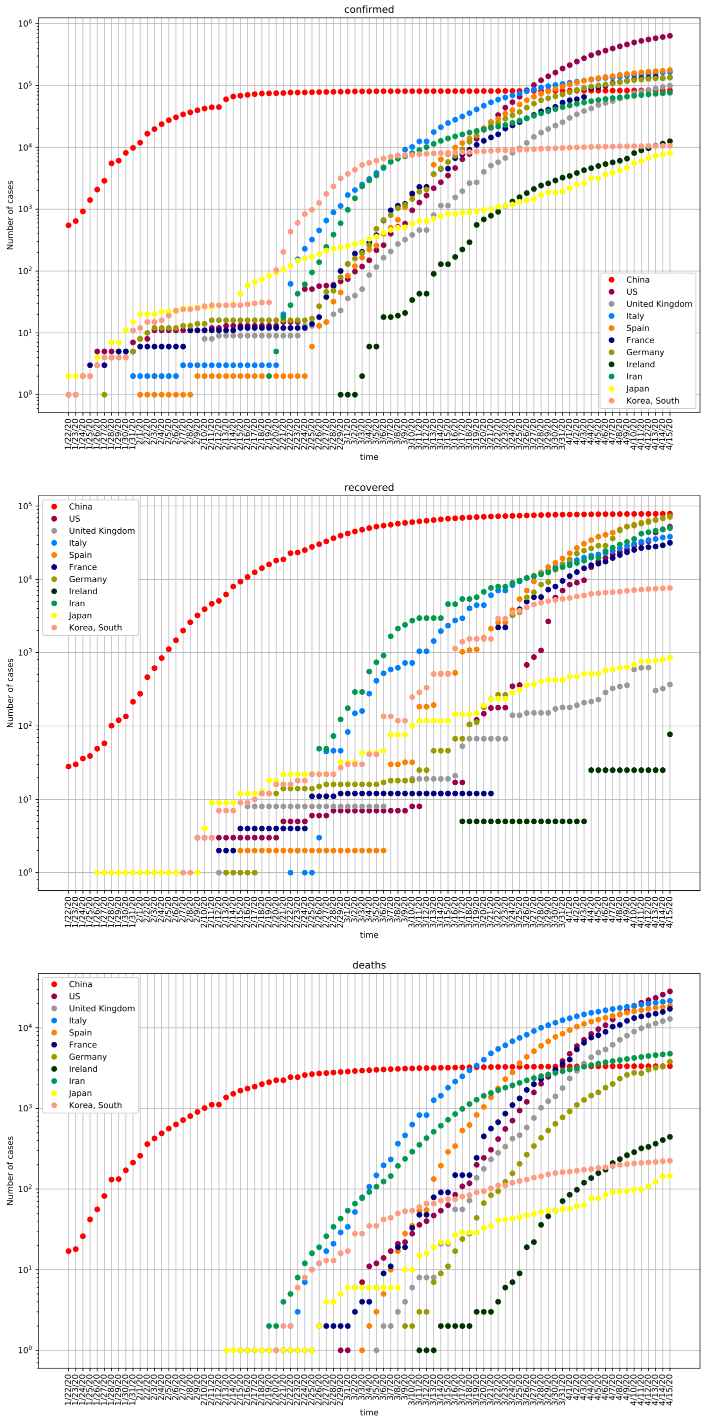


Figure 1: Data about COVID-19 pandemic in selected countries. Top: confirmed infections. Middle: recovered patients. Bottom: deaths.

Table 1: Parameter values for test case 1.

run	β	δ	γ	ρ	ϕ	P_0
(a)	$0.009\,\mathrm{y}^{-1}$	$0.0011\,\mathrm{y}^{-1}$	$0.2\,\mathrm{day}^{-1}$	$0.01\,\mathrm{y}^{-1}$	$0.001\,\mathrm{day}^{-1}$	10^9
(b)	idem	idem	idem	$0.05\,\mathrm{y}^{-1}$	idem	idem
(c)	idem	idem	idem	$0.1\,\mathrm{y}^{-1}$	idem	idem

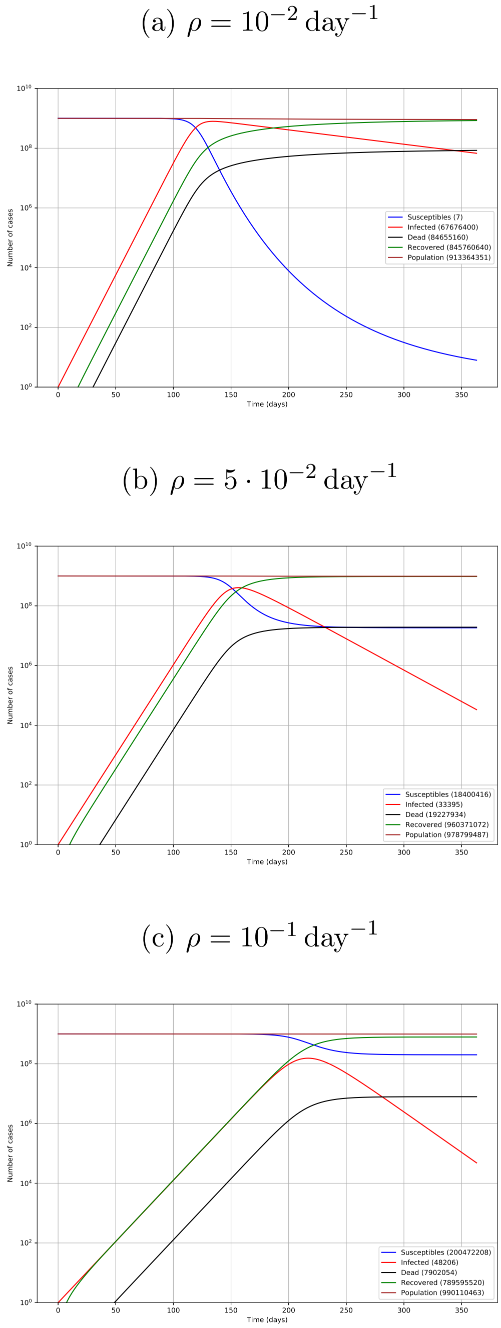
volume. The values of population, birth and death rate of each country, for which the model has been tested, are included in **d**. They are used to fix the values of β and δ, which are expressed on a daily basis, and to provide a first estimate of P_0.

3 Results

3.1 Model results

First of all the behavior of the model is shown with test case 1, which includes three model runs for which all the model parameters are kept fixed, but ρ: the list of parameter values is given in Table 1; the results of the model for a one-year-long simulation period are shown in Figure 2. The general behavior shows an exponential increase in the number of infected persons (notice that the vertical axis is in logarithmic scale) followed by an exponential decrease but with a longer characteristic time. The number of deaths obviously decreases if ρ increases and in particular, we have three different situations for the three runs: (a) for the smallest value of ρ, the curve of susceptible persons dramatically decreases from some days before the peak of infections and reaches very small values after few weeks; (b) for the intermediate value of ρ, the chosen values of model parameters yield a stationary conditions after about 8 months from the start of the epidemic for the number of susceptible and dead people, which reach almost the same value; (c) for the highest value of ρ, the number of susceptible people decreases with time, but remains consistent. Notice that, for this test case, the reduction of the total population is limited, less than 10%, and after one year almost all the living population is recovered. It is important to stress that this test case has the goal of showing how the model can predict different behavior and these results should not be considered as a forecast of the actual behavior of any real pandemic.

SIR models are sometimes applied using the ratio of the number of individuals in each category with respect to the total population as a variable. Even if test case 1 showed that for three sets of model parameters, which differ only for the value of ρ, the total population shows only a limited variation, nevertheless, the term used to compute the infection rate introduces a non-linearity in the model. Therefore test case 2 is designed to assess the effect of P_0 on model results. P_0 values span four orders of magnitude,



(a) $\rho = 10^{-2}\,\mathrm{day}^{-1}$

(b) $\rho = 5 \cdot 10^{-2}\,\mathrm{day}^{-1}$

(c) $\rho = 10^{-1}\,\mathrm{day}^{-1}$

Figure 2: Model results for test case 1. Numbers in the legends refer to the values at the end of the simulation period.

from 10^6 to 10^9, whereas the other parameters are fixed at the values of run (a) of test case 1. The results are shown in Figure 3 as functions of the normalized quantities versus time. The values of each function at the end of the simulation period are very similar. The main differences are in the evolving phase, for which the response of a small population appears to be more rapid than that of a large population. Roughly speaking, the curves corresponding to high populations show a delay with respect to the curve for the smallest population of about 15 days per an increase in P_0 by an order of magnitude. This remark, if confirmed by runs with more reliable parameter sets, could have fundamental consequences in the design of early warning systems.

3.2 Model calibration

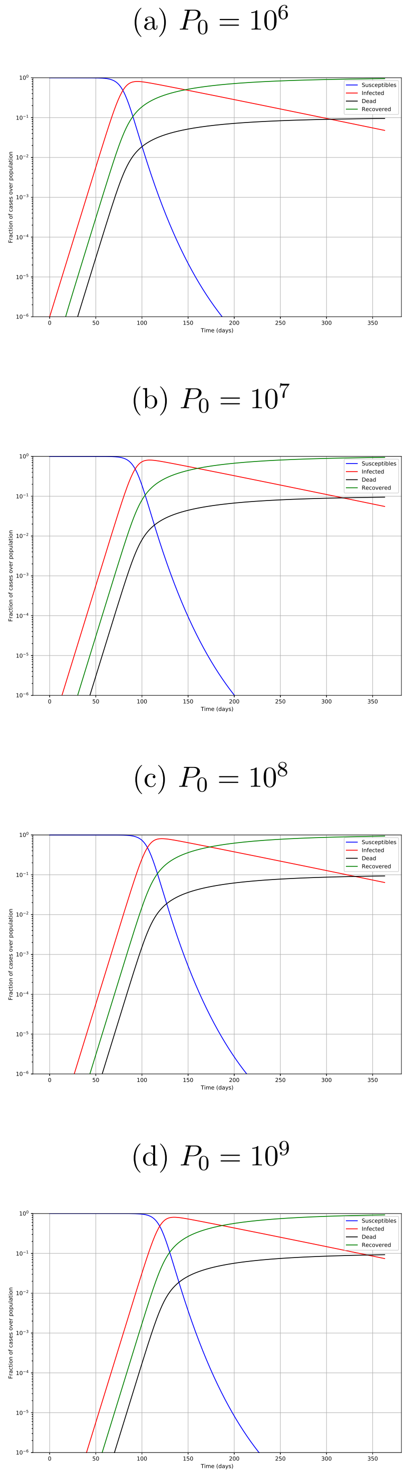
Model calibration for the COVID-19 pandemic by solution of the inverse problem is a very challenging problem. This is not surprising at all, because the comparison of the trends of the model time series (Figure 2) with those observed from the data and drawn in Figure 1 shows that the SIR model can hardly reproduce the observed trend.

In particular, this paper is focused on the results obtained with data from Italy, but the same qualitative remarks apply also to other cases.

The best "trial-and-error" estimate is shown in Figure 4 and is obtained for the following values of the parameters to be calibrated: $\gamma = 0.33\,\mathrm{day}^{-1}$, $\rho = 0.015\,\mathrm{day}^{-1}$, $\phi = 0.0025\,\mathrm{day}^{-1}$, $t_0 = 10$.

Starting from this initial set of parameters, minimization of the objective function O was performed with a SciPy function which implements several methods to find a minimum, also by taking into account possible bounds on the independent variable of the objective function. Several tests have been conducted and have shown that the best results were obtained with the L-BFGS-B method, which is a variation of the BroydenFletcherGoldfarb-Shanno (BFGS) algorithm [5] to reduce memory requirements and to handle simple constraints. The bounds assigned for the parameters to be calibrated are listed in Table 2. If the whole set of data is used, i.e., $t_1 = t_0 = 10$, $t_2 = N_{\mathrm{obs}}$, the application of this algorithm leads to a parameter set for which γ is very small, in fact close to the lower bound.

The application of `differential_evolution`, an algorithm of global minimization [16], yields the following set of parameters when $\xi = 10^6$, which is equivalent to considering root-mean-squared error for O_X: $\gamma = (0.1958 \pm 2 \cdot 10^{-5})\,\mathrm{day}^{-1}$, $\rho = (1.289 \cdot 10^{-2} \pm 2 \cdot 10^{-5})\,\mathrm{day}^{-1}$, $\phi = (8.14 \cdot 10^{-3} \pm 2 \cdot 10^{-5})\,\mathrm{day}^{-1}$, $t_0 = -9$, $P_0 = 218,200 \pm 200$. The mean value and its standard deviation of each parameter has been estimated from 10 runs of this stochastic algorithm, which introduces small variations in the returned results. The comparison between observed and fitted time series is shown in Figure 5. Two facts should be mentioned: $t_0 < 0$, i.e., it seems that the


(a) $P_0 = 10^6$

(b) $P_0 = 10^7$

(c) $P_0 = 10^8$

(d) $P_0 = 10^9$

Figure 3: Model results for test case 2.

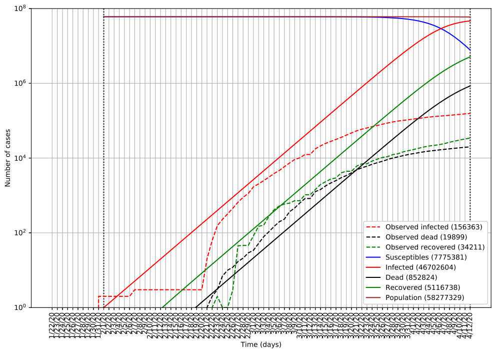


Figure 4: Comparison of observations for Italy and modelled data with the parameters obtained from subjective "trial-and-error" calibration.

Table 2: Intervals of variation fixed for the parameters to be calibrated for inversion of data referred to Italy.

	γ	ρ	ϕ	t_0	P_0
minimum	$10^{-4}\,\text{day}^{-1}$	$10^{-5}\,\text{day}^{-1}$	$10^{-6}\,\text{day}^{-1}$	-10	$2 \cdot 10^5$
maximum	$1\,\text{day}^{-1}$	$0.1\,\text{day}^{-1}$	$0.1\,\text{day}^{-1}$	30	10^8

infection started before the official appearance of the first confirmed case; P_0 is close to the lower bound, so that the model predicts that the population which has been involved in the infection could be relatively small.

These results suggested a further test case; in particular, the minimum admissible value for t_0 has been lowered to -30. The optimal set of values for this test are: $\gamma = (0.1543 \pm 1.5 \cdot 10^{-5})\,\text{day}^{-1}$, $\rho = (1.237 \cdot 10^{-2} \pm 1 \cdot 10^{-5})\,\text{day}^{-1}$, $\phi = (7.924 \cdot 10^{-3} \pm 2 \cdot 8.5 \cdot 10^{-6})\,\text{day}^{-1}$, $t_0 = -29$, $P_0 = 247,490 \pm 96$. The comparison between observed and modeled data is shown in Figure 6. The value of the function O decreases from $8.3 \cdot 10^{-3}$ for the results of Figure 5 to $5.2 \cdot 10^{-3}$ for those of Figure 6, but the visual inspection shows a moderate difference.

If $\xi = 1$, then $\gamma = (0.31 \pm 0.03)\,\text{day}^{-1}$, $\rho = (1.9 \pm 0.3) \cdot 10^{-2}\,\text{day}^{-1}$, $\phi = (1.7 \pm 0.4) \cdot 10^{-2}\,\text{day}^{-1}$, $t0 = 19 \pm 2$, $P_0 = 16,000,000 \pm 10,000,000$. Indeed, for this test case, the returned "optimal" value of O has a coefficient of variation equal to 13%, much larger than those obtained in the test cases for $\xi = 10^6$, which were less than 0.06%.

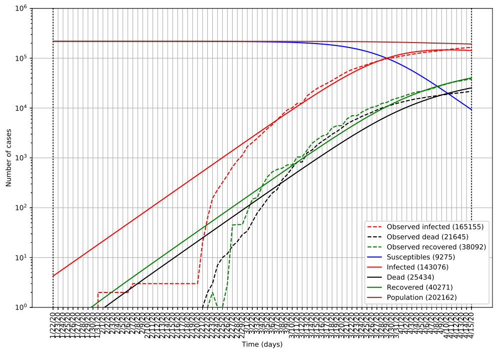


Figure 5: Comparison of observations for Italy and modelled data with the parameters obtained by solution of the inverse problem with a global minimization algorithm, for $t_0 = -9$. Vertical dotted line delimit the data set used for model calibration, i.e., they correspond to t_1 and t_2.

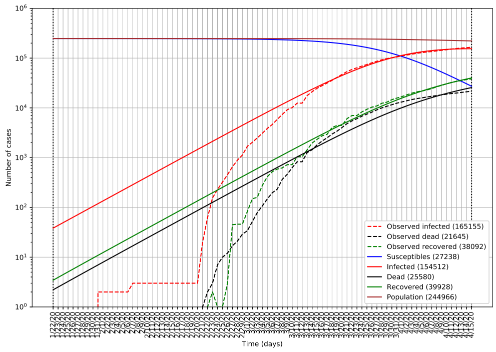


Figure 6: Comparison of observations for Italy and modelled data with the parameters obtained by solution of the inverse problem with a global minimization algorithm, for $t_0 = -29$. Vertical dotted line delimit the data set used for model calibration, i.e., they correspond to t_1 and t_2.

4 Discussion

4.1 Remarks about the model

Some basic assumptions, on which the model developed in this work is funded, deserve to be recalled and discussed.

The developed model basically assumes "homogeneity" of the population. In other words, no distinction is done in terms of sex, age, economic wealth, health and wellness, working conditions, life style, home state, and any other, including genetic background. Also, the model assumes that the population under study is a closed system, thus disregarding variations induced by short-time, touristic or business travels, by intermediate-time mobility of students and workers and by long-time effects of migrant fluxes.

The model is also independent of the climatic and environmental conditions, i.e., the processes considered by the model are assumed to be independent of the variability of weather conditions and environmental quality at any temporal and space scale. In particular, this means that neither sharp and rapid variations nor annual or seasonal cycling should affect these processes.

Epidemic models rarely consider birth and death rates, because the corresponding terms in the equations are usually negligible. In this work, these terms have been kept, in order to facilitate this discussion. In particular, following the assumption of population homogeneity, it is assumed that infected pregnant women give birth to infected babies and that this occur at the same rate as for susceptible women.

With regard to infection rate, which is described by the term $\gamma IS/P$ in (1), some remarks are in order. This term is computed by assuming that each infected individual has a given, constant number of contacts with other persons per unit time. The fraction of contacted persons which cannot be infected is given by $(I+R)/P$, which assumes that recovered people become immune to the virus, an aspect which is not confirmed by the scientific community (see, e.g., [15]). Moreover, recovered people are assumed to be not infectious, which is the case if the response of their immune system is so fast that, once they come in contact with the virus again, the virus is destroyed by the immune system before it can be spread to susceptible persons. On the other hand, the fraction of contacted individuals which can be infected is given by S/P. The γ coefficient, due to the "homogeneity" assumption, is considered to be independent of the factors which have been recalled at the beginning of this subsection; in particular, working and living conditions could control the distance and the duration of contacts of infected - and therefore infectious - individuals with other persons.

The so-called recovery and fatality coefficients ρ and ϕ are assumed to be constant. This is not based on the "homogeneity" assumption only. In fact, this implies that recovery and fatality are modeled as instantaneous processes, i.e., independent of the time passed since infection; moreover, no distinction is done among death or healing of infected people according to the strength of their symptoms and to the location where they are treated (home and hospitals non-intensive, or Intensive Care Units – ICUs). The latter condition could be modeled by subdividing the class of infected people among sub-classes, e.g., asymptomatic, with light symptoms, admitted to hospital non-intensive care units, admitted to ICUs.

Accounting for the time from infection is slightly more complex, but could be handled, for instance, by introducing functions ($\tilde{\phi}$, $\tilde{\rho}$) of elapsed time since infection. Such functions should enter in a deconvolution product involving the number of persons who have been infected at a given time and are still infected, i.e., are not yet recovered or passed away. With this approach, ϕI and ρI in (1) could be replaced by

$$\begin{gathered}\int_0^{\tau_{\max}} \tilde{\phi}(\tau)\tilde{I}(t-\tau)\,\mathrm{d}\tau \quad \text{and} \quad \int_0^{\tau_{\max}} \tilde{\rho}(\tau)\tilde{I}(t-\tau)\,\mathrm{d}\tau, \\ \text{where} \quad I(t-\tau) = \frac{\mathrm{d}I}{\mathrm{d}t}(t-\tau)\exp\left\{-\int_0^{\tau}\left[\tilde{\rho}(\tau') + \tilde{\phi}(\tau')\right]\,\mathrm{d}\tau' - \delta\tau\right\}.\end{gathered} \tag{20}$$

Notice that the fatality coefficient, ϕ, accounts for the deaths related to the pandemic, i.e., it represents the increase in the death rate due to the pandemic. The normal death rate is considered through δ.

4.2 Remarks about model calibration by solution of the inverse problem

The results presented in section 3.2 show some of the classical, well known difficulties of non-linear least-squares inversion, in particular the dependence of the solution on the starting values, related to the existence of multiple local minima, and the flatness of the objective function around the local minima.

Better results have been obtained with the "differential evolution" algorithm. Obviously, different algorithms for global optimization could be tested, like, e.g., genetic algorithms [2], particle swarm optimization [8], simulated annealing [11].

With reference to the specific example under study, it is necessary to stress some aspects, mostly related to the role of data on model calibration [6].

First of all, the solutions obtained by means of a global optimization algorithm for high values of the threshold ξ and different intervals of admissible values for t_0 (Figures 5 and 6), show that the optimal value of P_0 is smaller than the total Italian population. This parameter P_0 has been included in $\mathbf{p}^{(\mathrm{cal})}$ with the objective of assessing the extension of the reference population. In other words, including P_0 among the parameters to be calibrated might provide a, possibly very rough, estimate of the width of the initial population whose evolution is represented by the model. In this particular instance, the results suggest that the reference initial population does not cover the whole country, but only a limited portion.

The latter remark seems to go in tandem with the well-known fact that in the countries most affected by COVID-19, the epidemic spread of the virus had mostly concentrated in specific areas: the province of Hubei, and above all the city of Wuhan, in China; the Lombardy region, and above all the provinces of Bergamo, Brescia, Lodi and Milan, in Italy; the city of New York in the USA; Île-de-France in France; Madrid and Catalunya in Spain.

Finally, it is quite difficult to assess the quality of data on the COVID-19 pandemia, but their uncertainty is expected to be very high. For instance, the true number of infected people "remains unknown because asymptomatic cases or patients with very mild symptoms might not be tested and will not be identified", as recognized, e.g., by [1]. In an interview published on March 23rd 2020 by the Italian newspaper "La Repubblica", Angelo Borrelli, head of Dipartimento della Protezione civile (national civil protection department) stated that a ratio of one certified case out of every 10 total cases is credible. Furthermore, different criteria have been adopted by different countries and institutions to define the various categories of infected, recovered and deceased people by or with COVID-19. This fact has been widely recognised as a cause of uncertainty in the collected data. Finally, censorship on COVID-19 pandemics is reported by journalists and organization in some countries.

As a consequence, the use of official data to perform reliable estimates is questionable. In principle, stochastic approaches, e.g., the Bayesian framework, could be very helpful to handle discrepancies between model predictions and observations. Unfortunately, in this case the systematic and random errors could be so high as to make it very difficult to handle them even in a stochastic framework.

5 Conclusions

The modeling exercis conducted within this work supports a series of remarks, which are summarized in this conclusive section, together with some future perspectives.

Starting from some remarks about modeling aspects, the limitations of classical SIR models have been recalled. These should be always recalled and carefully considered especially for applications and when these models are used as engines of decision support systems.

The main limitation is related to the "homogeneity" assumption, accompanied by the steadiness of the recovery and fatality coefficients. The latter aspect could be taken care of as discussed in subsection 4.1 and might yield terms of the form given in (20).

The assumption of "homogeneity" could be relaxed by considering "distributed" models, similar to those applied for transport phenomena, e.g., for diffusion of contaminants in the environment. Those models can account for "diffusive" spread and for "advective" transport. However, the required parameterisation is often much finer than the one for lumped models, so that the number of parameters to be calibrated strongly increases, and therefore in absence of good quality data it could be difficult to perform a reliable calibration and validation of the model for a practical application.

Promising classes of models are given by the use of stochastic models, either under a Monte Carlo framework or by using assimilation techniques, e.g., the Ensemble Kalman Filter (EnKF, see, e.g., [4]). In principle, Monte Carlo models might be adapted in a relatively easy way to account for several phenomena and also to consider the role of some aspects (e.g., sex, age, health and wellness, etc.) on the probability of infection. On the other hand, EnKF could provide a firm theoretical framework to improve model predictions by means of uncertain data.

With regard to the specific application to COVID-19 epidemic, although it could be improvident to draw quantitative conclusions, it is nevertheless qualitatively confirmed that infection started quite earlier than the certain appearance of the first episodes of infection. The results of model inversion also suggest that the calibrated model could be reliable for a portion of the whole population. Somehow, the model itself, through its calibration, seems to suggest the width of the population for which its approximations could be valid.

Last, but not least for its practical importance, this paper has the ambition to provide further evidence about the great care that has to be given to the quality of pandemic data, when used to calibrate or validate epidemic models. In fact, poor quality data might yield unrealistic parameter values and, therefore, unreliable model predictions.

Acknowledgments

This work presents results of a purely curiosity-driven research, which has received support only through the standard working facilities of the authors' institutions.

References

[1] David Baud, Xiaolong Qi, Karin Nielsen-Saines, Didier Musso, Léo Pomar, and Guillaume Favre, *Real estimates of mortality following covid-19 infection*, The Lancet Infectious Diseases (2020), DOI:10.1016/S1473-3099(20)30195-X.

[2] Lawrence Davis, *Handbook of genetic algorithms*, Van Nostrand Reinhold, New York, 1991.

[3] Ensheng Dong, Hongru Du, and Lauren Gardner, *An interactive web-based dashboard to track covid-19 in real time*, The Lancet Infectious Diseases (2020), DOI:10.1016/S1473-3099(20)30120-1.

[4] Geir Evensen, *The ensemble kalman filter: theoretical formulation and practical implementation*, Ocean Dynamics **53** (2003), 343–367, DOI:10.1007/s10236-003-0036-9.

[5] Roger Fletcher, *Practical methods of optimization*, second edition ed., John Wiley & Sons, Ltd, 2013.

[6] Mauro Giudici, *Development, calibration and validation of physical models*, Geographic Information Systems and Environmental Modeling (M. C. Krane K. C. Clarke, B. O. Parks, ed.), Prentice-Hall, Upper Saddle River (NJ), 2001, pp. 100–121.

[7] Mauro Giudici, Fulvia Baratelli, Laura Cattaneo, Alessandro Comunian, Giovanna De Filippis, Cinzia Durante, Francesca Giacobbo, Silvia Inzoli, Mauro Mele, and Chiara Vassena, *A conceptual framework for discrete inverse problems in geophysics*, 2019, arXiv:1901.07937.

[8] James Kennedy and Russell C. Eberhart, *Particle swarm optimization*, Proceedings of ICNN'95 - International Conference on Neural Networks, vol. 4, 1995, pp. 1942–1948.

[9] William Ogilvy Kermack and Anderson G McKendrick, *A contribution to the mathematical theory of epidemics*, Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character **115** (1927), 700–721, DOI:10.1098/rspa.1927.0118.

[10] ———, *Contributions to the mathematical theory of epidemics. ii. the problem of endemicity*, Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character **138** (1932), 55–83, DOI:10.1098/rspa.1932.0171.

[11] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi, *Optimization by simulated annealing*, Science **220** (1983), 671–680, DOI:10.1126/science.220.4598.671.

[12] Department of Economic and Social Affairs, *2018 demographic yearbook annuaire démographique*, sixty-ninth issue/soixante-neuvième édition ed., United Nations, 2019.

[13] Ronald Ross, *An application of the theory of probabilities to the study of a priori pathometry.–part i*, Proceedings of the Royal Society of London.

Series A, Containing papers of a mathematical and physical character **92** (1916), 204–230, DOI:10.1098/rspa.1916.0007.

[14] Ronald Ross and Hilda P. Hudson, *An application of the theory of probabilities to the study of a priori pathometry.–part ii*, Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character **93** (1917), 212–225, DOI:10.1098/rspa.1917.0014.

[15] Yufang Shi, Ying Wang, Changshun Shao, Jianan Huang, Jianhe Gan, Xiaoping Huang, Enrico Bucci, Mauro Piacentini, Giuseppe Ippolito, and Gerry Melino, *Covid-19 infection: the perspectives on immune responses*, Cell Death & Differentiation (2020), DOI:10.1038/s41418-020-0530-3.

[16] Rainer Storn and Kenneth Price, *Differential evolution a simple and efficient heuristic for global optimization over continuous spaces*, Journal of Global Optimization **11** (1997), 341–359, DOI:10.1023/A:1008202821328.

Authors' affiliations

Mauro Giudici
Università degli Studi di Milano, Dipartimento di Scienze della
Terra “A.Desio”, Milano, Italy
mauro.giudici@unimi.it

Alessandro Comunian
Università degli Studi di Milano, Dipartimento di Scienze della
Terra “A.Desio”, Milano, Italy
alessandro.comunian@unimi.it

Romina Gaburro
University of Limerick, Department of Mathematics and Statistics,
Health Research Institute (HRI), Limerick, Ireland
romina.gaburro@ul.ie