



Underdetection in a stochastic SIR model for the analysis of the COVID-19 Italian epidemic

Antonella Bodini¹ · Sara Pasquali¹ · Antonio Pievatolo¹ · Fabrizio Ruggeri¹

Accepted: 30 July 2021 / Published online: 28 August 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

We propose a way to model the underdetection of infected and removed individuals in a compartmental model for estimating the COVID-19 epidemic. The proposed approach is demonstrated on a stochastic SIR model, specified as a system of stochastic differential equations, to analyse data from the Italian COVID-19 epidemic. We find that a correct assessment of the amount of underdetection is important to obtain reliable estimates of the critical model parameters. The adaptation of the model in each time interval between relevant government decrees implementing contagion mitigation measures provides short-term predictions and a continuously updated assessment of the basic reproduction number.

Keywords Susceptible-infected-removed · Underdetection · State-space SDE · Identifiability · Infection fatality rate · Basic reproduction number

1 Introduction

In early 2020 the outbreak of SARS-CoV-2 started in Wuhan, China and spread to several other countries, causing the respiratory disease named COVID-19. At the end of January 2020, WHO (World Health Organisation) declared the outbreak to be a Public Health Emergency of International Concern, the highest level of alarm (WHO World Health Organization 2020a).

One of the many challenges of the COVID-19 pandemic is the estimation of the fraction of infected people with mild or no symptoms that escape testing and tracing, as it has been estimated that asymptomatic persons account for approximately 40%–45% of SARS-CoV-2 infections and can transmit the virus to others for an extended period (Oran and Topol 2021). Therefore, the greater the underdetection (or underreporting) rate is, the more difficult it is to understand the real dynamics of the epidemic. This can affect both government actions and individual behaviour, reducing the adherence of people to containment measures, for instance, if the epidemic is perceived as less severe than it is described. There are two main causes

of underdetection, namely the specificity of the SARS-CoV-2 virus inducing asymptomatic or mildly symptomatic cases which are hardly monitored or detected in large populations, and the emergency preparedness of national health systems in making available, and carrying out, the necessary amount of tests, as recommended by WHO (2020b), to both control the spread of the pandemic and determine the level of community transmission. Moreover, it has been estimated that, at the beginning of the epidemic, even the number of symptomatic cases was largely under-ascertained (Russell et al. 2020; Pullano et al. 2021). In an epidemiological approach, seroprevalence studies, where the host immune response to SARS-CoV-2 infection is measured in blood samples as a proxy for previous infections, have been implemented worldwide to enable estimates of the true extent of infection (Byambasuren et al. 2021). However, these studies are time- and resource-intensive and the resulting estimates can be affected by a high, often overlooked, uncertainty (Brown and Walensky 2020). In addition to seroprevalence studies, underreporting assessments have been carried out, for instance, from the number of reported deaths (Flaxman et al. 2020; Phipps et al. 2020) comparing the case fatality rate (CFR) to an estimate of the key parameter, infection fatality rate (IFR) (Sánchez-Romero et al. 2021); from exported cases and air travel volume (Tuite et al. 2020);

✉ Antonella Bodini
antonella.bodini@cnr.it

¹ CNR IMATI “E. Magenes”, Milano, Italy

from epidemiological and population data using deterministic models (Krantz and Rao 2020); by post-stratification sampling techniques (Bassi et al. 2021) or by also using information on testing and countries characteristics (Kuster and Overgaard 2021). Depending on different approaches, countries and analysis period considered, the estimated fraction of underdetected infections can vary remarkably (Noh and Danuser 2021). Underdetection, however, should not only be estimated as a fixed fraction, but its dynamic nature should also be considered in any modelling approach aiming at understanding the dynamics of the epidemic and forecasting its short-term development, since the estimation of the true number of infected (symptomatic, mild and asymptomatic cases) is fundamental to decide the implementation or lifting of containment measures and their subsequent evaluation.

In this work we propose to couple a stochastic compartmental model for the description of the dynamics of the transmission of SARS-CoV-2 with an observation equation that relates the actual numbers of infected and removed to the observed ones. The proposed approach is illustrated on the Italian data.

Since the onset of the pandemic, the popular SIR model, introduced about 10 years after the 1918 influenza pandemic (Kermack and McKendrick 1927) has been widely applied. A recent systematic review found that about 46% of the studies resulting from a search of the keywords related to SARS-CoV-2, and its modelling and prediction over the time period from 1st January 2020 to 30th November 2020, were based on a simple compartmental model, namely SIR and SEIR, both as deterministic or stochastic models (Gnanvi et al. 2021). Moreover, as a parsimonious model able to allowing measurement and forecast of the impact of non-pharmacological interventions such as social distancing, the SIR model still maintains a primary role in the analysis of the early phase of COVID-19 outbreak (Bertozzi et al. 2020; Carletti et al. 2020). Then, we considered a simple SIR model to exemplify the proposed approach to deal with underdetection in a compartmental model. A few other approaches have been found in literature. In Calafiore et al. (2020) the initial number of susceptible individuals, as well as the proportionality factor relating the detected number of positives with the actual (and unknown) number of infected individuals, were included among the parameters of a SIRD model. A Poisson model was used in Hong and Li (2020) to link the reported numbers of infected and removed to the true numbers of infected and removed as described by a time-varying SIR. The SIR model has also been extended by introducing several different compartments (Wang et al. 2020) in an attempt to capture the whole complexity of the pandemic. Some extended SIR models, in particular, include the compartment of

asymptomatic or some other distinction between detected and undetected infected (Giordano et al. 2020; Deo and Grover 2021; Gatto et al. 2020; Liu et al. 2021) to take into account underreporting. However, in previous work (Bilge et al. 2015) it was shown that, while the parameters of the SIR model can be uniquely determined from the temporal evolution of the normalised curve of removed individuals, the same is not true for more complex models. Thus, the SEIR and other models should not be used in the absence of additional information that could be obtained from clinical studies. The lack of clinical information had a significant impact on the early modeling of Italian data, given that Italy was the third country in the world and the first in the Western world to incur the pandemic (Berardi et al. 2020). Moreover, in Italy as in other countries, at the beginning of the epidemic underdetection was very high due to limited testing capacity and inadequate availability of both personal protective equipment and ventilators that forced the Italian government to restrict testing to people with severe cases and priority risk groups (Bosa et al. 2021).

The article is organised as follows. In Sect. 2 we introduce the stochastic SIR model, and we propose to model the fraction of reported cases as a random variable with a beta distribution and suggest a way to parameterise the beta distribution that relies on the infection fatality rate and its crude estimate, the case fatality rate. In Sect. 3 we suggest a method to compare the simulated and the observed epidemic dynamics using the model estimated via the Rao-Blackwellised particle filter (RBPF) algorithm reported in “Appendix 1” (Doucet et al. 2000). We also present a way to obtain short-term forecast on the dynamics. In Sect. 4, we apply the filtering algorithm to a simulated dataset to assess its behaviour and to introduce notation and terminology to be used later. We also make a few considerations on the problems arising in parameter estimation. In Sect. 5 we apply our method to the Italian data of both the first and the second infection wave, obtaining a good fit, along with a forecast that could be valid only in the short-term. We also consider the problem of assigning the correct observation error distribution using available information on the infection fatality rate and compare the simulated dynamics to the result of a sample serological survey carried out by Istat (Italian national statistical office) and the Italian Ministry of Health between May and July 2020. This comparison indicates that our model, when properly calibrated, provides a realistic assessment of the state of the epidemic. A section with some final remarks concludes the article.

2 Stochastic SIR model with underdetection

Consider a population and denote by S_t the fraction of susceptible individuals at time t , by I_t the fraction of infected individuals and by R_t the fraction of removed individuals (survivors and dead). We suppose that the population is closed, then $S_t + I_t + R_t = 1$ for every time t . The deterministic SIR model can be written

$$\begin{cases} \frac{dS_t}{dt} = -\beta I_t S_t \\ \frac{dI_t}{dt} = \beta I_t S_t - \gamma I_t \\ \frac{dR_t}{dt} = \gamma I_t \end{cases} \quad (1)$$

where β is the disease transmission rate, that is, the fraction of all contacts, between infected and susceptible people, that become infectious per unit of time and per individual in the population, and γ is the removal rate. The parameters β and γ allow us to approximate the basic reproduction number R_0 (or ratio, also called basic reproductive number or ratio) that can be thought of as the expected number of infected people generated by an infected individual in a population where all individuals are susceptible to infection. Despite its conceptual simplicity, R_0 is usually estimated with complex mathematical models developed using various sets of assumptions (Delamater et al. 2019). In the above SIR model it holds

$$R_0 = \frac{\beta}{\gamma},$$

where the parameters β and γ are unknown and have to be estimated. We suppose that these parameters are subject to uncertainty and change in time as follows

$$\beta_t = \beta_0 + \sigma w_t^{(1)} \quad \gamma_t = \gamma_0 + \eta w_t^{(2)} \quad (2)$$

with $w_t^{(1)}$ and $w_t^{(2)}$ independent Wiener noises. That is, β_t is supposed normally distributed with mean β_0 and variance $\sigma^2 t$ and γ_t is normally distributed with mean γ_0 and variance $\eta^2 t$. For alternative ways to introduce stochasticity, see Ganyani et al. (2021). The parameters σ and η measuring the noise intensity are assumed known and sufficiently small to obtain positive β_t and γ_t with probability approximately equal to one. Substituting the expression (2) for β and γ in system (1), we obtain the following stochastic SIR model:

$$\begin{cases} dS_t = -\beta_0 I_t S_t dt - \sigma I_t S_t dw_t^{(1)} \\ dI_t = (\beta_0 I_t S_t - \gamma_0 I_t) dt + \sigma I_t S_t dw_t^{(1)} - \eta I_t dw_t^{(2)} \\ dR_t = \gamma_0 I_t dt + \eta I_t dw_t^{(2)} \end{cases} \quad (3)$$

The introduction of noise in the parameters β and γ no

longer grants the condition $S_t + I_t + R_t = 1$. We can enforce it by replacing S_t by $1 - I_t - R_t$ in the second and third equations and removing the first equation to obtain the reduced system

$$\begin{cases} dI_t = (\beta_0 I_t (1 - I_t - R_t) - \gamma_0 I_t) dt + \sigma I_t (1 - I_t - R_t) dw_t^{(1)} - \eta I_t dw_t^{(2)} \\ dR_t = \gamma_0 I_t dt + \eta I_t dw_t^{(2)} \end{cases} \quad (4)$$

Denoting by $X_t = (I_t, R_t)^T$ the *state* vector, by $W_t = (w_t^{(1)}, w_t^{(2)})^T$ the vector of independent Wiener processes and by $\theta_0 = (\beta_0, \gamma_0)^T$ the parameter vector, we can rewrite the system (4) in vectorial form:

$$dX_t = h(X_t)\theta_0 dt + g(X_t)dW_t \quad (5)$$

where

$$\begin{aligned} h(X_t) &= \begin{bmatrix} I_t(1 - I_t - R_t) & -I_t \\ 0 & I_t \end{bmatrix}; \\ g(X_t) &= \begin{bmatrix} \sigma I_t(1 - I_t - R_t) & -\eta I_t \\ 0 & \eta I_t \end{bmatrix}. \end{aligned} \quad (6)$$

We call X_t the state of the system, which for COVID-19 is unobservable, and introduce Y_t to denote what can actually be observed, in accordance with the terminology derived from state-space modelling.

We suppose that each component of the observation vector Y_{t+1} is given by the product of the corresponding component of X_{t+1} and a random variable:

$$\begin{bmatrix} Y_{t,1} \\ Y_{t,2} \end{bmatrix} = \begin{bmatrix} U_{t,1} X_{t,1} \\ U_{t,2} X_{t,2} \end{bmatrix} \quad (7)$$

where $U_{t,1}$ and $U_{t,2}$ are independent and identically beta distributed random variables with shape parameters a and b . (In the following, by U , Y and X with no subscript we mean scalar random variables distributed as $U_{t,i}$, $Y_{t,i}$, and $X_{t,i}$, respectively, $i = 1, 2$). The observation error in SIR models has been considered in other frameworks using different formulations (see, for example, Stocks et al. (2021)). Finally, we assume that the initial distribution of θ_0 is Gaussian with mean μ_0 and covariance matrix Σ_0 .

2.1 Assigning parameters of the observation error distribution

The choice of a and b at the beginning of a new epidemic, in the absence of epidemiological information, is very critical and we propose to refer to the infection fatality ratio (IFR), the ratio of COVID-19 deaths to total infections of SARS-CoV-2, including asymptomatic and undiagnosed infections. The IFR is a fundamental quantity to estimate the severity of the epidemic, but it can only be

known at the end of the epidemic and after testing the entire population. An accurate estimation is therefore challenging in general and especially for COVID-19 (Mallapaty 2020), and the case fatality ratio (CFR), the ratio of COVID-19 deaths to confirmed cases, is usually considered as a rough estimate. By definition, the CFR is greater than the IFR. At any time t an estimate CFR_t of the CFR and an estimate IFR_t of the IFR are related by the simple relationship

$$CFR_t = \frac{D_t}{R_{obs,t}} = \frac{D_t}{R_t} \times \frac{R_t}{R_{obs,t}} = IFR_t \times \frac{R_t}{R_{obs,t}} \tag{8}$$

where D_t denotes all deaths by time t and $R_{obs} = Y_2 = U_2 X_2$ (Ghani et al. 2005).

Since by (8)

$$R_{obs,t} = \frac{IFR_t}{CFR_t} \times R_t, \tag{9}$$

the ratio IFR_t/CFR_t can be regarded as the underreporting factor that we have modelled as the U beta random variable introduced earlier.

Given estimates IFR_t as $t = 1, \dots, T$ and the corresponding observed sequence of CFR_t , we would obtain a sample $u_1 = IFR_1/CFR_1, \dots, u_T = IFR_T/CFR_T$ and an estimate of a and b by any established method. Using the method of moments, for example, and considering an estimate IFR of the IFR to substitute IFR_t , we would get

$$\begin{aligned} \hat{a} &= \bar{u} \left(\frac{\bar{u}(1-\bar{u})}{s_u^2} - 1 \right) \\ &= \bar{u}_{1/CFR} \left\{ \frac{\bar{u}_{1/CFR}(1-IFR\bar{u}_{1/CFR})}{s_{1/CFR}^2} - IFR \right\} \\ \hat{b} &= (1-\bar{u}) \left(\frac{\bar{u}(1-\bar{u})}{s_u^2} - 1 \right) \\ &= \frac{1-IFR\bar{u}_{1/CFR}}{IFR} \\ &\quad \left\{ \frac{\bar{u}_{1/CFR}(1-IFR\bar{u}_{1/CFR})}{s_{1/CFR}^2} - IFR \right\} \end{aligned} \tag{10}$$

where \bar{u} and s_u^2 are the sample mean and variance of (u_1, \dots, u_T) and $\bar{u}_{1/CFR}$ and $s_{1/CFR}^2$ are the sample mean and variance of $1/CFR_1, \dots, 1/CFR_T$. These equations show that the choice of IFR affects both parameters.

The fatality ratio approach has the advantage that the IFR is a pure number and information on its value can be gathered from different populations. Then, in practice, we may estimate the sample mean and variance of $1/CFR_1, \dots, 1/CFR_T$ from the observed fatality and removal data, and assign \hat{a} and \hat{b} for a selected IFR . If a range of values is available for the IFR from another source, such as a confidence or a credibility interval, we

may repeat the analysis with the IFR varying within the interval and evaluate the sensitivity of the results.

3 Parameter estimation, filtering, forecasting, and goodness-of-fit

To estimate the parameter θ_0 we propose a Rao-Blackwellised particle filter (RBPf) algorithm based on the Euler discretisation of the stochastic system (5):

$$X_{t+1} = X_t + h(X_t)\theta_0\Delta t + g(X_t)\Delta W_t, \quad t = 0, 1, 2, \dots \tag{11}$$

where we also use t for discrete time to save notation. The RBPf algorithm is described in ‘‘Appendix 1’’. This algorithm allows us to jointly calculate, at each time step, the estimated parameter and the state of the system using a noisy observation of the state as input.

To visualise how well the SIR model fits the observations, we need a way to compare the simulated trajectories with the observed data. We have supposed that the observations are smaller than the true value of the state X ; therefore we have to scale them by a factor that makes them comparable to the filtered state. A scaling factor is suggested by constructing a prediction interval of the state X at each observation time. Note that from (7) the random variable Y/X is a pivotal quantity with beta distribution and we may state that

$$\begin{aligned} 1-q &= P\left(u_{\frac{q}{2}} \leq \frac{Y}{X} \leq u_{1-\frac{q}{2}}\right) \\ &= P\left(\frac{Y}{u_{1-\frac{q}{2}}} \leq X \leq \frac{Y}{u_{\frac{q}{2}}}\right) \end{aligned} \tag{12}$$

where $u_{\frac{q}{2}}$ and $u_{1-\frac{q}{2}}$ are the $\frac{q}{2}$ and the $1-\frac{q}{2}$ percentiles of the beta distribution of U . Then, the corresponding prediction interval for X , after observing y , is

$$\left(\frac{y}{u_{1-\frac{q}{2}}}, \frac{y}{u_{\frac{q}{2}}} \right) \tag{13}$$

and a natural scaling factor for a point prediction of X is the median of U , yielding that the simulated trajectories of the model can be compared to the observed data by scaling them as $y/u_{0.5}$. The feature of (13) is that it does not depend on the SIR modelling assumption, but only on the observation error assumption, and, therefore, it offers a way to see how well the SIR dynamic follows the (transformed) data.

The model can also be used to predict the future behaviour of the epidemic. Let $y_{1:t}$ be the time series of observations up to time t ; for a fixed initial state x_0 , the RBPf algorithm provides a sample $x_{0:t}^{(i)}$, $i = 1, \dots, M$, to

approximate the posterior distribution of the state $p(x_{0:t}|y_{1:t})$. Furthermore, the conditional distributions of θ_0 given $x_{0:t}$, $p(\theta_0|x_{0:t})$, is Gaussian with mean $\mu_t = E(\theta_0|x_{0:t})$ and covariance matrix $\Sigma_t = Cov(\theta_0|x_{0:t})$. The RBPF algorithm produces a sample $(\mu_t^{(i)}, \Sigma_t^{(i)})$ of conditional mean vectors and covariance matrices given $x_{0:t}^{(i)}$. To forecast X_{t+k} given $y_{1:t}$, we aim at computing $E(X_{t+k}|y_{1:t})$. If we fix θ_0 and $x_{0:t}$, and run model (11) for k time steps, we obtain a value for X_{t+k} as a function $f_k(x_{0:t}, \theta_0, \xi)$, in which ξ indicates the sequence of increments ΔW_s , $s = t + 1, \dots, t + k$. Using $f_k(x_{0:t}, \theta_0, \xi)$, the conditional expectation is

$$E(X_{t+k}|y_{1:t}) = \int f_k(x_{0:t}, \theta_0, \xi) p(\xi) p(\theta_0|x_{0:t}) p(x_{0:t}|y_{1:t}) d\xi d\theta_0 dx_{0:t} \tag{14}$$

where conditional independence of θ_0 on $y_{1:t}$ given $x_{0:t}$ allows for substitution of $p(\theta_0|x_{0:t}, y_{1:t})$ by $p(\theta_0|x_{0:t})$. Then, if for each i we draw $\theta_0^{(i)}$ from $p(\theta_0|x_{0:t}^{(i)})$ and $\xi^{(i)}$ from the distribution of the Wiener process increment, the predictive expectation of X_{t+k} is approximated by

$$E(X_{t+k}|y_{1:t}) \simeq \frac{1}{M} \sum_{i=1}^M f_k(x_{0:t}^{(i)}, \theta_0^{(i)}, \xi^{(i)}) . \tag{15}$$

4 Model assessment with synthetic data

To check the convergence of the method and describe how to apply the estimation method we fix a parameter value and simulate the observations (or *data*).

We start from an initial condition of 1% infected and 0.1% removed. We simulate data for the parameters $\beta_0 = 0.3$ and $\gamma_0 = 0.1$. The parameters σ and η in (2) are 0.03 and 0.01, respectively.

To represent the initial phase of the epidemic, we run model (4) to generate 67 daily step states (circles in Fig. 1). Then, we use Eq. (7) with $a = 10$ and $b = 40$, meaning $u_{0.5} = 19.6\%$, for the beta distribution of the observation error obtaining the observations (asterisks in Fig. 1). The first 60 data will be used for the estimation procedure and the other 7 to check the goodness of the forecast.

We apply the RBPF algorithm described in ‘‘Appendix 1’’ with 200,000 particles and with a time step of 1/24 day. The choice of considering hourly observations while having daily ones is motivated by the need to improve the precision of the algorithm. For this purpose we imputed new hourly observations by linear interpolation between two consecutive daily observations (asterisks in Fig. 1). The imputed observations are no longer independently distributed conditionally on the states; however this

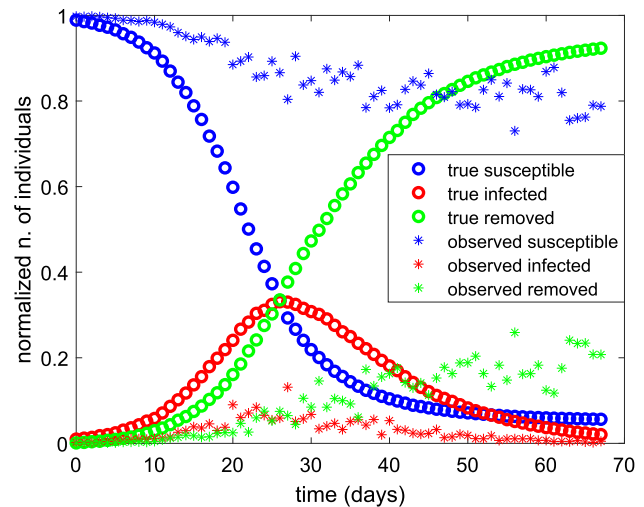


Fig. 1 True states (circles) obtained from Eq. (4) and observations (asterisks) obtained from true states applying formula (7), where $U_{t,1}$ and $U_{t,2}$ have a beta distribution with parameters $a = 10$ and $b = 40$, meaning $u_{0.5} = 19.6$

approximate procedure keeps the effective sample size of the RBPF algorithm at large values with no appreciable difference in results.

The initial guess for the parameter $\theta_0 = (\beta_0, \gamma_0)^T$ is $\mu_0 = (0.5, 0.5)^T$ and the prior covariance matrix $\Sigma_0 = diag(0.05, 0.02)$. The mean trajectories of infected and removed people over all the particles obtained by running the RBPF algorithm are represented with solid lines in the left panel of Fig. 2 where the circles represent simulated states before the introduction of the observation error. The susceptible individuals are obtained as $S_t = 1 - I_t - R_t$ and then the goodness of fit is a consequence of the fit for the other two compartments.

We denote the filtered or forecasted states as \hat{I}_t and \hat{R}_t , where the value of t determines whether we are filtering or forecasting, that is, if our observation period ends at time s , then \hat{I}_t and \hat{R}_t are forecasts when $t > s$; otherwise they are filtered states. From the RBPF we get the filtered states and, using (15), we get a forecast of the dynamics. \hat{I}_t and \hat{R}_t are compared to the true states in the left panel of Fig. 2, where we see that both the fit up to day 60 and the forecast on days 61–67 are satisfactory. In particular, the forecast well represents the state trend.

In real cases the true states (circles in Fig. 1) are unobserved, and we can only compare the filtered states to the observations, taking into account underdetection. Therefore, starting from the observations (asterisks in Fig. 1), we compute the daily prediction intervals for infected I_t and removed R_t , as in (13) with $q = 0.025$. In the right panel of Fig. 2 the prediction intervals are represented by vertical lines, while the thin solid lines represent the ratio between the observations and the median of

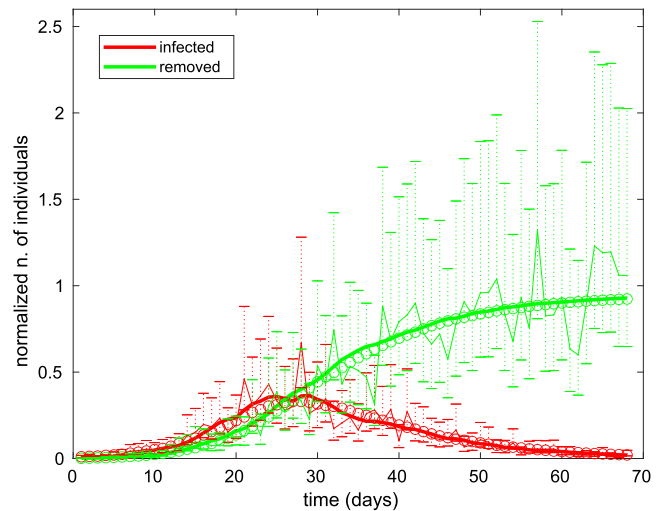
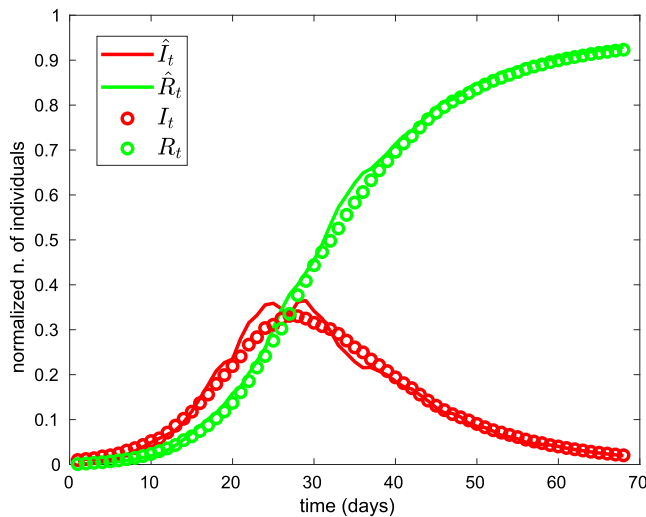


Fig. 2 Application of the RBPF algorithm. Left panel: trajectories of infected (red line) and removed (green line) compared with the true state (circles). The dynamics up to day 60 are the filtered states, while the dynamics from day 61 to day 67 are forecasts. Right panel: true states (circles), filtered state and forecast (thick lines) and adjusted

observations (thin line) with 95% prediction intervals (13). The thick lines up to day 60 are the filtered states, while those from day 61 to day 67 are forecasts. The thin lines are the observations divided by $u_{0.5}$, the median of the observation error distribution

the distribution of the observation error (which we may call the adjusted observations). The width of the prediction intervals reflects the dispersion of the observation error distribution for which $q_{0.025} = 0.10$ and $q_{0.975} = 0.32$. The true states (circles in Fig. 2) are inside the intervals and cross the adjusted observations (thin lines), then, if the adjusted observations and the filtered state (thick lines) agree with each other, this is a necessary condition for the filtered state to follow the unknown true state.

We denote the estimates of β_0 and γ_0 with information up to time t by $\hat{\beta}_t$ and $\hat{\gamma}_t$, see (21). Their time behaviour is shown in the top panel of Fig. 3. The behaviour of $\hat{R}_0(t)$, the estimated basic reproduction number (22), is displayed in the bottom panel of Fig. 3 and it can be seen that it well approaches the true R_0 .

Before considering the application to a real case, we would like to remark two aspects regarding the parameter estimation in stochastic systems: sample variability and identifiability. When considering a stochastic SIR model, the filtered states and the estimated parameters are affected by the variability in the data generated from system (4). As a consequence, for some datasets, the estimated parameters, although giving a good fit of the dynamics, can be far from the values used for generating the observations, keeping however the ratio $\hat{\beta}_t/\hat{\gamma}_t$ close to the true value of R_0 . The second aspects is the identifiability. The stochastic SIR models considered here are structurally identifiable (Bellman and Åström 1970; Piazzola et al. 2021), but the parameters can be practically non identifiable, that is there might be two different pairs of parameters giving a satisfactory fit of the dynamics. It is then fundamental to

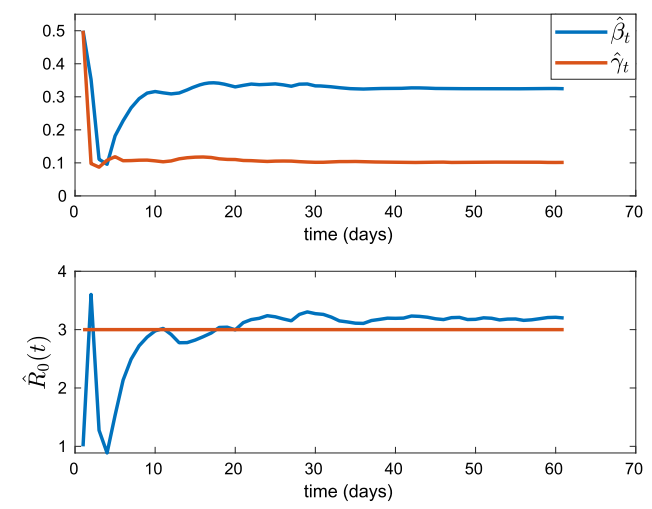


Fig. 3 Top panel: behaviour of $\hat{\beta}_t$ (blue) and $\hat{\gamma}_t$ (red) obtained from (21). Bottom panel: behaviour of $\hat{R}_0(t)$ obtained from (22) (blue line) and value of true $R_0 = \beta_0/\gamma_0 = 0.3/0.1$ (red line)

suitably choose the beta distribution of the observation error in the collection of infected and removed people to avoid practical nonidentifiability. Assessment of sample variability and identifiability is carefully discussed in “Appendix 2”.

5 The Italian data

In this section we apply the proposed approach to analyse the data of the first wave of COVID-19 in Italy, and we show how our analysis can capture the start of the second

wave of the epidemic. Data were collected by Protezione Civile (Civil Protection Department) from 24th February 2020 (Morettini et al. 2020). We consider the data for the entire Italy up to 26th November 2020. Available data are the number of infected, dead, and recovered individuals. Removed people can be obtained by adding up dead and recovered people. In Italy, all deaths of people infected with SARS-CoV-2 have been classified as due to COVID-19 (Riccardo et al. 2020). The infected and removed individuals in Italy from 24th February to 26th November are represented in Fig. 4. The total resident population as of 31st December 2019 is 60,244,639 people, as certified by Istat.

To deal with underdetection we consider observations as generated by (7), where we recall that $U_{t,1}$ and $U_{t,2}$ are independently beta distributed with common shape parameters a and b .

As we can see from Fig. 4, the first wave officially began on 24th February and lasted until mid-summer, when the number of infected people started to rise again, as in the rest of Europe (Bontempi 2020). The second wave is distinguished from the first also by the increased test capacity (https://www.covid19healthsystem.org/countries/italy/livn_ghit.aspx?Section=1.5%20Testing&Type=Section, last access 17th July 2021). Hence we consider the two waves as different models, with respect to both the SIR parameters and the observation error distribution parameters and we conventionally set the start of the second wave on 1st August.

For Italy, we may use an indirect method to indicate a plausible value of the IFR taking advantage of a seroprevalence survey targeting IgG antibodies conducted from May to July by Istat, the Italian national statistical office,

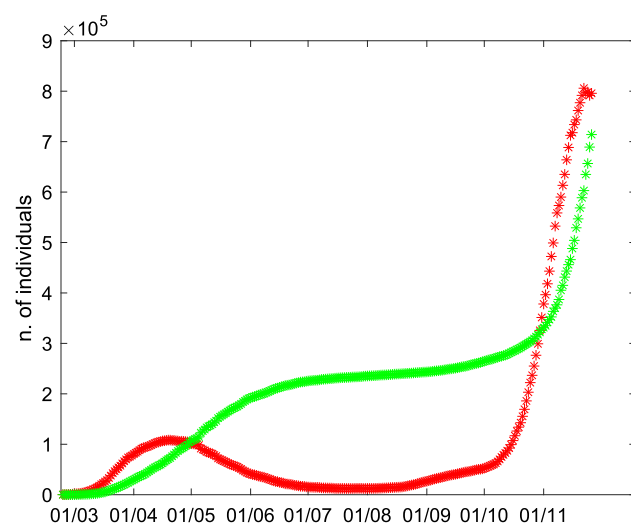


Fig. 4 Infected (red asterisks) and removed (green asterisks) in Italy from 24th February (time 0) to 26th November 2020. Data from Protezione Civile

and the Italian Ministry of Health. Preliminary results obtained from 64,660 people were presented in early August (Istituto Nazionale di Statistica 2020). According to them, almost 1.5 million people in Italy or 2.5% of the population had developed coronavirus antibodies, a figure six times larger than official numbers reported. In short, the idea is to compare the 2.5% figure of people who developed antibodies to the healed people (who have antibodies) estimated from the filtered state \hat{R}_t in an appropriate time interval. The infected compartment may also contain seropositive individuals; however, the fraction of people in this compartment had become small when Istat’s survey started, so we consider only the removed compartment. The reasoning behind this comparison is that if the assumed *IFR* is correct, then the observation error distribution derived from (10) is correct and the filtered states are realistic and they should be in agreement with the Istat survey result.

To be more specific, let $R_t = H_t + D_t$, where H_t and D_t are the fractions (over the population) of healed and dead people by time t , respectively. Healed people can be seronegative if IgG antibodies are no longer in their system, but we can safely assume that a person enters the healed record soon so s/he can be considered as seropositive when they do. Now, H_t includes all healed individuals since the start of the epidemic, hence a fraction of H_t can be seronegative, depending on the duration d of seropositivity. Hence we should compare 2.5% to $H_t - H_{t-d}$, where $H_{t-d} = 0$ if $t - d < 1$. The true values of H_t are unknown. We may recover them from \hat{R}_t and the available data on the fraction of deaths as $\hat{H}_t = \hat{R}_t - D_t/u_{0.5}$. Since Istat’s survey was carried out between 25th May and 15th July 2020, we compare 2.5% to

$$\bar{H} = \frac{1}{52} \sum_{t=25\text{May}}^{15\text{July}} (\hat{H}_t - \hat{H}_{t-d}) . \tag{16}$$

The duration of IgG antibodies is still today largely debated, (Röltgen et al. 2020; Scozzari et al. 2021) and here we consider three months as a plausible value of d (Duysburgh et al. 2021). This procedure rests on several assumptions and we only regard it as a way to check for gross deviations of our model from reality.

5.1 State and parameter estimation

We run the RBPF algorithm with 20,000 particles, time discretisation step of 1/24 day as done for the synthetic data, starting from the first day with at least 100 removed (1st March). The initial values are $(\beta_0, \gamma_0)^T = \mu_0 = (0.3, 0.1)^T$ and $\Sigma_0 = \text{diag}(0.002, 0.001)$. Moreover $\sigma = 0.03$ and $\eta = 0.01$.

The Centre for Evidence-Based Medicine (CEBM) at the University of Oxford bases its timely updates of IFR point estimate on a continuously evolving meta-analysis based on CFR data. The IFR point estimate is then obtained by halving the lower bound of the 95% prediction interval of the CFR and the current estimate sets the IFR at 0.51% (Oke and Heneghan 2020). In Brazeau et al. (2020) low and high income countries were discussed separately; in a typical high income country, with a higher concentration of elderly individuals, an overall IFR of 1.15% (0.78–1.79 95% prediction interval) is estimated. An estimate of 1.3% has been obtained using data from the closed population of passengers in the Diamond Princess cruise ship (Russell et al. 2020). The meta-analysis carried out by Meyerowitz-Katz and Merone (2020) of published research data on COVID-19 infection fatality rates, with last search on 16/06/2020, has indicated a point estimate of IFR of 0.68% (0.53%–0.82%) with high heterogeneity, and suggested that in many populations the IFR would be $> 1\%$ if excess mortality was taken into account.

For the first wave, we then computed a and b from (10) for a range of IFR values from 0.1%–6%, where the minimum value is the lowest we found in the relevant literature, besides having been suggested as lower bound of IFR in Europe by CEBM. The maximum value is still inspired by CEBM and the considerations in Meyerowitz-Katz and Merone (2020). Indeed, in Italy an estimated initial CFR of about 11–19% has been reported (De Natale et al. 2020; Tosi et al. 2020; Riccardo et al. 2020). This suggested to consider IFR=6% as a possible maximum initial value, accounting for the lack of knowledge at the beginning of the first wave. Moving too far from the highest value gives a large discrepancy between Istat's 2.5% estimated seropositivity in the population and (16). Then we present here results for $IFR = 4.5\%$ (on 31st March 2020), for which $\bar{H} = 2.4\%$. The parameters of the corresponding beta density are $a = 11.9$ and $b = 93.17$, meaning a median underdetection value $u_{0.5} = 11.1\%$ (95% range: 6–18%). This result is in line with Gatto et al. (2020), and is more optimistic than the estimated 4% in Kuster and Overgaard (2021) while noticeably lower than the estimated 40% in Liu et al. (2021). These differences are also due to the time interval considered. The initial condition for I_t (R_t), for each trajectory, is given by the normalised number of infected (removed) people collected by Protezione Civile on 1st March divided by the median of this beta distribution to take into account the underdetection on 1st March data.

To account for government actions, we split the first wave into subintervals considering the DPCMs¹ with the greatest impact on social organisation allowing for 10 days

for the DPCM to have an effect on the epidemic (that is, change points are the DPCM dates plus 10 days). In particular, we considered the following DPCM dates: 11th March, 22nd March, 26th April and 3rd June, so that the change points are on 21st March, 1st April, 3rd May and 13th June.

For each time interval, except the first, we used the filtered state \hat{x}_t (20) at the end of the previous interval as initial state and the values of $\hat{\beta}_t$ and $\hat{\gamma}_t$ at the end of the previous interval as starting parameters. Then the discontinuity in the update was determined only by the initial covariance matrix. From a preliminary study on a single interval we found small values for the matrix Σ . Then we took $\Sigma_0 = \text{diag}(0.002, 0.001)$ for all the time intervals. This choice also allowed us to avoid big jumps in the trajectories of the parameters at the change points. The dynamics of the five different SIR models are represented as a whole dynamics in Fig. 5. The filtered states of infected and removed individuals are represented with thick lines in Fig. 5 where also the prediction intervals for both infected and removed are reported. The prediction intervals are computed from (13) and the thin lines in Fig. 5 represent the adjusted observations obtained considering the ratio between the observations and the median $u_{0.5} = 11.1\%$ of the observation error distribution.

The trajectories $\hat{\beta}_t$, $\hat{\gamma}_t$ and $\hat{R}_0(t)$ show jumps at the change points (Fig. 6), which are not very pronounced due to the choice of a small Σ_0 . After a few steps from each jump, the trajectories stabilise following a regular trend. The dynamics of infected individuals fits very well the observed infected divided by $u_{0.5} = 11.1\%$. At the beginning of the epidemic, $\hat{R}_0(t)$ is higher than 3. This reflects both the uncertainty in the Italian data and the recognised high value of R_0 describing the early temporal spread of SARS-CoV-2 at global and local levels, (Katul et al. 2020; Linka et al. 2020; Yu et al. 2021). Values around 3 at the end of March (around day 30) reflect the current estimates for that period, before the effects of the general stay-at-home recommendation (22nd March), reinforcing the national lockdown (11th March), are felt (Gatto et al. 2020; D'Arienzo and Coniglio 2020; Allietta et al. 2021). After day 66 (corresponding to 26th April), $\hat{\beta}_t$ is smaller than $\hat{\gamma}_t$, so $\hat{R}_0(t) < 1$. This result is in agreement with the effective reproduction number published for the first time by Istituto Superiore di Sanità (Italian National Institute of Health, ISS) on 30th April (Istituto Superiore Sanità 2020a): the effective reproduction numbers were reported for every Italian region (except for two because of bad quality data) and they were all smaller than one. It is worth pointing out that our simulation study showed that the method is capable of estimating the basic reproduction number with greater accuracy and precision than the infection and

¹ DPCM: Italian acronym for government decrees. For a summary of the DPCMs related to the COVID-19 emergency, see <http://www.governo.it/it/coronavirus-misure-del-governo>.

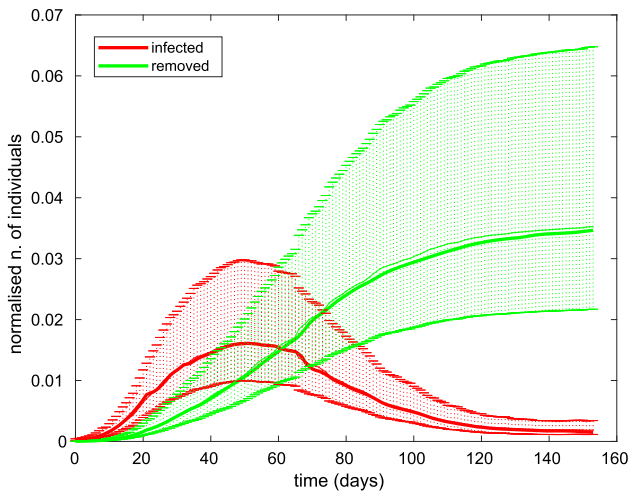


Fig. 5 Filtered states for infected (thick red line) and removed (thick green line) from five different SIRs in the intervals [0, 20], [20, 31], [31, 66], [66, 104], [104, 160]. Parameters of the beta observation error distribution are $a = 11.9$ and $b = 93.17$. The prediction intervals are computed from (13) with $q = 0.025$. The thin lines are the observed infected (red) and removed (green) divided by $u_{0.5} = 11.1\%$. Time 0 is 1st March 2020

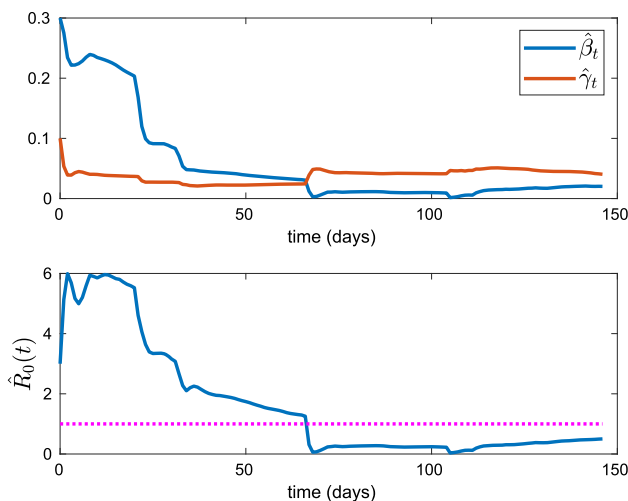


Fig. 6 Top panel: plots of $\hat{\beta}_t$ and $\hat{\gamma}_t$ from (21). Bottom panel: plot of $\hat{R}_0(t)$ from (22). Time 0 is 1st March 2020

removal rate parameters (see “Appendix 2.1”), which is relevant to public health decisions.

We then analysed the forecast of the infected and removed dynamics for the first wave, computed as in (15). We considered observations up to 20 days after the second or third change points, and we tried to forecast the dynamics for the seven following days (Fig. 7).

It can be observed from Fig. 7 that the forecast is satisfactory 20 days after the third change point, that is when the estimate of β_t has approximately reached a constant

value, and it is less precise when β_t has a decreasing trend (see Fig. 6).

We then focused on the second wave, starting on 1st August and we considered three change points 10 days after the following measures: on 1st September (partial opening of museums and stadiums, and increased occupancy in public transportation); on 21st October (curfew in Lombardy, the most affected region); on 3rd November (DPCM establishing red, orange and yellow scenarios to classify regions from highest to lowest risk and introducing tiered restrictions). We run the RBPF algorithm, with $\mu_0 = (0.05, 0.03)^T$ as initial value for $(\beta_0, \gamma_0)^T$ and $\Sigma_0 = \text{diag}(0.002, 0.001)$. Moreover, $\sigma = 0.03$ and $\eta = 0.01$. The state is formed by infected individuals and newly removed individuals since 1st August, that is, the difference between the removed at each time and the removed on 31st July.

For the second wave the underdetection error of infected and removed people is smaller, because of an increase in resources for taking swab tests. For this reason it is appropriate to recalculate the parameters of the beta observation error distribution from (10) only with data since 1st August. Unfortunately, we lacked a benchmark such as the serological survey during the first wave, and therefore we present the results obtained by considering four different *IFR* values: 1.15%, 1.3%, 1.5%, 1.75% according to the most recent studies. The beta densities obtained for these values (on 31st August 2020) are represented in Fig. 8 and compared with the beta density used for the first wave. We excluded smaller values of *IFR* because the corresponding observation error beta distributions were located on small values, like for the first wave.

It can be seen from Fig. 8 that both mean and variance of the beta distribution increase as the *IFR* increases. The corresponding dynamics are shown in Fig. 9. The normalised numbers of individuals decrease as the *IFR* increases, because the observations are divided by the median of the beta distribution which is increasing with the *IFR*. The width of the prediction intervals also decreases as the *IFR* increases. The filtered dynamics well reconstruct the behaviour of the adjusted data in all the cases. The estimated parameters $\hat{\beta}_t$ and $\hat{\gamma}_t$ are very similar for the different cases and only the parameters obtained using *IFR* = 1.3% are reported in Fig. 10 with the corresponding $\hat{R}_0(t)$.

We observed a large jump in $\hat{R}_0(t)$ on the date of the second change point, 10 days after the curfew in Lombardy, followed by a slow decrease, indicating that this measure did not produce the desired effect. Then it was followed by the measure of 3rd November that allowed $\hat{R}_0(t)$ to accelerate its decrease, approaching one, in

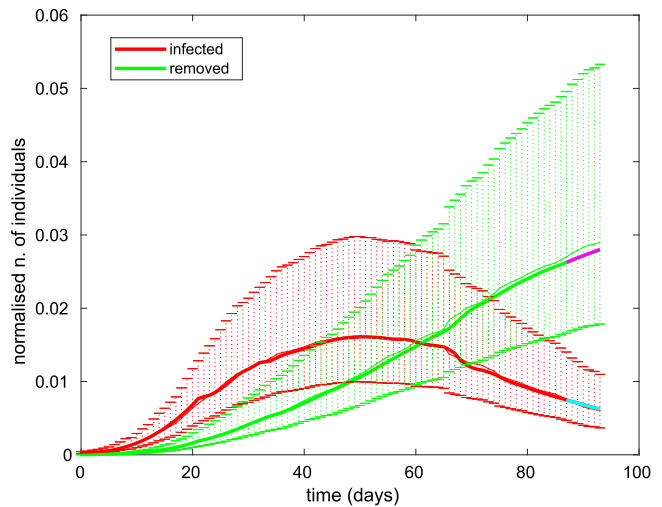
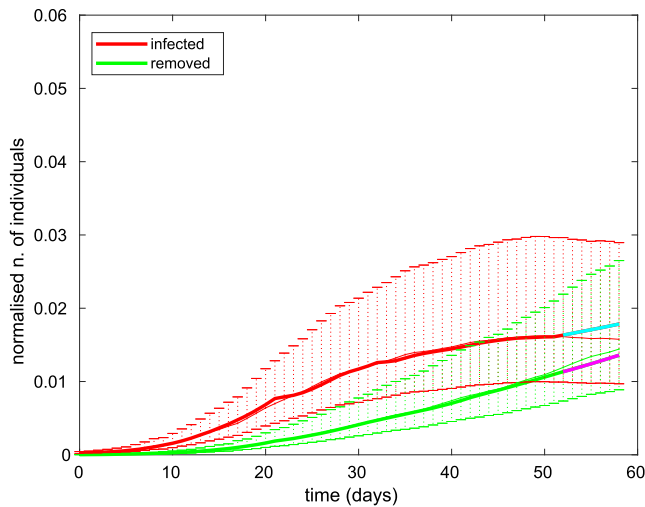


Fig. 7 Dynamics of infected and removed individuals with forecasts during an increasing phase (left panel) or a decreasing phase (right panel). The forecasts start 20 days after the second change point (left

panel) and 20 days after the third change point (right panel). Forecasts of infected and removed individuals are highlighted with different colours

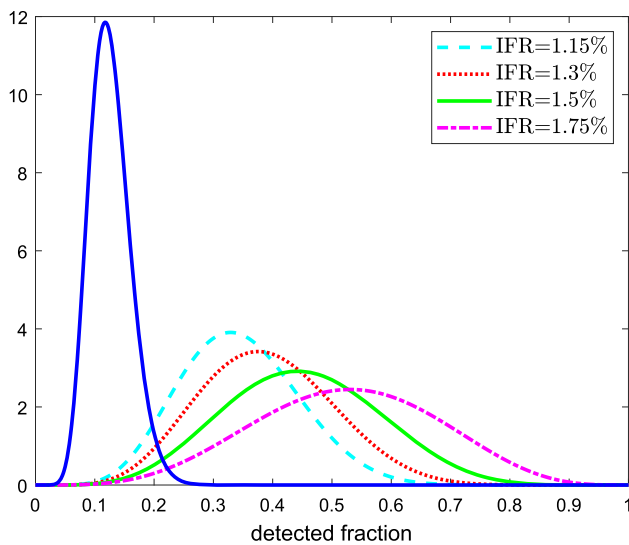


Fig. 8 Beta distributions of the observation error for different values of IFR: 1.15% (light blue dashed line), 1.3% (red dotted line), 1.5% (green continuous line), 1.75% (magenta dashed-dotted line). For comparison also the beta density used for the first wave is reported (blue continuous line)

agreement with what the ISS reported in its 25th November bulletin (Istituto Superiore Sanità 2020b).

6 Concluding remarks

In this work we have proposed to explicitly model the underdetection of SARS-CoV-2 infected and removed subjects by introducing a beta distribution for the observation error in a SIR model. In particular, a piecewise stochastic SIR model with change points has been fitted to

the COVID-19 data in Italy from 1st March to 26th November 2020, using the dates of the measures taken by the government to control the epidemic to define the change points. This strategy allowed us to estimate the actual dynamics of the epidemic by correcting the observed data. The stochastic SIR model, coupled with the RBPF algorithm to estimate the parameters, has improved the description of the dynamics that can be obtained using a piecewise deterministic SIR model with maximum likelihood estimation of parameters, as shown in “Appendix 3”. By particle filtering and parameter learning algorithm, the model can produce short-term predictions of the population in each compartment and continuously updated estimates of key quantities such as the basic reproduction number, on which decision makers can act. We have obtained a rather large basic reproduction number in the initial phase of the first wave, progressively decreasing in the following phases, in line with the current literature.

The adaptation of a simple SIR model to the Italian data at the beginning of the epidemic has been here supported by the lack of clinical information. However, even as more specific information emerged, the appropriateness of using simple models has not been entirely questioned. From about 9,000 laboratory-confirmed cases reported outside Hubei in mainland China from mid-January to mid-February 2020, a mean incubation period of 5.2 days (1.8–12.4) and an almost coinciding mean serial interval at 5.1 days (1.3–11.6) were estimated (Zhang et al. 2020). This means that the infection can occur in the presymptomatic phase. Moreover, an analysis of the first 6,000 laboratory confirmed cases in Italy showed that the viral load does not significantly differ with the type of symptoms (Cereda et al. 2020), so isolation of infected persons should

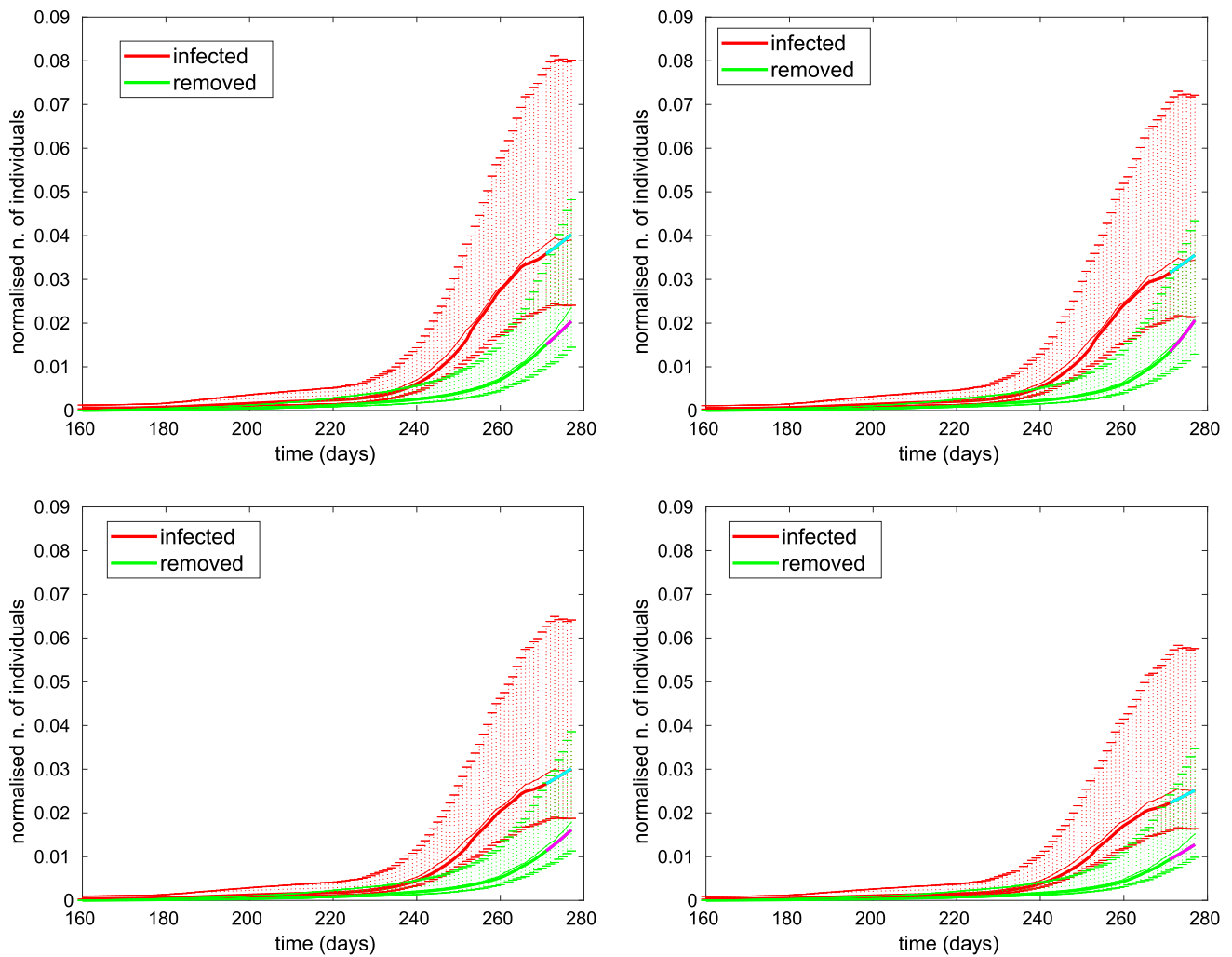


Fig. 9 Filtered states for infected (thick red line) and removed (thick green line) in the second wave for different IFR: 1.15% (top left), 1.3% (top right), 1.5% (bottom left), 1.75% (bottom right). The prediction intervals are computed from (13) with $q = 0.025$. The thin

lines are the observed infected (red) and removed (green) divided by $u_{0.5}$. Forecast of infected and removed individuals are highlighted with different colors. Time 160 is 1st August

be performed regardless of symptoms (Lee et al. 2020). All these results together suggest that the specification of many compartments among the infected may also not help clarify the dynamics.

The stochastic SIR model might be used to evaluate the effect of mitigation measures by extending the predictions to a horizon of several weeks beyond the date of the next government decree, as an answer to the question: “what would have been the mid-term evolution of the epidemic if this specific measure had not been taken?”. But, given the complexity of the phenomenon, which is only partially captured by the model, great caution is required in doing so. One possible way out is to enrich the model by adding compartments, but, as our model identifiability study has shown (“Appendix 2.2”), solid prior information or data relevant to the required additional parameters are needed to obtain meaningful results. Indeed, in our simulation study

we have obtained that the model describes reliable dynamics provided the parameters of the observational error, assumed as known, are correctly assigned (“Appendix 2.1”).

With respect to prior information, our approach to the selection of the observation error distribution depends on an estimate of the IFR. For the first wave we have chosen larger values than for the second wave (up to 6% against a maximum of 1.75%), which is in agreement with an ISS report (Fabiani et al. 2020), published after we finished our analysis, where the monthly standardised CFR has been calculated. When standardised with respect to the age and sex structure of the Italian population, the CFR is close to 9% in February-March 2020 and close to 4.5% in April. Then it falls to around 1% in June and July and increases to above 2% in October.

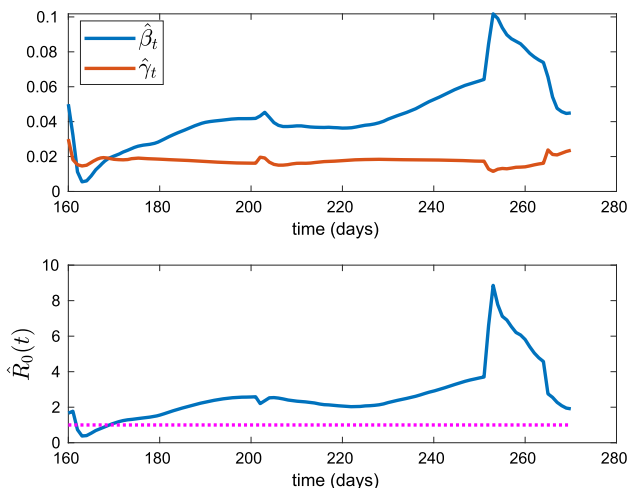


Fig. 10 Top panel: plots of $\hat{\beta}_t$ and $\hat{\gamma}_t$ from (21). Bottom panel: plot of $\hat{R}_0(t)$ from (22). The IFR is 1.3%. Time 0 is 1st August

Appendix 1: RBPf algorithm

To estimate the parameter $\theta_0 = (\beta_0, \gamma_0)^T$ and the state X_t we propose to apply a Rao-Blackwellised particle filter (RBPf) algorithm. We consider the Euler discretisation of the stochastic system (5) reported in Eq. (11). Since the system is linear in θ_0 , we can apply the Kalman filter. Suppose that $\theta_0 = (\beta_0, \gamma_0)^T$ has a Gaussian prior distribution with mean μ_0 and covariance matrix Σ_0 , then the distribution of θ_0 , given $x_{0:t+1} = (x_0, x_1, \dots, x_{t+1})$ after $t + 1$ time steps, as $t = 0, 1, 2, \dots$, is Gaussian with mean μ_{t+1} and covariance matrix Σ_{t+1} given by

$$\begin{cases} \mu_{t+1} = \mu_t + S_{t+1}^T [x_{t+1} - x_t - h(x_t)\mu_t \Delta t] \\ \Sigma_{t+1} = \Sigma_t - S_{t+1}^T h(x_t) \Sigma_t \Delta t \\ S_{t+1}^T = \Sigma_t h^T(x_t) \Delta t [h(x_t) \Sigma_t h^T(x_t) (\Delta t)^2 + g(x_t) g^T(x_t) \Delta t]^{-1} \end{cases} \tag{17}$$

The distribution of X_{t+1} given $x_{0:t}$ is Gaussian with mean B_{t+1} and covariance matrix G_{t+1} given by

$$\begin{cases} B_{t+1} = x_t + h(x_t)\mu_t \Delta t \\ G_{t+1} = h(x_t) \Sigma_t h^T(x_t) (\Delta t)^2 + g(x_t) g^T(x_t) \Delta t. \end{cases} \tag{18}$$

Recalling that the observations are obtained multiplying the state X_t for the beta-distributed observation error term, as defined in Eq. (7), the RBPf algorithm can be summarised as follows:

Step 1

- At time $t = 0$, we draw M initial values $x_0^{(i)}, i = 1, 2, \dots, M$ of X_0 from its prior distribution $\pi(x_0)$ or, alternatively, we put x_0 equal to the initial observation.

- We consider a Gaussian prior distribution $\mathcal{N}(\mu_0, \Sigma_0)$ for the parameter θ_0 , where μ_0 is a vector of initial parameters, and Σ_0 is a diagonal covariance matrix.
- To obtain candidate values of the state at importance sampling steps, we use the distribution implied by the state-transition Eq. (11) after marginalising it with respect to θ_0 . At step one, a value for $\tilde{x}_1^{(i)}$, conditional on $x_0^{(i)}$, is sampled from a Gaussian distribution with mean $B_1^{(i)}$ and covariance matrix $G_1^{(i)}$, for $i = 1, \dots, M$, given by (18) with $k = 0$.
- Denoting by $y_1 = (y_{1,1}, y_{1,2})$ the observation at time $k = 1$, we compute weights for each particle from the likelihood at $\tilde{x}_1^{(i)} = (\tilde{x}_{1,1}^{(i)}, \tilde{x}_{1,2}^{(i)})$

$$v_1^{(i)} = L(\tilde{x}_1^{(i)}; y_1) = p(y_{1,1} | \tilde{x}_{1,1}^{(i)}) \times p(y_{1,2} | \tilde{x}_{1,2}^{(i)})$$

where

$$p(y|x) = \frac{\left(\frac{y}{x}\right)^{a-1} \left(1 - \frac{y}{x}\right)^{b-1}}{B(a, b)} \frac{1}{x} I_{[0,x]}(y). \tag{19}$$

In order to resample the particles, we need to normalise the weights:

$$v_1^{(i)} = \frac{\tilde{v}_1^{(i)}}{\sum_{i=1}^M \tilde{v}_1^{(i)}}.$$

- We update the posterior distribution of θ_0 given $\{\tilde{x}_1^{(i)}, x_0^{(i)}\}$ by taking one step of the Kalman filter of Eq. (17), obtaining the new mean vector $\tilde{\mu}_1^{(i)}$ and covariance matrix $\tilde{\Sigma}_1^{(i)}$.
- We resample M particles from a discrete distribution with support $\left\{ \left(\tilde{x}_1^{(i)}, \tilde{\mu}_1^{(i)}, \tilde{\Sigma}_1^{(i)} \right) \right\}_{i=1, \dots, M}$ and corresponding probabilities $\left\{ v_1^{(i)} \right\}_{i=1, \dots, M}$. We denote by $\left\{ \left(x_1^{(i)}, \mu_1^{(i)}, \Sigma_1^{(i)} \right) \right\}_{i=1, \dots, M}$ the resampled particles.

At time $t \geq 1$, we assume the sample $\left\{ \left(x_t^{(i)}, \mu_t^{(i)}, \Sigma_t^{(i)} \right) \right\}_{i=1, \dots, M}$ is available.

Step $t + 1$

- For $i = 1, \dots, M$, sample candidate particles $\tilde{x}_{t+1}^{(i)}$ from a Gaussian distribution with mean $B_{t+1}^{(i)}$ and covariance matrix $G_{t+1}^{(i)}$, given by (18).
- Compute the weights $\tilde{v}_{t+1}^{(i)}$ for each particle as the product of two distributions with density (19). Normalise the weights:

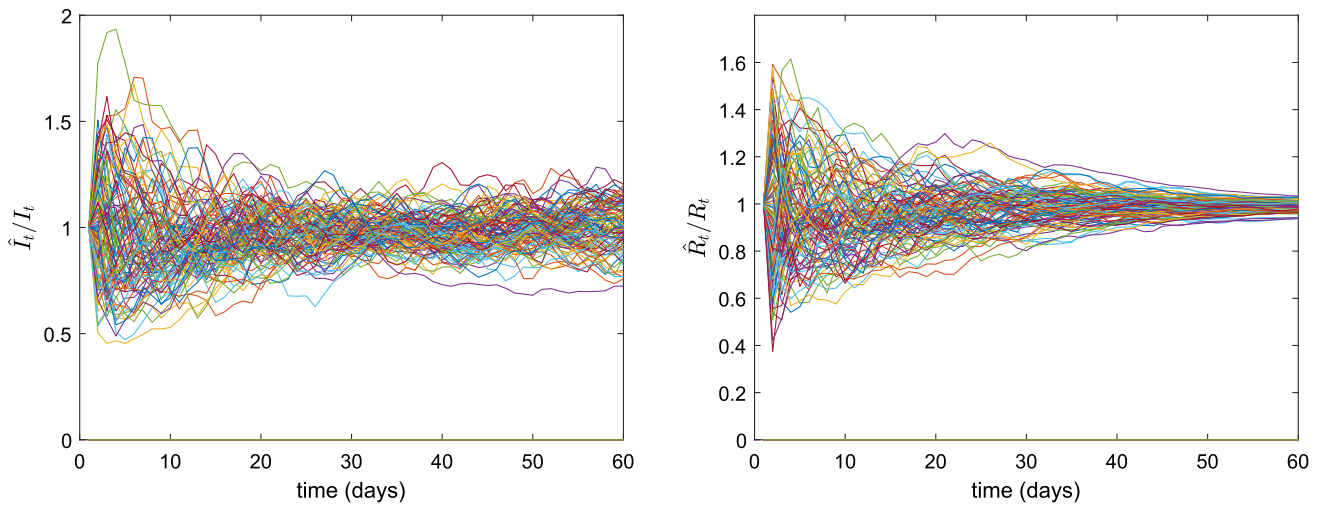


Fig. 11 Left panel: behaviour of \hat{I}_t/I_t for 100 different simulations. Right panel: behaviour of \hat{R}_t/R_t for 100 different simulations

$$v_{t+1}^{(i)} = \frac{\hat{v}_{t+1}^{(i)}}{\sum_{i=1}^M \hat{v}_{t+1}^{(i)}}.$$

- Update the posterior distribution of θ_0 given $x_{0:t+1}^{(i)}$, which is a Gaussian distribution with mean $\hat{\mu}_{t+1}^{(i)}$ and covariance matrix $\hat{\Sigma}_{t+1}^{(i)}$ given by Eq. (17).
- Resample M particles using the probabilities $\{v_{t+1}^{(i)}\}_{i=1, \dots, M}$ and denote the resampled particles by $\{(x_{t+1}^{(i)}, \mu_{t+1}^{(i)}, \Sigma_{t+1}^{(i)})\}_{i=1, \dots, M}$.

Particles $\{(x_t^{(i)}, \mu_t^{(i)}, \Sigma_t^{(i)})\}_{i=1, \dots, M}$ are a sample from the distribution of interest. In detail, the $x_t^{(i)}$'s are a sample from $p(x_t|y_{1:t})$ and, by keeping track of the resampling history, the entire sample $x_{0:t}^{(i)}, i = 1, \dots, M$ is potentially available, hence a sample from $p(x_{0:t}|y_{1:t})$. The mean of $x_t^{(i)}$ over the particles approximates $E(x_t|y_{1:t})$ and we call it the filtered state:

$$\hat{x}_t = \frac{1}{M} \sum_{i=1}^M x_t^{(i)}. \tag{20}$$

The $\mu_t^{(i)}$'s and $\Sigma_t^{(i)}$'s are a sample of conditional means and covariance matrices, that is, $E(\theta_0|x_{0:t}^{(i)})$ and $Cov(\theta_0|x_{0:t}^{(i)})$. Therefore, an estimate of $E(\theta_0|y_{1:t})$ is

$$(\hat{\beta}_t, \hat{\gamma}_t)^T = \hat{\theta}_t = \frac{1}{M} \sum_{i=1}^M \mu_t^{(i)} \tag{21}$$

and, by sampling M values from M Gaussian distributions $N(\mu_t^{(i)}, \Sigma_t^{(i)})$, $i = 1, \dots, M$, we produce a sample $(\theta_t^{(1)}, \dots, \theta_t^{(M)})$ from $p(\theta_0|y_{1:t})$.

The basic reproduction number is defined as $R_0 = \beta_0/\gamma_0$, therefore an estimate based on $y_{1:t}$ is $E(\beta_0/\gamma_0|y_{1:t})$, which is computed as

$$\hat{R}_0(t) = \frac{1}{M} \sum_{i=1}^M \frac{\beta_t^{(i)}}{\gamma_t^{(i)}} \tag{22}$$

where $(\beta_t^{(i)}, \gamma_t^{(i)})^T = \theta_t^{(i)}$. If the variances on the diagonal of $\Sigma_t^{(i)}$ are small, the additional sampling from the $N(\mu_t^{(i)}, \Sigma_t^{(i)})$ may be unnecessary and the following approximation might be used, corresponding to degenerate conditional distributions:

$$\hat{R}_0(t) = \frac{1}{M} \sum_{i=1}^M \frac{\mu_{1,t}^{(i)}}{\mu_{2,t}^{(i)}}. \tag{23}$$

$\hat{R}_0(t)$ can be regarded as our best estimate of the basic reproduction number in the light of the observed data.

Appendix 2: Numerical simulations

Appendix 2.1: Assessment of sample variability

We have analysed the variability of the filtered states and the estimated parameters due to the variability of the data generated from the system (4), in order to get an

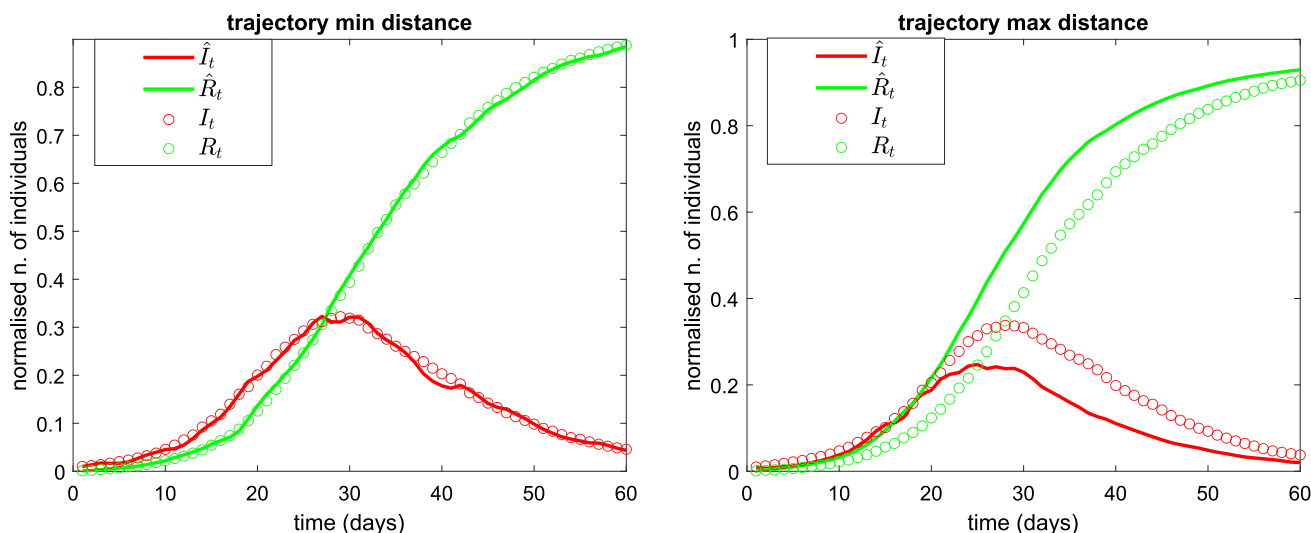


Fig. 12 Left panel: dynamics of filtered states \hat{I}_t and \hat{R}_t (continuous lines) in the case of minimum sum of root mean square error between filtered states and true states (circles). Right panel: dynamics of

filtered states \hat{I}_t and \hat{R}_t (continuous lines) in the case of maximum sum of root mean square error between filtered states and true states (circles)

impression of how far they can get from the true values, even if the true random underreporting error distribution is used. We have used the same parameters of Sect. 4, that is, $(\beta_0, \gamma_0) = (0.3, 0.1)$, $\sigma = 0.01$, $\eta = 0.03$, initial values $\mu_0 = (0.5, 0.5)^T$ and $\Sigma_0 = \text{diag}(0.05, 0.02)$ and initial state $X_0 = (1\%, 0.1\%)$.

We generated 500 dynamics from the system (4), from which we obtained 500 trajectories of observed infected and removed people. Then, we run the RBPF algorithm on each simulated dataset, with a time step of 1/24 day and for 200,000 particles. For every simulations we computed the trajectories \hat{I}_t/I_t and \hat{R}_t/R_t , where we recall that \hat{I}_t and \hat{R}_t are the estimated infected and removed individuals filtered by the RBPF algorithm, while I_t and R_t are the true states for the corresponding simulation. For a clear representation, in Fig. 11 we show one in every five trajectories of \hat{I}_t/I_t (left panel) and \hat{R}_t/R_t (right panel).

In both panels of Fig. 11, after a transient phase with larger dispersion, \hat{I}_t/I_t and \hat{R}_t/R_t end up fluctuating around 1, with a stable dispersion in the left panel and a decreasing dispersion in the right panel. Then we considered the sum of the root mean square error (RMSE) between \hat{I}_t and I_t and of the RMSE between \hat{R}_t and R_t for all the trajectories, as a measure of distance between the estimate and the truth. Among all the trajectories we represented the one with the smallest and the one with the largest distance in the left and in the right panels of Fig. 12, respectively. The latter picture shows that the fit may be very unsatisfactory.

Finally, we report the scatter plot of all the pairs $(\hat{\beta}_t, \hat{\gamma}_t)$ obtained in the 500 simulations (Fig. 13) at $t = 60$. We observe that they are dispersed around the true value of the

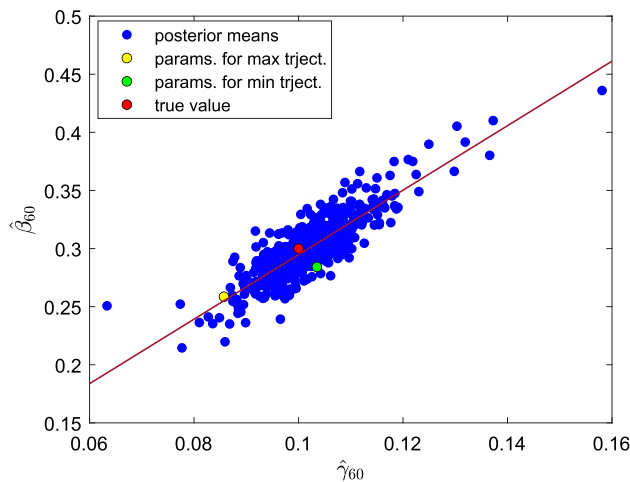


Fig. 13 Scatter plot of the parameters $(\hat{\gamma}_{60}, \hat{\beta}_{60})$ obtained for the different simulations. The red dot represents the true pair (0.1, 0.3). The green dot represents the pair of parameters relative to the filtered state trajectory with minimum distance from the true states (left panel of Fig. 12). The yellow dot represents the pair of parameters corresponding to the filtered state trajectory with maximum distance from the true states (right panel of Fig. 12). The line is the least squares fit of $\hat{\beta}_{60}$ against $\hat{\gamma}_{60}$ (slope 2.78)

parameter (0.3, 0.1). The pair corresponding to the trajectory of minimum distance (green dot) is closer to the true parameter (red dot) than the pair estimated from the trajectory of maximum distance (yellow dot).

An interesting feature of Fig. 13 is that the ratio $\hat{\beta}_t/\hat{\gamma}_t$ shows a smaller variability than $\hat{\beta}_t$ and $\hat{\gamma}_t$, around a straight line with slope close to 3, the true value of R_0 .

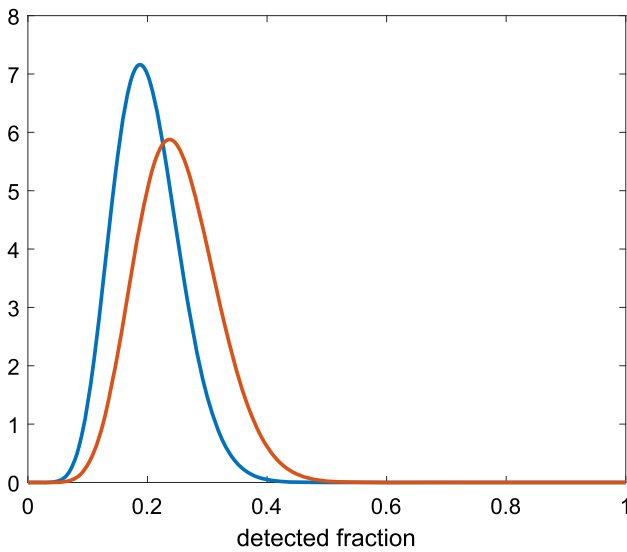


Fig. 14 Comparison between the two beta densities used to model the observation error

Appendix 2.2: Identifiability

A statistical model, belonging to a parametric family, is said to be identifiable if, for any two different values of parameters, there exists at least a measurable set in the sample space that is not assigned the same probability by the two members of the family, that is, given $\theta_{01} \neq \theta_{02}$, there exists at least one set B such that

$$\Pr(Y_{1:t} \in B; \theta_{01}) \neq \Pr(Y_{1:t} \in B; \theta_{02}) , \tag{24}$$

where $Y_{1:t}$ denotes a finite-length trajectory of observations from (7). For a deterministic model, this property is called structural identifiability, which holds if there exists a map from the parameter to the output $\theta_0 \mapsto y_{1:t}(\theta_0)$ which is

injective, that is, given $\theta_{01} \neq \theta_{02}$, the two models $y(\theta_{01})$ and $y(\theta_{02})$ describe different output trajectories. By a differential algebra approach, Piazzola et al. (2021) have shown that the following deterministic SIR model with its output

$$\begin{cases} \frac{dI_t}{dt} = \beta_0 I_t (N - I_t - R_t) - \gamma_0 I_t \\ \frac{dR_t}{dt} = \gamma_0 I_t \\ y_{1,t} = \frac{1}{K} I_t \\ y_{2,t} = \frac{1}{K} R_t \end{cases} \tag{25}$$

defined for a non-normalised population of size N , is structurally identifiable with respect to the unknown parameters β_0, γ_0 and K . The parameter $K > 1$ accounts for underreporting of infected and removed and has the same function as the U random variables in (7). Then, after adding noise to the output, the authors in Piazzola et al. (2021) went on to show that, despite structural identifiability, the parameters may not be practically identifiable, that is, a good or acceptable agreement between observations and fit is displayed for different values of the parameters when observations end before reaching the peak.

The way randomness has been included into this problem via model (11)-(7) is different from Piazzola et al. (2021), but we have also observed practical identifiability. The identifiability problem was also discussed in Ganyani et al. (2021) for stochastic SIR models. We generated state trajectories composed by 30 daily values using model (4) with parameters $\beta_0 = 0.1$ and $\gamma_0 = 0.03$. Then, to obtain the actual observations, we multiplied each value by a

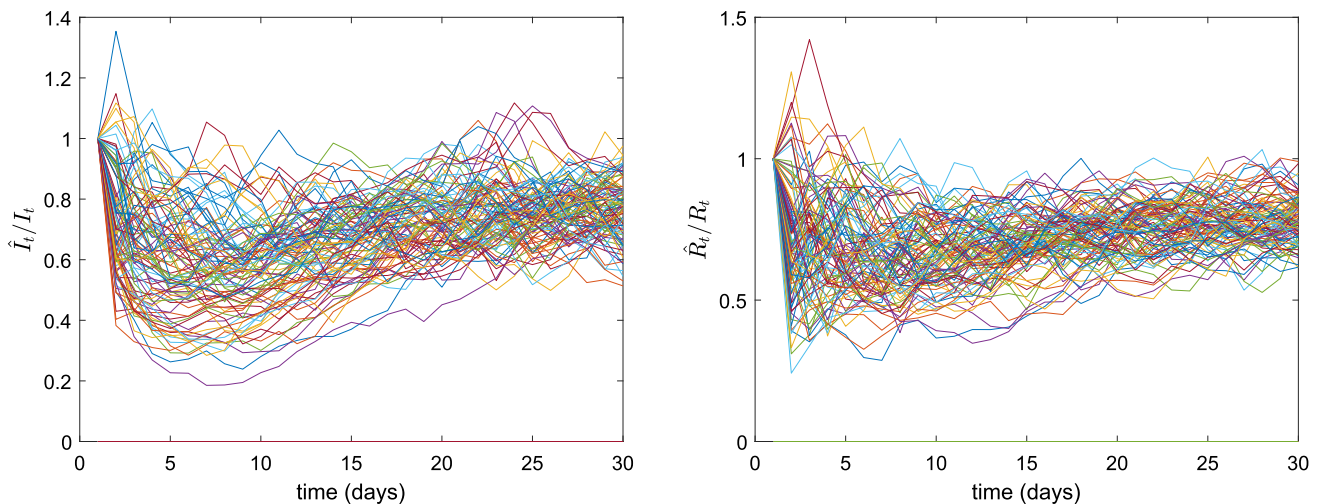


Fig. 15 Left panel: behaviour of \hat{I}_t/I_t for 100 different simulations. Right panel: behaviour of \hat{R}_t/R_t for 100 different simulations

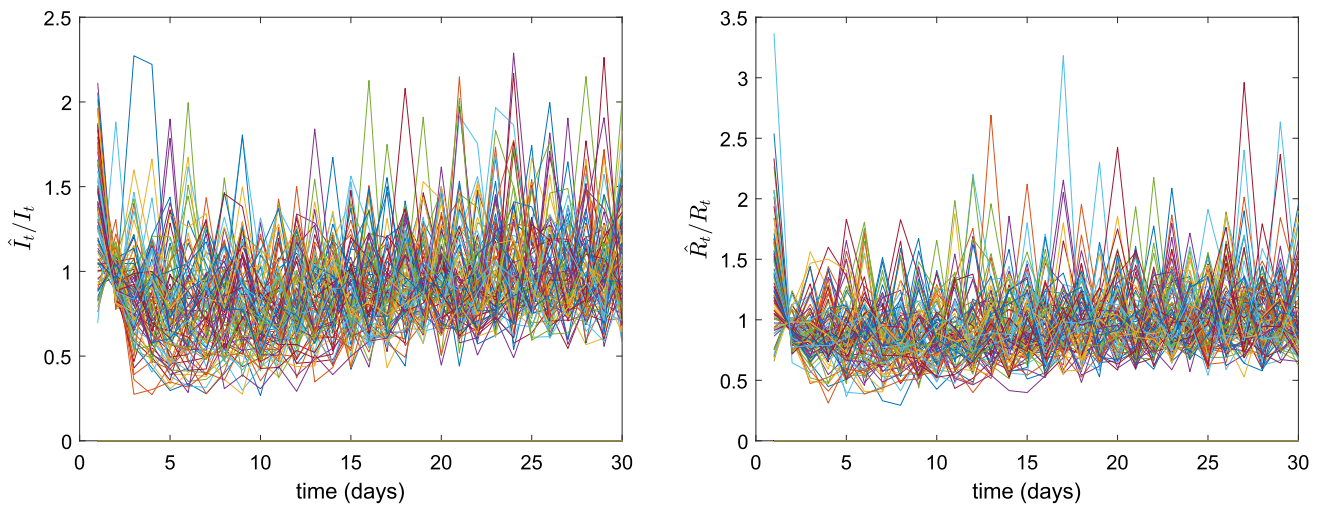


Fig. 16 Left panel: behaviour of $\hat{I}_t / (y_{t,1} / m)$ for 100 different simulations, where m is the true median of the observation error distribution. Right panel: behaviour of $\hat{R}_t / (y_{t,2} / m)$ for 100 different simulations

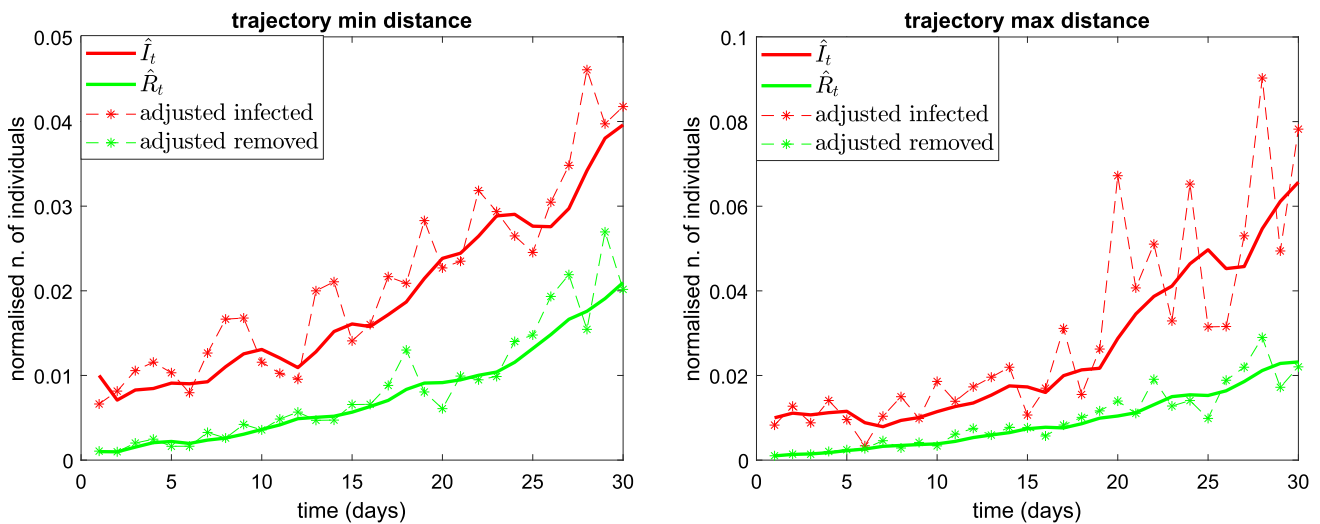


Fig. 17 Left panel: case of minimum sum of root mean square error between true and filtered state, dynamics of \hat{I}_t and \hat{R}_t (continuous lines) and adjusted observations (thin line with asterisks). Right panel:

case of maximum sum of root mean square error between true and filtered state, dynamics of \hat{I}_t and \hat{R}_t (continuous lines) and adjusted observations (thin line with asterisks)

number drawn from a beta distribution with parameters $a = 10$ and $b = 40$ (blue line in Fig. 14). After running the RBPf algorithm using the same beta distribution we obtained results analogous to those in the previous sections, that is, a satisfactory fit of the observed dynamics and a good estimate of the parameters β_0 and γ_0 .

Then we run the RBPf algorithm assuming an observation error with beta distribution with a mean larger than the truth, with parameters $a = 10$ and $b = 30$ (red line in Fig. 14). We compared the filtered states with both the true

ones and the observed data. First, we considered 500 simulations and looked at the ratio between the filtered state and the state. For a clear representation, in Fig. 15 we show one in five trajectories for both infected and removed individuals. These ratios are, generally, less than 1, denoting an underestimation of both infected and removed.

Then, we considered the ratio between the filtered states and the adjusted observations. The results of one in five trajectories are reported in Fig. 16 for both infected and removed individuals. These ratios fluctuate around 1,

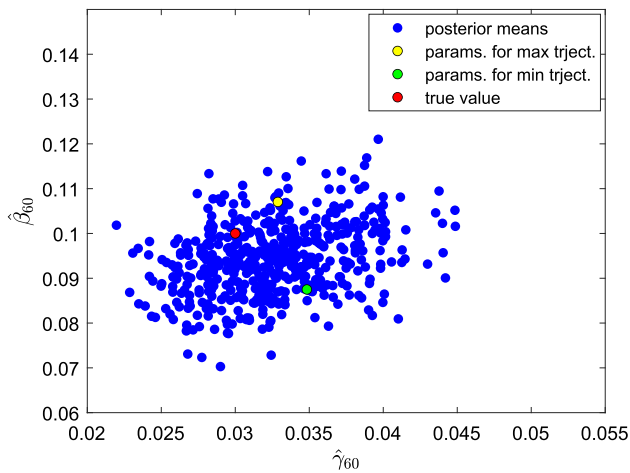


Fig. 18 Scatter plot of the parameters $(\hat{\gamma}_{60}, \hat{\beta}_{60})$, obtained for the different simulations, with wrong parameters a and b in the observation error distribution. The red dot represents the true pair $(0.03, 0.1)$. The green dot is the estimate corresponding to minimum distance case (left panel of Fig. 17). The yellow dot is the case corresponding to the maximum distance case (right panel of Fig. 17)

denoting a satisfactory fit to the observed data. Fig. 17 shows the dynamics of two simulations: the dynamics with minimum distance of the filtered state from the true state (intended as the minimum sum of the root mean square errors over the two components), in the left panel, and the dynamics with maximum distance, in the right panel. Both trajectories fit the (scaled) observed data reasonably well.

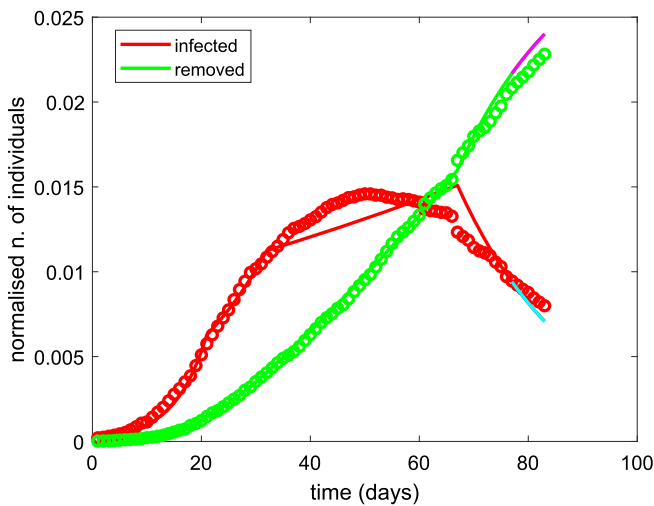


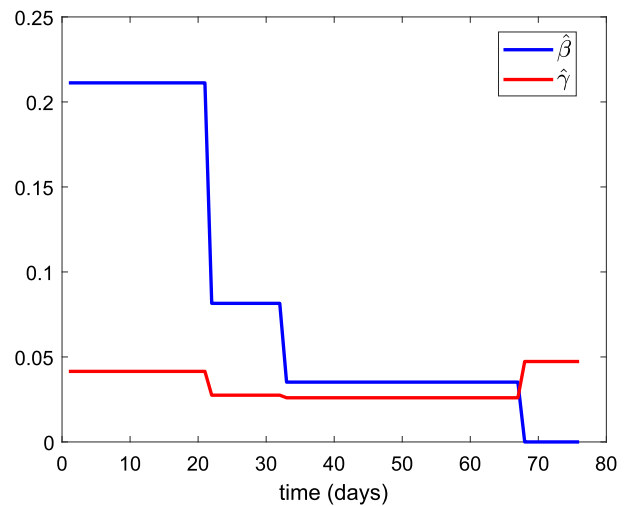
Fig. 19 First wave: piecewise deterministic SIR simulations with ML parameter estimates and adjusted observations and seven-day forecast (left) until 15th May; parameter estimates $(\hat{\beta}, \hat{\gamma})$ in each phase (right)

Even if the filtered states follow the adjusted observations, the values of the estimated parameters are not correct, as shown in Fig. 18. In fact, the pairs of estimated parameters in the 100 simulations are not equally dispersed around the true value (red point) but are placed mainly below the true value, denoting a bad estimation for β_0 . The estimation of γ_0 is better.

It follows that it is very important to suitably choose the beta distribution of the observation error (as we have done in Sect. 5) in the collection of infected and removed people to avoid practical nonidentifiability.

Appendix 3: A comparison with the SIR deterministic model

We considered for comparison the model (1) combined with the observation equations (7), that is, when the state dynamics is completely deterministic. We repeated part of the analysis done with the stochastic SIR model on the first wave data, using the same observation error distribution and the same time partition based on DPCMs, and estimated (β, γ) via maximum likelihood. The piecewise deterministic SIR model, see Fig. 19, follows the scaled observations but it is rather less flexible than its stochastic counterpart (see Fig. 5).



are $(0.21, 0.04)$, $(0.08, 0.03)$, $(0.04, 0.03)$, $(1.6 \times 10^{-8}, 0.05)$. Cut dates correspond to 10 days after the government decrees on 11th March, 22nd March, 26th April. Time 0 is 1st March

Acknowledgements The authors acknowledge the many fruitful discussions with several colleagues, and in particular the colleagues at CNR-IMATI that participated in the COVID-19 modelling study group.

Declarations

Conflict of interests The authors did not receive support from any organisation for the submitted work. The authors have no relevant financial or non-financial interests to disclose.

References

- Allieta M, Allieta A, Sebastiano DR (2021) COVID-19 outbreak in Italy: estimation of reproduction numbers over 2 months prior to phase 2. *J Public Health* 1:9
- Bassi F, Arbia G, Falorsi P (2021) Observed and estimated prevalence of Covid-19 in Italy: How to estimate the total cases from medical swabs data. *Sci Total Environ* 764:142799
- Bellman R, Åström KJ (1970) On structural identifiability. *Math Biosci* 7(3–4):329–339
- Berardi C, Antonini M, Genie MG, Cotugno G, Lanteri A, Melia A, Paolucci F (2020) The COVID-19 pandemic in Italy: policy and technology impact on health and non-health outcomes. *Health Policy Technol* 9(4):454–487
- Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D (2020) The challenges of modeling and forecasting the spread of COVID-19. *Proc Natl Acad Sci* 117(29):16732–16738
- Bilge AH, Samanlioglu F, Ergonul O (2015) On the uniqueness of epidemic models fitting a normalized curve of removed individuals. *J Math Biol* 71(4):767–794
- Bontempi E (2020) The Europe second wave of COVID-19 infection and the Italy “strange” situation. *Environ Res* 193:110476
- Bosa I, Castelli A, Castelli M, Ciani O, Compagni A, Galizzi MM, Garofano M, Ghislandi S, Giannoni M, Marini G et al (2021) Response to COVID-19: was Italy (un) prepared? *Health Econ Policy Law*. <https://doi.org/10.1017/S1744133121000141>
- Brazeau NF, Verity R, Jenks S, Fu H, Whittaker C, Winskill P, Dorigatti I, Walker P, Riley S, Schnekenberg RP et al (2020) COVID-19 infection fatality ratio: estimates from seroprevalence. Technical report. Imperial College London. <https://doi.org/10.25561/83545>
- Brown TS, Walensky RP (2020) Serosurveillance and the COVID-19 epidemic in the US: undetected, uncertain, and out of control. *JAMA* 324(8):749–751
- Byambasuren O, Dobler CC, Bell K, Rojas DP, Clark J, McLaws M-L, Glasziou P (2021) Comparison of seroprevalence of SARS-CoV-2 infections with cumulative and imputed COVID-19 cases: systematic review. *PLoS ONE* 16(4):e0248946
- Calafiore GC, Novara C, Possieri C (2020) A Modified SIR Model for the COVID-19 Contagion in Italy. In: 2020 59th IEEE Conference on Decision and Control (CDC), pp 3889–3894. <https://doi.org/10.1109/CDC42340.2020.9304142>
- Carletti T, Fanelli D, Piazza F (2020) COVID-19: The unreasonable effectiveness of simple models. *Chaos Solit Fract* X 5:100034
- Cereda D, Tirani M, Rovida F, Demicheli V, Ajelli M, Poletti P, Trentini F, Guzzetta G, Marziano V, Barone A et al (2020) The early phase of the COVID-19 outbreak in Lombardy, Italy. arXiv preprint [arXiv:2003.09320](https://arxiv.org/abs/2003.09320)
- D’Arienzo M, Coniglio A (2020) Assessment of the SARS-CoV-2 basic reproduction number, R_0 , based on the early phase of COVID-19 outbreak in Italy. *Biosaf Health* 2(2):57–59
- De Natale G, Ricciardi V, De Luca G, De Natale D, Di Meglio G, Ferragamo A, Marchitelli V, Piccolo A, Scala A, Somma R et al (2020) The COVID-19 Infection in Italy: a Statistical Study of an Abnormally Severe Disease. *J Clin Med* 9(5):1564
- Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH (2019) Complexity of the Basic Reproduction Number (R_0). *Emerg Infect Dis* 25(1):1
- Deo V, Grover G (2021) A new extension of state-space SIR model to account for Underreporting—An application to the COVID-19 transmission in California and Florida. *Results Phys* 24:104182
- Doucet A, Godsill S, Andrieu C (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat Comput* 10(3):197–208
- Duysburgh E, Mortgat L, Barbezange C, Dierick K, Fischer N, Heyndrickx L, Hutse V, Thomas I, Van Gucht S, Vuylsteke B et al (2021) Persistence of IgG response to SARS-CoV-2. *Lancet Infect Dis* 21(2):163–164. Erratum in: *Lancet Infect Dis*. 2020
- Fabiani M, Onder G, Boros S, Spuri M, Minelli G, Mateo Urdiales A, Andrianou X, Riccardo F, Del Manso M, Petrone D, Palmieri L, Vescio MF, Bella A, Pezzotti P (2020) Il case fatality rate dell’infezione SARS-CoV-2 a livello regionale e attraverso le differenti fasi dell’epidemia in Italia. Technical Report 1/2021, Istituto Superiore di Sanità. [Online; accessed 28 June 2021; in Italian]
- Flaxman S, Mishra S, Gandy A et al (2020) Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* 584:257–261
- Ganyani T, Faes C, Hens N (2021) Simulation and analysis methods for stochastic compartmental epidemic models. *Ann Rev Stat Appl* 8:19.1-19.20
- Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, Rinaldo A (2020) Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures. *Proc Natl Acad Sci* 117(19):10484–10491
- Ghani AC, Donnelly CA, Cox DR, Griffin JT, Fraser C, Lam TH, Ho LM, Chan W-S, Anderson RM, Hedley AJ et al (2005) Methods for estimating the case fatality ratio for a novel, emerging infectious disease. *Am J Epidemiol* 162(5):479–486
- Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, Colaneri M (2020) Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat Med* 26(6):855–860
- Gnanvi J, Salako KV, Kotanmi B, Kakai RG (2021) On the reliability of predictions on Covid-19 dynamics: a systematic and critical review of modelling techniques. *Infect Dis Modell* 6:258–272
- Hong HG, Li Y (2020) Estimation of time-varying reproduction numbers underlying epidemiological processes: a new statistical tool for the COVID-19 pandemic. *PLoS ONE* 15(7):e0236464
- Istituto Nazionale di Statistica (2020) Primi risultati dell’indagine di sieroprevalenza sul SARS-CoV-2. Technical report. [Online; accessed 28 June 2021; in Italian]
- Istituto Superiore di Sanità (2020a) Epidemia COVID-19. Aggiornamento nazionale 28 aprile 2020 – ore 16:00. https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_28-aprile-2020.pdf. [Online; accessed 28 June 2021; in Italian]
- Istituto Superiore di Sanità (2020b) Epidemia COVID-19. Aggiornamento nazionale 25 novembre 2020 – ore 16:00. https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_25-novembre-2020.pdf. [Online; accessed 28-June-2021; in Italian]
- Katul GG, Mrad A, Bonetti S, Manoli G, Parolari AJ (2020) Global convergence of COVID-19 basic reproduction number and estimation from early-time SIR dynamics. *PLoS ONE* 15(9):e0239800

- Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proc R Soc Lond Ser A* 115(772):700–721
- Krantz SG, Rao ASS (2020) Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: preliminary retrospective results based on wavelets and deterministic modeling. *Infect Control Hosp Epidemiol* 41(7):857–859
- Kuster AC, Overgaard HJ (2021) A novel comprehensive metric to assess effectiveness of COVID-19 testing: Inter-country comparison and association with geography, government, and policy response. *PLoS ONE* 16(3):e0248176
- Lee S, Kim T, Lee E, Lee C, Kim H, Rhee H, Park SY, Son H-J, Yu S, Park JW et al (2020) Clinical course and molecular viral shedding among asymptomatic and symptomatic patients with SARS-CoV-2 infection in a community treatment center in the Republic of Korea. *JAMA Intern Med* 180(11):1447–1452
- Linka K, Peirlinck M, Kuhl E (2020) The reproduction number of COVID-19 and its correlation with public health interventions. *Comput Mech* 66(4):1035–1050
- Liu Z, Magal P, Webb G (2021) Predicting the number of reported and unreported cases for the COVID-19 epidemics in China, South Korea, Italy, France, Germany and United Kingdom. *J Theor Biol* 509:110501
- Mallapaty S (2020) How deadly is the coronavirus? Scientists are close to an answer. *Nature* 582:467–468
- Meyerowitz-Katz G, Merone L (2020) A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. *Int J Infect Dis* 101:138–148
- Morettini M, Sbröllini A, Marcantoni I, Burattini L (2020) Covid-19 in Italy: dataset of the Italian civil protection department. *Data Brief* 30:105526
- Noh J, Danuser G (2021) Estimation of the fraction of COVID-19 infected people in US states and countries worldwide. *PLoS ONE* 16(2):e0246772
- Oke J, Heneghan C (2020) Global Covid-19 case fatality rates. <https://www.cebm.net/covid-19/global-covid-19-case-fatalityrates>. [Online; accessed 28 June 2021]
- Oran DP, Topol EJ (2021) Prevalence of asymptomatic SARS-CoV-2 infection. *Ann Intern Med* 174(2):286–287
- Phipps SJ, Grafton RQ, Kompas T (2020) Robust estimates of the true (population) infection rate for COVID-19: a backcasting approach. *Royal Soc Open Sci* 7(11):200909
- Piazzola C, Tamellini L, Tempone R (2021) A note on tools for prediction under uncertainty and identifiability of SIR-like dynamical systems for epidemiology. *Math Biosci* 332:108514
- Pullano G, Di Domenico L, Sabbatini CE, Valdano E, Turbelin C, Debin M, Guerrisi C, Kengne-Kuetche C, Souty C, Hanslik T et al (2021) Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature* 590(7844):134–139
- Riccardo F, Ajelli M, Andrianou XD, Bella A, Del Manso M, Fabiani M, Bellino S, Boros S, Urdiales AM, Marziano V et al (2020) Epidemiological characteristics of COVID-19 cases and estimates of the reproductive numbers 1 month into the epidemic, Italy, 28 January to 31 March 2020. *Euro Surveill* 25(49):2000790
- Röltgen K, Powell AE, Wirz OF, Stevens BA, Hogan CA, Najeeb J, Hunter M, Wang H, Sahoo MK, Huang C et al (2020) Defining the features and duration of antibody responses to SARS-CoV-2 infection associated with disease severity and outcome. *Sci Immunol* 5(54):eabe0240
- Russell TW, Golding N, Hellewell J, Abbott S, Wright L, Pearson CA, van Zandvoort K, Jarvis CI, Gibbs H, Liu Y et al (2020) Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Med* 18(1):1–9
- Russell TW, Hellewell J, Jarvis CI, Van Zandvoort K, Abbott S, Ratnayake R, Flasche S, Eggo RM, Edmunds WJ, Kucharski AJ et al (2020) Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Euro Surveill* 25(12):2000256
- Sánchez-Romero M, di Lego V, Prskawetz A, Queiroz BL (2021) An indirect method to monitor the fraction of people ever infected with COVID-19: An application to the United States. *PLoS ONE* 16(1):e0245845
- Scozzari G, Costa C, Migliore E, Coggiola M, Ciccone G, Savio L, Scarmozzino A, Pira E, Cassoni P, Galassi C et al (2021) Prevalence, persistence, and factors associated with SARS-CoV-2 IgG seropositivity in a large cohort of healthcare workers in a Tertiary Care University Hospital in Northern Italy. *Viruses* 13(6):1064
- Stocks T, Britton T, Höhle M (2021) Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. *Biostatistics* 21(3):400–416
- Tosi D, Verde A, Verde M (2020) Clarification of misleading perceptions of COVID-19 fatality and testing rates in Italy: data analysis. *J Med Internet Res* 22(6):e19825
- Tuite AR, Ng V, Rees E, Fisman D (2020) Estimation of COVID-19 outbreak size in Italy. *Lancet Infect Dis* 20(5):537
- Wang N, Fu Y, Zhang H, Shi H (2020) An evaluation of mathematical models for the outbreak of COVID-19. *Precision Clin Med* 3(2):85–93
- World Health Organization (2020a) COVID-19 Public Health Emergency of International Concern (PHEIC) global research and innovation forum: towards a research roadmap. [https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-\(pheic\)-global-research-and-innovation-forum](https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum). [Online; accessed 2 July 2021]
- World Health Organization (2020b) WHO Director-General's opening remarks at the media briefing on COVID-19 - 16 March 2020. <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-16-march-2020>. [Online; accessed 28 June 2021]
- Yu C-J, Wang Z-X, Xu Y, Hu M-X, Chen K, Qin G (2021) Assessment of basic reproductive number for COVID-19 at global level: a meta-analysis. *Medicine* 100(18):e25837
- Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen X, Li M, Zheng W, Yi L, Chen X et al (2020) Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect Dis* 20(7):793–802

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.