Annual Reviews in Control 51 (2021) 460–476

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

Forecasting COVID-19 cases based on a parameter-varying stochastic SIR model

João P. Hespanha [a,*], Raphael Chinchilla [a], Ramon R. Costa [b], Murat K. Erdal [a], Guosong Yang [a]

[a] University of California, Santa Barbara, USA
[b] Federal University of Rio de Janeiro, Brazil

ARTICLE INFO

Keywords:
Epidemic models
System identification
Estimation

ABSTRACT

We address the prediction of the number of new cases and deaths for the coronavirus disease 2019 (COVID-19) over a future horizon from historical data (forecasting). We use a model-based approach based on a stochastic Susceptible–Infections–Removed (SIR) model with time-varying parameters, which captures the evolution of the disease dynamics in response to changes in social behavior, non-pharmaceutical interventions, and testing rates. We show that, in the presence of asymptomatic cases, such model includes internal parameters and states that cannot be uniquely identified solely on the basis of measurements of new cases and deaths, but this does not preclude the construction of reliable forecasts for future values of these measurements. Such forecasts and associated confidence intervals can be computed using an iterative algorithm based on nonlinear optimization solvers, without the need for Monte Carlo sampling. Our results have been validated on an extensive COVID-19 dataset covering the period from March through December 2020 on 144 regions around the globe.

1. Introduction

The recent global epidemics of SARS first reported in 2003, Swine flu in 2009, MERS in 2012, Western African Ebola in 2013, Zika in 2015, and COVID-19 in 2019, identified building epidemic models specifically aimed at forecasting the propagation of contagious diseases as a key need to guide epidemic response (Morgan, 2019; Shearer, Moss, McVernon, Ross, & McCaw, 2020). Motivated by this need, the main goal of this paper is to compute forecasts for the number of new cases and deaths due to the COVID-19 outbreak to aid decision makers in provisioning healthcare resources or imposing nonpharmaceutical interventions. In this context, forecasts must reliably provide measures of confidence to enable decision makers to plan for worst-case scenarios. We construct such forecasts using a time-varying stochastic Susceptible–Infected–Removed (SIR) epidemic model, that is fitted to daily measurements of new cases and deaths.

SIR epidemic models are based on the assumption that each individual of a population is in one of three basic states: *susceptible* to infection, but not yet infected by the virus; *infective* and thus contagious; and *removed* from the infective state either because the individual developed antibodies and is no longer susceptible to the infection or because the individual passed away. In this paper, we consider *compartmental* SIR models that are focused on counting the number of individuals in each of the states (also known as compartments). The estimation of the state and parameters of an SIR models is typically based on a measurement model that maps the model's state to time series of measurements that typically include daily counts of newly discovered infected patients and deaths.

The SIR dynamics and measurements depend on a number of key parameters that include the *infection rate* that can be interpreted as the probability that a susceptible individual will become infected due to interactions with infected individuals; the *removal rate* that can be regarded as the probability that an individual leaves the infective state; the *deaths reporting rate* that can be regarded as the probability that a new death is reported; and the *new-cases reporting rate* that can be regarded as the probability that a new infection is reported. These parameters are strongly influenced by biological properties of the pathogen that causes the disease, such as its ability to travel from individual to individual, the body's natural ability to fight it, and whether or not the disease exhibits noticeable symptoms when an individual is infected. However, all these parameters are also strongly influenced by external sources, which from the perspective of an SIR model, often dominate in determining their values:

1. The infection rate is greatly modulated by the degree to which the population is engaging in social distancing, which in turn

* Corresponding author.
E-mail addresses: hespanha@ece.ucsb.edu (J.P. Hespanha), raphaelchinchilla@ucsb.edu (R. Chinchilla), ramon@coep.ufrj.br (R.R. Costa), m_erdal@ucsb.edu (M.K. Erdal), guosongyang@ucsb.edu (G. Yang).

https://doi.org/10.1016/j.arcontrol.2021.03.008
Received 30 December 2020; Received in revised form 22 February 2021; Accepted 25 March 2021
Available online 8 April 2021

depends on media coverage of the epidemic and nonpharmaceutical interventions.
2. The removal rate expresses the rate by which patients leave the state in which they can pass the disease to susceptible patients, which may be quite different from the rate at which they get cured (or die). This rate is thus strongly effected by social behavior and policies regarding quarantine, contact tracing and testing, since a patient can stop infecting others much earlier than cure/death.
3. In the absence of testing of asymptomatic individuals, the reporting rate is essentially the fraction of symptomatic patients; otherwise it will depend strongly on the policies in effect regarding the testing of asymptomatic individuals.

The strong dependence of the SIR parameters on social factors prevents extrapolating their values across time and space: one cannot estimate the value of a parameter in one region/time and expect it to remain the same in a different region or even in the same region at a later time. These observations motivate two key choices behind this work: All SIR parameters need to be learned from data collected on a relatively homogeneous region of space and the parameter values must be allowed to drift over time.

Motivated by the observations above, we introduce in Section 2 a stochastic SIR model with three sources of stochasticity: First, we introduce a stochastic component to the number of individuals that transition between states; second, we assume that daily measurements for the number of new cases and deaths are corrupted by (stochastic) noise; and finally, we take the key model parameters as random walks that drift over time. The variances associated with all these stochastic components need to be learned from data and, for the reasons outlined above, we do not extrapolate their values across different countries/regions.

From a methodological perspective, we regard forecasting as computing the a posteriori distribution of the random variables that we want to estimate, and subsequently extracting from those distributions point estimates and associated confidence intervals. The a posteriori distributions depend on unknown parameters that are estimated using maximum likelihood. The key challenge of this approach is that, because the SIR model is nonlinear, it is not possible to compute in closed form the likelihood function and the corresponding a posteriori distributions for the variables that we want to forecast. We overcome this by essentially using Laplace's method to approximate the integral that appears in the formula of the likelihood function. We show in Section 3 that this approach enables the computation of the maximum likelihood values for the unknown parameters and the a posteriori point estimates and associated error covariances through a single deterministic optimization that can be solved numerically. For large datasets this optimization may be computationally difficult, so we propose an iterative algorithm that alternates between two smaller optimizations for which the computation required by a 2nd order numerical solver scales linearly with the length of the dataset and the forecasting horizon. This approach is quite general and can be used for much more general estimation/forecasting problems.

Our modeling and forecasting approaches were validated on an extensive collection of COVID-19 datasets. For 144 regions around the world, we computed weekly 7, 14, and 21 days-ahead forecasts from March to December of 2020 for the number of new cases and deaths due to COVID-19. These forecasts and all the associated initial conditions and parameter estimates were computed solely using past data and then compared with the actual (future) data. We started to produce forecasts with as little as 21 days of data, but it generally took 4–5 weeks of data to start getting confidence intervals that are somewhat tight.

We show formally in Section 2.3 and observe numerically through the results in Section 4, that an SIR model that includes an unknown reporting rate is unidentifiable, in the sense that multiple sets of parameters can explain the same observed data with equal likelihood (in the sense of maximum likelihood). This means that there is a fundamental ambiguity in estimating SIR model parameters from measurements of new cases and deaths; an observation that is often ignored but had been made in prior work (Comunian, Gaburro, & Giudici, 2020). However, this ambiguity does not prevent the computation of reliable forecasts for the daily number of news cases and deaths. This is because the ambiguity exists in parameter space but not on the space of the variables that we are trying to forecast. In essence, while multiple sets of parameters may have the same likelihood, they result in consistent forecasts.

To study the importance of parameter drift, we compare our proposed stochastic SIR model with a hierarchy of simpler models for which we take the infection, new-cases and/or deaths reporting rates to be constant, rather than time-varying; and also consider a more complex model with a time-varying removal rate. Across a large number of countries/regions, we conclude that taking all of these parameters, or even just the infection rate, to be constant leads to poor results; either resulting in gross violations of the confidence intervals or to overly wide confidence intervals for the forecasts. This is especially noticeable (and not surprising) in regions that show multiple waves of infection, which could never be explained by a constant parameter SIR model. The numerical results also show that, in general, taking all parameters to be time varying does not result in forecasts that are significantly better than those obtained by assuming that the death rate is constant. In fact, for most countries/regions assuming constant removal and death rates result in more accurate forecast.

Related work

The basic SIR epidemic model with the number of new cases proportional to the product of the numbers of susceptible and infected individuals can be traced as far back as the work of Hamer (1906). Since then, SIR models have evolved in multiple directions, including incorporating stochastic effects (Andersson & Britton, 2000; Ball & Neal, 2002; Beretta, Capasso, & Rinaldi, 1988; Beretta, Kolmanovskii, & Shaikhet, 1998; Beretta & Takeuchi, 1995; Ji & Jiang, 2014; King, Domenech de Cellès, Magpantay, & Rohani, 2015; Tornatore, Buccellato, & Vetro, 2005), the addition of new compartments corresponding to different states of the disease (Capasso, 2008; Efimov & Ushirobira, 2020; Giordano et al., 2020; IHME COVID-19, 2020; Keeling & Rohani, 2008; Köhler et al., 2020; Peng, Yang, Zhang, Zhuge, & Hong, 2020; Xia, Zhang, Xue, Sun, & Jin, 2015), and considering a network of interacting populations (Della Rossa et al., 2020; Mei, Mohagheghi, Zampieri, & Bullo, 2017; y Piontti, Perra, Rossi, Samay, & Vespignani, 2019; Stolerman, Coombs, & Boatto, 2015; Youssef & Scoglio, 2011). The reader is referred to (Hethcote, 2000) for an historical perspective on deterministic SIR-like epidemic models and their analysis and to the monograph (Brauer, Castillo-Chavez, & Feng, 2019) for the application of such models to several diseases.

A common feature to many recent models is the addition of states to address the existence of individuals that are infected and can transmit the disease, but are asymptomatic and thus are not accounted for as infected in official reports (Capasso, 2008; Giordano et al., 2020; IHME COVID-19, 2020; Köhler et al., 2020; Li et al., 2020; Xia et al., 2015; Zou et al., 2020). The addition of states has also been used to account for individuals under quarantine (Li et al., 2020; Peng et al., 2020), infected but not yet infective (Efimov & Ushirobira, 2020; IHME COVID-19, 2020; Köhler et al., 2020; Li et al., 2020; Peng et al., 2020), asymptomatic but diagnosed through testing (Giordano et al., 2020; Köhler et al., 2020), and hospitalized (Li et al., 2020; Xia et al., 2015). The inclusion of more states and the associated parameters that determine the rates of transfer between states, facilitates matching measurements with the model outputs. However, it also makes the model identification problem more formidable, especially because most of these parameters may change as the epidemic evolves. In fact, it was shown in Roda, Varughese, Han, and Li (2020) that adding an

"exposed" (E) state (i.e., infected but not yet infective) to a basic SIR model, actually results in a worse value for the Akaike Information Criterion. This essentially means that, while an SEIR model can better represent the data, this improvement does not suffice to justify the additional model complexity.

It is widely accepted that beyond the initial outbreak, the parameters of an SIR model vary in response to changes in social behavior and medical advances. A common approach to address this consists of breaking the epidemic into stages and identifying a different set of model parameters for each stage (Giordano et al., 2020; Xia et al., 2015), with the times of the transitions between stages typically selected to coincide with the introduction of nonpharmaceutical measures. The model used in Srivastava, Xu, and Prasanna (2020) assumes piecewise constant infection rates that remain constant over an intervals of length J, that is learned from data. The model in Calafiore, Novara, and Possieri (2020) expresses the time-varying parameters through a linear combination of pre-specified time functions, with the coefficients of these linear combinations identified from data. This permits more realistic smooth variations of the parameters, but makes the forecast highly dependent on the choice of the basis functions, which must be pre-specified and not learned from data. In Al-Salti, Al-Musalhi, Elmojtaba, and Gandhi (2020), the infection rate is assumed to be monotone decreasing, evolving according to a deterministic differential equation that depends on 3 parameters that can be adjusted. The model (IHME COVID-19, 2020) considers a time-varying infection rate that is assumed to be a linear combination of a set of explanatory covariates that include seasonality, mobility, testing rates, and mask use. The coefficients of this linear combination are estimated from data.

Stochastic SIR models appeared in many flavors: Beretta et al. (1988) and Beretta and Takeuchi (1995) considered stochasticity in the delay from the time an individual gets infected until he/she becomes infective, leading to an integral differential equation with delays; Ji and Jiang (2014) introduced stochasticity in the form of an additive stochastic perturbation, resulting in a stochastic differential equation; and both these sources of stochasticity appear combined in Beretta et al. (1998) and Tornatore et al. (2005). A fundamentally different stochastic model was proposed in Ball and Neal (2002), where individuals enter and exit the infective state at points on Poisson processes. It considers different rates for the Poisson processes regarding on whether individuals share the same household, workplace, etc. A more conventional SEIR model in the form of a continuous-time Markov process was considered in King et al. (2015). As in Beretta et al. (1998), Ji and Jiang (2014) and Tornatore et al. (2005), our paper considers additive stochastic perturbations, whose variances are estimated from data. However, we shall see that our numerical results for COVID-19 indicate that this stochastic effect rapidly becomes negligible as the epidemic progresses. The key stochastic component to our model will turn out to be the SIR model parameters, which we regard as realizations of Gaussian random walks with unknown variances that must be learned from data.

The identification of SIR models based on fitting *cumulative* data of the total number of cases and recoveries since the start of the epidemic has been widely used in the literature (Calafiore et al., 2020; Comunian et al., 2020; Efimov & Ushirobira, 2020; Giordano et al., 2020; Köhler et al., 2020; Peng et al., 2020; Stolerman et al., 2015; Xia et al., 2015; Zou et al., 2020). However, it was shown in King et al. (2015) that the use of cumulative data can lead to non-independent successive errors, resulting in confidence intervals that suggest a degree of precision that is not consistent with the data. In view of this, and as in Roda et al. (2020) and Srivastava et al. (2020), our identification procedure is based on daily counts of new patients and deaths, rather than cumulative counts of infected, removed, and dead patients.

An additional aspect in which our work differs from a large number of previous works on epidemic forecast is that we learn all model parameters from data, whereas much of the prior work relies on a combination of fitted parameters with "clinical information" (Giordano et al., 2020; Köhler et al., 2020; Li et al., 2020; Peng et al., 2020; Xia et al., 2015). As noted above, we opted to avoid relying on external data as all SIR parameters are strongly dependent on social behavior that is hard to extrapolate over time and space. Notable works that are mostly data driven include (Calafiore et al., 2020; Comunian et al., 2020; IHME COVID-19, 2020). However, IHME COVID-19 (2020) brings to the epidemic model a large corpus of external data in the form of covariates which are assumed to "explain" the future evolution of model parameters.

2. SIR stochastic modeling

Denoting by $\nu(t)$ the number of patients that were infected during day t and by $\rho(t)$ the number of removed patients on day t (i.e., patients that exited the infective state either through death or recovery), we have that

$$S(t+1) = S(t) - \nu(t), \tag{1a}$$

$$I(t+1) = I(t) + \nu(t) - \rho(t), \tag{1b}$$

$$R(t+1) = R(t) + \rho(t), \tag{1c}$$

where $S(t)$, $I(t)$, and $R(t)$ denote the cumulative numbers of individuals susceptible to the infection, infective patients, and removed patients, respectively, at the start of day t. A *classical SIR model* postulates that

$$\rho(t) = \gamma I(t), \qquad \nu(t) = \frac{\beta\, I(t)}{N_0} S(t), \tag{2}$$

where γ denotes the *removal rate*, which corresponds to the fraction of patients that leave the infective state on a particular day; N_0 the total population; and $\beta I(t)/N_0$ the fraction of susceptible individuals that become infected on day t. This model assumes that this fraction is proportional to the fraction $I(t)/N_0$ of the population that is infective and the proportionality constant β is known as the *infection rate*.

We consider two key deviations from this classical SIR model

(i) We add stochasticity, by regarding the daily number of new infective patients $\nu(t)$ and the daily number of removed patients $\rho(t)$ as random variables whose means are given by (2) but exhibit day-to-day stochastic variability.
(ii) We take the infection rate $\beta(t)$ to be time-varying and the realization of a random process that reflects the changes in population behavior over time.

These modifications lead to a *stochastic SIR model* that replaces (2) by

$$\rho(t) = \gamma I(t) + d_\rho(t), \qquad \nu(t) = \beta(t)\frac{I(t)}{N_0} S(t) + d_\nu(t),$$

where $d_\nu(t)$ and $d_\rho(t)$ are zero-mean independent Gaussian random variables that account for the daily stochastic variability of $\nu(t)$ and $\rho(t)$; and the infection rate $\beta(t)$ is also a random process. In Section 4.4, we also consider a variation of this model with a stochastic time-varying removal rate $\gamma(t)$, but we shall see that this does not appear to introduce significant improvements to the quality of our forecasts.

2.1. Measurement model

To identify the dynamics (3) and produce forecasts we use (noisy) measurement of daily new cases, of the form

$$y_\nu(t) = \phi(t)\nu(t) + w_\nu(t),$$

where the $w_\nu(t)$ denote zero-mean independent random variables that account for stochastic errors in the daily counts of new reported cases and $\phi(t) \in (0, 1]$ the fraction of infected patients that are reported as new cases on day t. The need to consider values $\phi(t) < 1$ arises from the observation that a significant fraction of the newly infected patients may not be reported because they are asymptomatic, they have not been tested, or simply because their disease has not been reported to

the entity that is keeping track of new cases. It is important to model $\phi(t)$ as a time-varying parameter because, as a pandemic progresses, one should expect significant variations in the number of asymptotic people that get tested and reported. In the sequel, we refer to $\phi(t)$ as the *new-cases reporting rate*, with the understanding that this parameter actually depends on a large number of factors aside from the actual testing rate of the population.

The use of a measurement model based on daily new cases, rather than on the cumulative number of cases, is strongly supported by the results in King et al. (2015) showing that identification based on cumulative measurements with uncorrelated noise leads to an underestimate of uncertainty.

In addition to daily counts of new cases, we also assume that we have available measurements that are roughly proportional to the number of infections, such as the daily number of deaths, the number of hospitalized patients, or the number of patients in intensive care units (ICU). The results presented here use only the number of deaths, which is available for a very large number of countries and regions. This measurement model takes the form:

$$y_D(t) = \omega(t)I(t) + w_D(t),$$

where $\omega(t)$ denotes the *deaths reporting rate*, which corresponds to the expected value of the fraction of infective patients that is likely to be reported as dead due to the epidemic on day t; and where the $w_D(t)$ are zero-mean independent random variables that account for stochastic errors. In practice, a fraction of the pandemic-related deaths may not be reported as such because of asymptomatic cases, so the deaths reporting rate $\omega(t)$ must actually reflect the fraction of infected patients that died *and* whose death was associated with the pandemic. Variability in $\omega(t)$ thus arises from a combination of factors that include medical advances in treating the disease, load on the healthcare system that may limit the patients' access to healthcare resources, as well as testing and the policy used to determine which deaths are attributed to the pandemic.

2.2. Full time-varying model

The conservation law

$$S(t) + I(t) + R(t) = S(1) + I(1) + R(1) =: N_0.$$

enable us to eliminate one of the three state variables in (1) and write the full model presented above in terms of the numbers of removed $R(t)$ and unsusceptible $U(t) := R(t) + I(t)$, $\forall t$ individuals, leading to

$$R(t+1) = R(t) + \gamma\big(U(t) - R(t)\big) + d_\rho(t), \tag{3a}$$

$$U(t+1) = U(t) + \beta(t)\big(U(t) - R(t)\big)\Big(1 - \frac{U(t)}{N_0}\Big) + d_\nu(t), \tag{3b}$$

$$y_\nu(t) = \phi(t)\Big(\beta(t)\big(U(t) - R(t)\big)\Big(1 - \frac{U(t)}{N_0}\Big) + d_\nu(t)\Big) + w_\nu(t), \tag{3c}$$

$$y_D(t) = \omega(t)\big(U(t) - R(t)\big) + w_D(t), \tag{3d}$$

where we used the facts that $I(t) = U(t) - R(t)$, $S(t) = N_0 - U(t)$, $\forall t$. This selection of states has the benefit that the dynamics in (3a)–(3b) have independent disturbances $d_\rho(t)$ and $d_\nu(t)$, which would not be the case, e.g., if we were to work with the states $R(t)$ and $I(t)$.

The key problem under consideration is to use measurements $y_\nu(t)$, $y_D(t)$ taken over a window of time $t \in \{1, 2, \dots, T\}$ to produce forecasts for the values of the same measurements on a future horizon $t \in \{T+1, T+1, \dots, T+P\}$. To solve this problem we use the model (3) to compute the a posteriori distribution of the forecasts given the available measurements, with all model parameters in (3) learned from the available measurements. These parameters include the removal rate γ, the original population N_0, the infection rate $\beta(t)$, the deaths reporting rate $\omega(t)$, and the new-cases reporting rate $\phi(t)$. We take the time-varying parameters to be Gaussian random walks of the form

$$\beta(t+1) = \beta(t) + d_\beta(t), \tag{4a}$$

$$\phi(t+1) = \phi(t) + d_\phi(t), \tag{4b}$$

$$\omega(t+1) = \omega(t) + d_\omega(t), \tag{4c}$$

where the $d_\beta(t)$, $d_\omega(t)$, $d_\phi(t)$ are independent zero-mean Gaussian processes with unknown variances, which are independent of the disturbances and measurement noise in (3). In practice, the specific realizations of these random processes will depend on a multitude of events, including the population's behavior, the enforcement of non-pharmaceutical interventions, the availability and policies regarding testing of symptomatic and asymptomatic individuals, quarantine policies and practice, etc.

Under the model (4a) with zero-mean Gaussian increments, given a specific value for the parameter β at day t, the most likely value of β at day $t+1$ is still $\beta(t)$, *in the absence of any additional information.* However, given numerical values for future measurements (3c)–(3d), the *a posteriori* distribution of $\beta(t)$ will change and $\beta(t)$ will generally not be the most likely value for $\beta(t+1)$. It is thus important to emphasize that the random walk model in (4) simply encodes an a priori assumption on the evolution of the time-varying parameters. We shall see in the numerical results shown in Section 4 that the estimates for these parameters derived from their a posteriori distributions will not have zero-mean increments.

Zero-mean independent increments for the parameters can be viewed as a very weak model that makes no a priori assumptions on how each parameter will vary from one day to the next. This assumption reflects the desire expressed in the introduction to make as few assumptions as possible on the evolution of the epidemic parameters.

In terms of forecasting, the use of zero-mean increments in (4) that are independent across time means that the past measurements collected up to time T provide no information about the value of future increments $d_\beta(t)$, $d_\phi(t)$, $d_\omega(t)$, $t > T$ and thus their most likely a posteriori value is still zero. Consequently, the mean a posteriori values of $\beta(t)$, $\phi(t)$, and $\omega(t)$ will remain constant after time T. However, the past data does provide information about the (unknown) variances of these increments and therefore the future forecasts will take into account that the a posteriori variance of the parameters grows linearly with time at a rate determined by the estimated increments' variances, which will be directly reflected in the confidence intervals associated with the forecasts.

Remark 1 (*Hospitalization and ICU Use*)**.** When we have available daily measurements of the number of hospitalized patients $y_H(t)$ and/or the number of patients in ICU units $y_{\text{ICU}}(t)$, the model (3) can be expanded to include measurements of the form

$$y_H(t) = \omega_H(t)\big(U(t) - R(t)\big) + w_H(t),$$
$$y_{\text{ICU}}(t) = \omega_{\text{ICU}}(t)\big(U(t) - R(t)\big) + w_{\text{ICU}}(t),$$

and all the results in this paper extend trivially to this enlarged set of measurements. □

2.3. Identifiability and forecastability

For every value of the constants $\bar N_0, \bar\phi(1) > 0$, making the change of variables

$$N_0 \to \bar N_0, \qquad \beta(t) \to \frac{\bar\phi(1)\bar N_0}{\phi(1)N_0}\beta(t), \tag{5a}$$

$$\phi(t) \to \frac{\bar\phi(1)}{\phi(1)}\phi(t), \qquad \omega(t) \to \frac{\bar\phi(1)}{\phi(1)}\omega(t), \tag{5b}$$

$$R(t) \to \bar N_0 - \frac{\phi(1)}{\bar\phi(1)}(N_0 - R(t)), \qquad U(t) \to \bar N_0 - \frac{\phi(1)}{\bar\phi(1)}(N_0 - U(t)), \tag{5c}$$

in the model (3) results in precisely the same measurements $y_\nu(t)$, $y_D(t)$ for the same measurement noise $w_\nu(t)$, $w_D(t)$ and disturbances

$$d_\rho(t) \to \frac{\phi(1)}{\bar\phi(1)}d_\rho(t), \qquad d_\nu(t) \to \frac{\phi(1)}{\bar\phi(1)}d_\nu(t), \tag{6a}$$

$$d_\beta(t) \to \frac{\bar\phi(1)\bar N_0}{\phi(1)N_0} d_\beta(t), \qquad d_\phi(t) \to \frac{\bar\phi(1)}{\phi(1)} d_\phi(t), \qquad d_\omega(t) \to \frac{\bar\phi(1)}{\phi(1)} d_\omega(t), \tag{6b}$$

which means that the joint distribution of the outputs of the model (3) would not change under the given transformation, if we were to adjust the (unknown) variances of the disturbances to match (6). It turns out that (5)–(6) corresponds to the only time-invariant affine transformation of $U(t)$ and $R(t)$ that preserves the structure of the SIR model (3) without changing the outputs.

A key consequence of the above observation is that, in the absence of additional information, the model outputs do not permit the identification of the original population N_0 nor the initial new-cases reporting rate $\phi(1)$; as the true values of these parameters could be replaced by arbitrary values $\bar N_0$, $\bar\phi(1)$ without changing the outputs distribution. Moreover, any estimates of the remaining parameters can only be known up to the transformation in (5)–(6). In spite of this, and precisely because this transformation does not affect the output's probability distribution, it remains possible to produce forecasts of future outputs, even if states estimates will have fundamental ambiguities. This argument shows that the model (3) is *not identifiable*, in the sense that the measurements available do not suffice to identify unique values for the states and parameters (Comunian et al., 2020; Miao, Xia, Perelson, & Wu, 2011).

The lack of state/parameter identifiability of (3), enable us to simplify this model by setting arbitrary values for N_0 and $\phi(1)$, *without compromising the quality of the forecasts*. We shall see in our numerical results that, once this ambiguity has been resolved, we obtain a posteriori probability density function of future outputs with finite error covariances, which means that the outputs of (3) are *forecastable*.

Remark 2. It should be noted that the transformation in (5) can result in (not physically meaningful) negative values for $R(t)$ and $U(t)$ if we pick $\bar N_0$ too small. It may also result in rates ϕ larger than 1, if the initial rate $\phi(1)$ was smaller than 1 and it increased to values above that initial one. However, regardless of whether or not these estimates are "physically meaningful" the forecasts will remain unchanged. □

3. Nonlinear system identification and forecasting

We are interested in predicting the future states and outputs of a general stochastic nonlinear system of the form[1]

$$x_{t+1} = f(x_t;\theta) + d_t, \quad \forall t \in \{1,2,\dots\}, \tag{7a}$$

based on a finite set of measurements

$$y_t = g(x_t;\theta) + w_t, \quad \forall t \in \{1,2,\dots\}, \tag{7b}$$

where $x_t \in \mathbb{R}^{n_x}$ denotes the state of the system, $y_t \in \mathbb{R}^{n_y}$ the measured output, $d_t \in \mathbb{R}^{n_x}$ a stochastic disturbance, and $w_t \in \mathbb{R}^{n_y}$ a stochastic measurement noise. The system dynamics in (7a), the measurements equation in (7b), and the probability distributions of d_t and w_t depend on an unknown parameter vector θ taking value in a given set $\Theta \subset \mathbb{R}^{n_\theta}$.

Measurements are available for (past) times $t \in \{1,\dots,T\}$ for some integer $T > 0$ and our goal is to forecast the state and output for future times $t \in \{T+1,\dots,T+P\}$ for some integer $P > 0$. Towards this goal, we need to compute a maximum likelihood estimate $\hat\theta$ for θ

$$\hat\theta = \arg\max_{\theta\in\Theta} \log p_y(y_1,\dots,y_T;\theta), \tag{8}$$

where $p_y(\cdot)$ denotes the likelihood function, and then use this estimate to compute the a posteriori conditional distributions of the past and future states and future outputs, given the measurements:

$$p_{x,y^+|y}(x_1,\dots,x_{T+P},y_{T+1},\dots,y_{T+P}|y_1,\dots,y_T;\hat\theta). \tag{9}$$

[1] To shorten the formulas, in this section we denote time dependence through a subscript, as in x_t rather than $x(t)$.

This could be accomplished by first computing the estimate $\hat\theta$ that minimizes (8) and subsequently using an extended Kalman filter (or a variation of it, like an unscented Kalman filter) to compute the conditional distribution of the future states and outputs in (9). The procedure proposed here jointly computes (8) and (9), without explicitly computing the output likelihood function that appears in (8).

3.1. Maximum likelihood estimation

For nonlinear systems, it is generally hard to compute the probability distribution of the measured outputs that appears in the maximum likelihood optimization in (8). However, even though the dynamics in (7) are nonlinear, it is straightforward to compute the joint distribution of the state and output for this model under mild Markovian assumptions, as noted in the following result, which will be proved in Section 3.3.

Lemma 1. *Assume that the disturbances and noise at each time t are conditionally independent of all past disturbances and noise, given the state at time t, specifically:*

$$p_{d_t,w_t}(d_t,w_t|x_t,d_{t-1},\dots,d_1,w_{t-1},\dots,w_1;\theta) = p_{d_t,w_t}\big(d_t,w_t|x_t;\theta\big), \qquad \forall t \in \{1,2,\dots\} \tag{10}$$

where $p_{d_t,w_t}(\cdot|\cdot)$ denotes the joint conditional probability density function (pdf) of d_t and w_t. In this case, the logarithm of the joint probability density function of the output and state sequences for the model (7) is given by

$$\log p_{x,y}(y_1,\dots,y_{T+P},x_1,\dots,x_{T+P+1};\theta) = \sum_{t=1}^{T+P} \log p_{d_t,w_t}\big(x_{t+1} - f(x_t;\theta), y_t - g(x_t;\theta) \mid x_t;\theta\big). \quad \square \tag{11}$$

The *marginal distribution* of the measurements needed for maximum likelihood estimation in (8) can be obtained from the state and output *joint distribution* in (11), using

$$p_y(y_1,\dots,y_T;\theta) = \int p_{x,y}(y_1,\dots,y_{T+P},x_1,\dots,x_{T+P+1};\theta)\, dy_{T+1}\cdots dy_{T+P}dx_1\cdots dx_{T+1}, \tag{12}$$

which requires an integration of the joint distribution with respect to all the state variables and future outputs; an operation that generally cannot be done in closed form for nonlinear systems. However, the following result (proved in Section 3.3) provides a procedure to avoid this integration based on Laplace's method to approximate integrals (MacKay, 2003): Consider a measurement vector Y with a probability density function of the form

$$p_Y(Y;\theta) = \int p_{Y,Z}(Y,Z;\theta)dZ,$$

where θ is a vector of unknown parameters taking values in some set Θ and $p_{Y,Z}(Y,Z;\theta)$ is the joint distribution of Y and a latent random variable $Z \in \mathbb{R}^{n_Z}$ that needs to be integrated out.

Lemma 2. *Assume that, for every θ and Y, the conditional distribution of Z given Y is a multivariable Gaussian. Then the Hessian matrix*

$$H(Y,Z;\theta) := \frac{\partial^2 \log p_{Y,Z}(Y,Z;\theta)}{\partial Z^2} \tag{13}$$

does not depend on Z,

$$\mathrm{E}[Z|Y] = \arg\max_{Z\in\mathbb{R}^{n_Z}} \log p_{Y,Z}(Y,Z;\theta), \qquad \mathrm{CoV}[Z|Y] = -H(Y,Z;\theta)^{-1}, \tag{14}$$

and

$$\log p_Y(Y;\theta) = \frac{n_Z\log(2\pi)}{2} - \frac{\log\det\big(-H(Y,Z;\theta)\big)}{2}$$

$$+ \max_{Z\in\mathbb{R}^{n_Z}} \log p_{Y,Z}(Y,Z;\theta). \tag{15}$$

Consequently, the maximum likelihood estimator for θ can be obtained by solving

$$\arg\max_{\theta\in\Theta}\Big(-\frac{1}{2}\log\det\big(-H(Y,Z;\theta)\big) + \max_{Z\in\mathbb{R}^{n_Z}} \log p_{Y,Z}(Y,Z;\theta)\Big), \tag{16}$$

and, in view of (14), the associated minimum variance estimator of Z is

$$\hat{Z} := \arg\max_{Z\in\mathbb{R}^{n_Z}} \log p_{Y,Z}(Y,Z;\theta)$$

with associated error covariance equal to

$$\mathrm{E}[(Z-\hat{Z})(Z-\hat{Z})'|Y] = -H(Y,Z;\theta)^{-1}. \quad \square \tag{17}$$

Making the following associations

$$Y := (y_1,\dots,y_T) \in \mathbb{R}^{n_y T},$$
$$Z := (y_{T+1},\dots,y_{T+P},x_1,\dots,x_{T+P+1}) \in \mathbb{R}^{n_Z}, \quad n_Z := n_y P + n_x(T+P),$$

with the joint probability density function of these variables given by (11) in Lemma 1, the formula (16) in Lemma 2 provides a method to compute the maximum likelihood estimate for θ. As a side product, we also obtain the mean and covariance matrix of the a posteriori distribution in (9). This enable us to simultaneously obtain the maximum likelihood estimate of θ and the corresponding a posteriori distribution of the state and future output.

When the functions f and g in (7) are nonlinear and/or the disturbance and noise distributions in (10) are not multivariable Gaussian, the conditional distribution of the state given the output measurements will likely also not be Gaussian and therefore (15) should be taken as an approximation of the true marginal distribution. We shall see in the proof of Lemma 2 that the Gaussian assumption is used to justify truncating the Taylor series of $Z \mapsto \log p_{z|y}(Z|Y;\theta)$ at its second term since this function is quadratic for Gaussian distributions. For non-Gaussian distributions, this truncation will introduce an error, but it is possible to establish a bound on this error. This is due to the fact that the Taylor series of the joint distribution $Z \mapsto \log p_{Y,Z}(Y,Z)$ and the conditional distribution $Z \mapsto \log p_{z|y}(Z|Y)$ have exactly the same terms and we have an explicit formula (11) for the joint distribution (see Remark 4 in Section 3.3.).

Remark 3 (*Numerical Issues Due to the Lack of Identifiability*)**.** When trying to apply Lemma 2 to models that are *not identifiable*, in the sense that multiple realizations for Z lead to the same value of the joint distribution $p_{Y,Z}(Y,Z;\theta)$, the Hessian matrix $H(Y,Z;\theta)$ may be singular, leading to "infinite" error covariance in (17).

For poorly identifiable models, the matrix $H(Y,Z;\theta)$ may be non-singular, but with very small singular values, making the log-determinant in (16) strongly negative and often causing numerical issues. To avoid this, one can replace the optimization in (16) by

$$\arg\max_{\theta\in\Theta}\Big(-\frac{1}{2}\log\det\big(\epsilon I - H(Y,Z;\theta)\big) + \max_{Z\in\mathbb{R}^{n_Z}} \log p_{Y,Z}(Y,Z;\theta)\Big), \tag{18}$$

for a small constant $\epsilon > 0$. The addition of the constant term ϵI will have little affect on the estimate of θ, as long as this variable does not affect significantly the kernel of $H(Y,Z;\theta)$, which can be numerically verified. $\square$

3.2. A scalable iterative algorithm

For the problem at hand, (16) involves a joint optimization with respect to n_θ parameters and $n_Z := n_y P + n_x(T+P)$ state variables. For large time horizons $T+P$, the computation complexity of such optimization can be greatly reduced by using an iterative block coordinate ascent algorithm:

Algorithm 1. Given a measurement vector Y and a tolerance ϵ_{tol}, the following algorithm returns estimates $\hat{\theta}_k$ and $\hat{Z}_k$:

1. Pick initial values for $\hat{\theta}_0$, $\hat{Z}_0$ and set $k = 1$.
2. Update estimates using:

$$\hat{Z}_k := \arg\max_{Z\in\mathbb{R}^{n_Z}} \log p_{Y,Z}(Y,Z;\hat{\theta}_{k-1}) \tag{19}$$

$$\hat{\theta}_k := \arg\max_{\theta\in\Theta}\Big(-\frac{1}{2}\log\det\Big(-H(Y,\hat{Z}_k;\theta)\Big) + \log p_{Y,Z}(Y,\hat{Z}_k;\theta)\Big) \tag{20}$$

3. Increment k and go back to step 2 until $\|\hat{\theta}_k - \hat{\theta}_{k-1}\| \leqslant \epsilon_{\text{tol}}$.

In general, block coordinate ascent/descent algorithms are not guaranteed to terminate (Bertsekas, 1999). However, upon a successful termination, the pair $(\hat{Z}_k,\hat{\theta}_{k-1})$ satisfies the first order optimality conditions of the optimization in (19) and the pair $(\hat{Z}_k,\hat{\theta}_k)$ satisfies the first order optimality conditions of the optimization in (20). Since the first order optimality conditions of (16) are precisely the union of the first order optimality condition of (19) and (20), we conclude that the pair $(\hat{Z}_k,\hat{\theta}_k)$ satisfies the first order optimality conditions of (16) up to the ϵ_{tol} discrepancy between $\hat{\theta}_{k-1}$ and $\hat{\theta}_k$. In general, this does not guarantee that Algorithm 1 will find a global maximum of the likelihood function, but it does guarantees that termination can only take place at a local maximum (up to the ϵ_{tol} error). In practice, we have observed that constraining the parameter values θ and states/outputs Z to physically meaningful sets consistently leads to the same optimum regardless of how we initialize the numerical solvers.

The key advantage of the iterative approach in Algorithm 1 with respect to solving the single-shot full optimization in (16) is that (i) the number n_Z of optimization variables in (19) scales linearly with the number of time instants of interest and (ii) the number of optimization variables in (20) is equal to the number n_θ of parameters, *regardless of the horizon length*. Furthermore, while the total number of entries of the Hessian matrix (13) scales quadratically with n_Z, the number of *non-zero* entries of this matrix only scales linearly with n_Z. This is because, the structure of (11) leads to

$$\frac{\partial^2 \log p_{Y,Z}(Y,Z;\hat{\theta}_{k-1})}{\partial x_t \partial x_{t+k}} = 0,$$

for every $k \notin \{-1,1,0\}$. In practice, this means that we can use second order Newton methods to solve (19) with computation times that only grow linearly with n_Z, provided that we compute the Newton direction using sparse solvers for linear equations (Davis, Gilbert, Larimore, & Ng, 2004; Hespanha, 2017). Regarding the optimization in (20), while the number of optimization variables is typically small and independent of the time horizon length, one still needs to compute the log-determinant of a large matrix. Also here, we can explore the sparsity of (13) by performing a sparse LDL factorization and obtain the determinant by simply multiplying the entries of the diagonal matrix, or adding their logarithms to directly obtain the log-determinant of the matrix, which is numerically much more stable.

3.3. Proof of Lemmas 1 and 2

Proof of Lemma 1. For each $t \geqslant 1$, we can expand the joint probability density function of the state up to time $t+1$ and measurements up to time t as

$$p(y_1,\dots,y_t,x_1,\dots,x_{t+1}) = p(x_{t+1},y_t \mid x_1,\dots,x_t,y_1,\dots,y_{t-1}) \\ p(y_1,\dots,y_{t-1},x_1,\dots,x_t), \quad \forall t \geqslant 0,$$

where, for simplicity of notation, we omitted all dependencies on the parameter vector θ. In view of (7) and the independence assumption (10), we have that

$$\begin{aligned} &p(x_{t+1},y_t \mid x_1,\dots,x_t,y_1,\dots,y_{t-1}) \\ &\quad = p_{d_t,w_t}\big(x_{t+1} - f(x_t;\theta), y_t - g(x_t;\theta) \mid x_1,\dots,x_t,y_1,\dots,y_{t-1}\big) \\ &\quad = p_{d_t,w_t}\big(x_{t+1} - f(x_t;\theta), y_t - g(x_t;\theta) \mid x_t\big), \end{aligned} \tag{21}$$

and therefore

$$p(y_1,\dots,y_t,x_1,\dots,x_{t+1}) = p_{d_t,w_t}\big(x_{t+1}-f(x_t;\theta),y_t-g(x_t;\theta)\mid x_t\big)p(y_1,\dots,y_{t-1},x_1,\dots,x_t).$$

Iterating this from $t=1$ to $t=T+P$ leads to

$$p(y_1,\dots,y_{T+P},x_1,\dots,x_{T+P+1}) = \prod_{t=1}^{T+P} p_{d_t,w_t}\big(x_{t+1}-f(x_t;\theta),y_t-g(x_t;\theta)\mid x_t\big),$$

from which (11) follows by taking logarithms. ■

Proof of Lemma 2. Denoting the conditional distribution of Z given Y by

$$p_{Z|Y}(Z|Y;\theta) := \frac{p_{Y,Z}(Y,Z;\theta)}{p_Y(Y;\theta)}. \tag{22}$$

we have that

$$\log p_Y(Y;\theta)+\log p_{Z|Y}(Z|Y;\theta) = \log p_{Y,Z}(Y,Z;\theta) \tag{23}$$

and therefore

$$\frac{\partial^2 \log p_{Z|Y}(Z|Y;\theta)}{\partial Z^2} = \frac{\partial^2 \log p_{Y,Z}(Y,Z;\theta)}{\partial Z^2}, \quad \forall Z, \tag{24a}$$

$$\log p_Y(Y;\theta)+\max_Z \log p_{Z|Y}(Z|Y;\theta) = \max_Z \log p_{Y,Z}(Y,Z;\theta). \tag{24b}$$

Moreover, since $p_{Z|Y}(Z|Y;\theta)$ is a multivariable Gaussian, we must have

$$\log p_{Z|Y}(Z|Y;\theta) = -\frac{n_Z}{2}\log(2\pi) + \frac{1}{2}\log\det\Sigma_{Z|Y}^{-1} - \frac{1}{2}(Z-\mu_{Z|Y})'\Sigma_{Z|Y}^{-1}(Z-\mu_{Z|Y}), \tag{25}$$

where $\mu_{Z|Y}$ and $\Sigma_{Z|Y}$ denote its mean and covariance matrix. This allows us to conclude that

$$\arg\max_Z \log p_{Z|Y}(Z|Y;\theta) = \mu_{Z|Y}, \tag{26a}$$

$$\frac{\partial^2 \log p_{Z|Y}(Z|Y;\theta)}{\partial Z^2} = -\Sigma_{Z|Y}^{-1}, \tag{26b}$$

$$\max_Z \log p_{Z|Y}(Z|Y;\theta) = -\frac{n_Z}{2}\log(2\pi)+\frac{1}{2}\log\det\Sigma_{Z|Y}^{-1} = -\frac{n_Z}{2}\log(2\pi) + \frac{1}{2}\log\det\Big(-\frac{\partial^2 \log p_{Z|Y}(Z|Y;\theta)}{\partial Z^2}\Big). \tag{26c}$$

We conclude from (24a) and (26b) that

$$\frac{\partial^2 \log p_{Y,Z}(Y,Z;\theta)}{\partial Z^2} = \frac{\partial^2 \log p_{Z|Y}(Z|Y;\theta)}{\partial Z^2} = -\Sigma_{Z|Y}^{-1}, \quad \forall Z,$$

does not depend on Z, and then from (24b) and (26c) that

$$\log p_Y(Y;\theta) = -\max_Z \log p_{Z|Y}(Z|Y;\theta)+\max_Z \log p_{Y,Z}(Y,Z;\theta) = \frac{n_Z}{2}\log(2\pi) - \frac{1}{2}\log\det\Big(-\frac{\partial^2 \log p_{Y,Z}(Y,Z;\theta)}{\partial Z^2}\Big) + \max_Z \log p_{Y,Z}(Y,Z;\theta),$$

from which the result follows. ■

Remark 4 (*Non-Gaussian a Posteriori*)**.** When the a posteriori distribution is not a multivariable Gaussian, one can view (25) as a second order truncation of the Taylor series of $Z \mapsto \log p_{Z|Y}(Z|Y;\theta)$, which means that this formula will have an error due to higher order terms in the series. In fact, such a truncation of this Taylor series is at the basis for the Laplace method to approximate integrals (MacKay, 2003). In view of (23), when $Z \mapsto \log p_{Z|Y}(Z|Y;\theta)$ and $Z \mapsto \log p_{Y,Z}(Y,Z;\theta)$ are analytic, both functions have the same Taylor series terms of order higher than 0 and therefore one can estimate the error in the truncation (25) by computing the terms in the Taylor series of $Z \mapsto \log p_{Y,Z}(Y,Z;\theta)$ of order higher than 2. □

4. Numerical results

We now summarize the results obtained by applying the identification/forecasting procedure outlined in Section 3 to the SIR stochastic model in Section 2. The measurements used include time series of COVID-19 daily cases and daily deaths obtained from the following sources:

1. the European Center for Disease Prevention and Control (ECDPC) (European Center for Disease Prevention and Control, 0000) for worldwide data outside the United States of America, Brazil, and Portugal;
2. the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, 2020) for State and City data within the United States of America;
3. the Portuguese Direção-Geral de Saúde for data in Portugal (Portuguese Direção-Geral de Saúde, 2020).
4. Brasil.IO based on data collected from the Secretarias Estaduais de Saúde for State and City data within Brazil (CoronaCidades.org, 0000).

We produced forecasts for every country represented in the ECDPC dataset that, at some point in time, reported more than 10 new COVID-19 cases in one day. For a few countries, including the United States of America and Brazil, we had available data at the state/province/city level and produced forecasts at finer regional levels. The ECDPC and CSSE time series go as far back as February 2020 and contain daily results up to the present. Our full set of results is available at (COVID-19, 2020) and includes 144 countries and regions around the world, but here we only show a subset of results for Italy, United Kingdom, Germany Portugal, Japan, India, and the US states of New York, California, Texas, Illinois, and Montana. This selection covers a representative set of countries/regions in terms of population size and density, timing and scope of nonpharmaceutical measures, population behavior, etc.

For a very large number of countries/regions, the time series with the daily number of cases exhibit large weekly variations, typically with a smaller number of cases reported during the weekends. In fact, for a few countries/time-periods no new cases/deaths are reported in the weekend and a very large number of cases are reported every Monday,[2] leading to a larger delay between the time a patient becomes infective and the case is reported. To remove this day-of-the-week effect, all our forecasts were computed weekly with data up to the latest Wednesday. In addition, all time series were pre-filtered with a 7-day moving average filter that tries to equalize the delay between infectiveness/death and reporting. For consistency, we have done this for every country/region, regardless of whether or not the data showed day-of-the-week effects. For countries/regions where the weekend effect is negligible or where this effect only appeared during a fairly brief period of time, the introduction of 7-day averaging makes little difference in terms of the forecasts. But for countries where this effect has been persistent, removing the averaging inflates the estimates of the noise/disturbances since this effect has to be explained through the stochastic elements of the model.

4.1. Methodology

All estimates and forecasts reported in this section were obtained using the Algorithm 1 with the vector Y containing daily measurements of new cases $y_v(t)$ and deaths $y_D(t)$ over a given time range $t \in \{1,2,\dots,T\}$. The latent random variable Z contains the full state $R(t), U(t), \beta(t), \phi(t), \omega(t)$ of the model (3)–(4) over an extended time range $t \in \{1,2,\dots,T+P\}$ that includes forecasts up to $P = 21$ days

[2] See, e.g., ECDPC dataset for France during the month of July 2020.

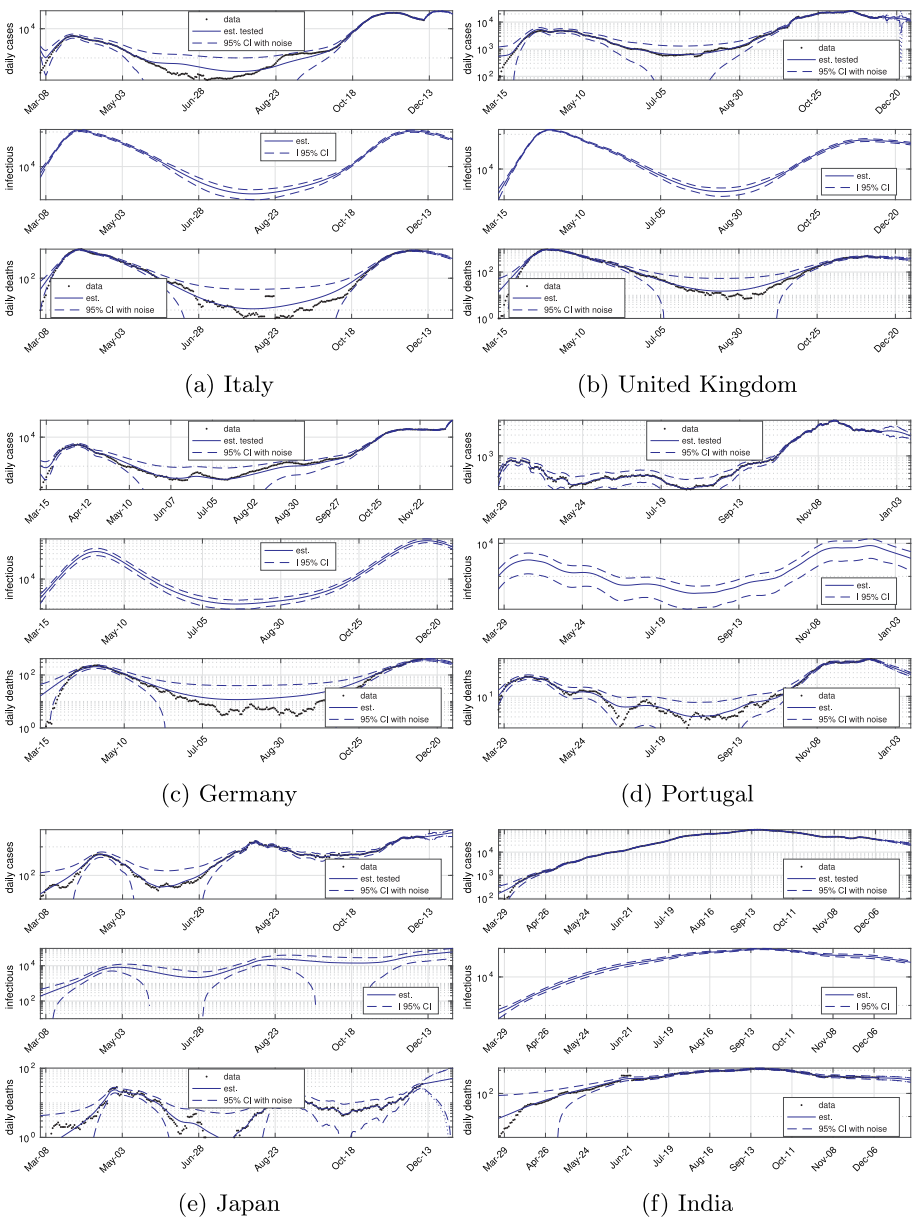


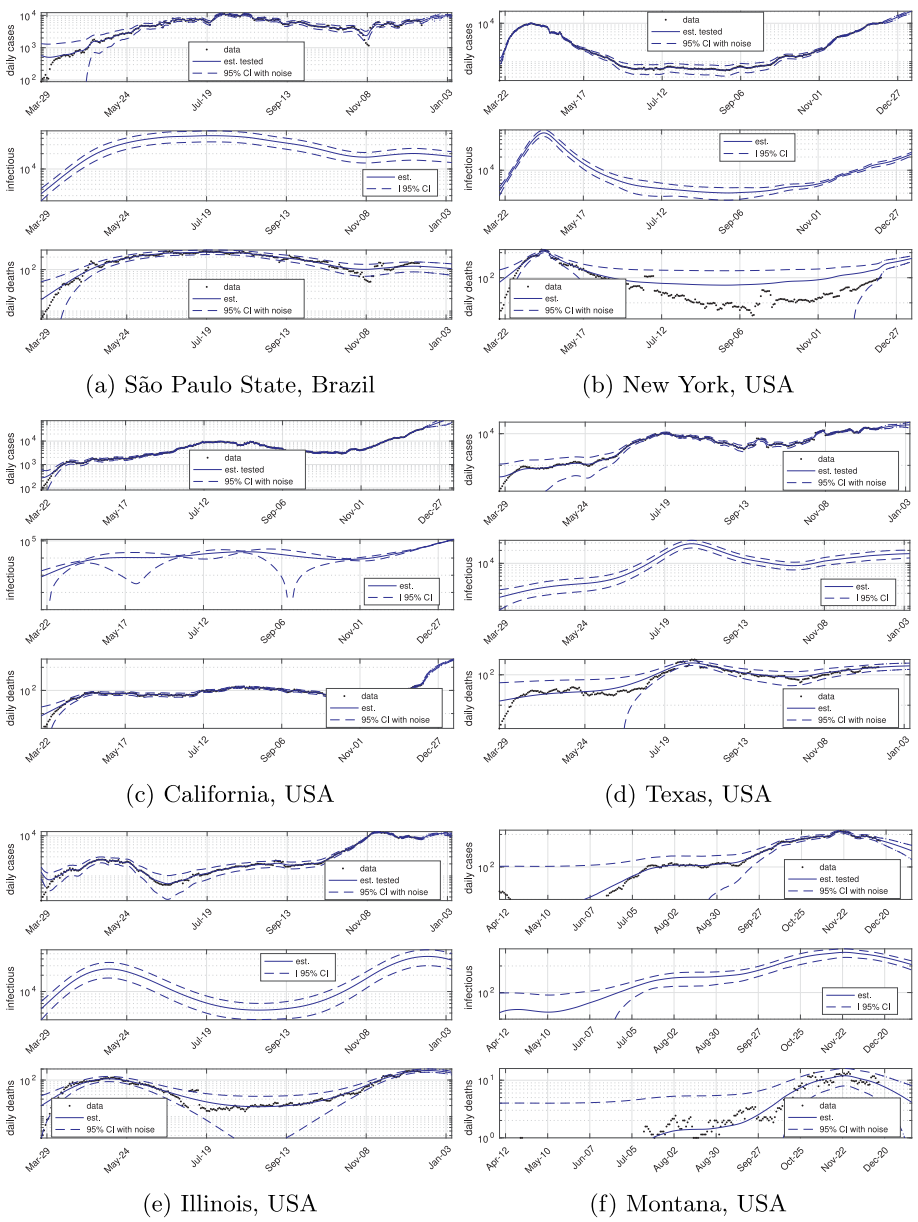
Fig. 1. State estimates and forecasts for the daily numbers of new cases and deaths for the model (3)–(4), based on data available up to mid-to-late December (depending on the source of data). The forecasts extend for 3 weeks past the available data. In all plots, the dots correspond to the daily measurements of new cases and deaths, the solid lines to a posteriori state estimates, and the dashed lines to 95% confidence intervals.

into the future; the removal rate γ; and measurements forecasts for new cases $y_v(t)$ and deaths $y_D(t)$ over $t \in \{T+1, \dots, T+P\}$. The parameter vector θ includes the variances of the state disturbances $d_\rho(t)$, $d_v(t)$; the variances of the parameter increments $d_\beta(t)$, $d_\omega(t)$, $d_\phi(t)$; and the variances of the noises $w_v(t)$ and $w_D(t)$. All 95% confidence intervals reported are based on the a posteriori covariance matrix computed using (17), which we marginalize for the different variables to obtain a posteriori standard deviations.

For essentially all datasets, we observed that Algorithm 1 converged to estimates corresponding to zero variances for the disturbances $d_\rho(t)$, $d_v(t)$ (or to negligible values) so we eventually removed those parameters from θ and set them to fixed values that were sufficiently low not to affect any of the other estimates. This improved numerical conditioning, because very small disturbance variances result in poorly conditioned Hessian matrices $H(Y, \hat{Z}_k; \theta)$ in (20).

The optimizations in (19)–(20) were carried out by primal–dual interior-point solvers built using the `TensCalc` toolbox (Hespanha, 2017). The solvers generated by `TensCalc` explore sparsity of the Hessian matrix, resulting in computation times that scale linearly with the horizon length (see Section 3.2). Algorithm 1 was initialized with a rough state estimated obtained as follows:

1. The new-cases reporting rate was initialized at $\phi(t) = 1$, $\forall t \geqslant 1$. Since $\phi(1)$ is not identifiable (see Section 3), the initial rate $\phi(1)$ was fixed at 1 and only subsequent $\phi(t)$, $t > 1$ were optimized.
2. The removal rate was initialized at the (somewhat arbitrary) value $\gamma = 1/21$ (21 days time constant).
3. The initialization for the state $U(t)$ was obtained by neglecting noise and disturbances from (3b)–(3c), which leads
$$U(t+1) = U(t) + y_v(t) \quad \Rightarrow \quad U(t) = \sum_{\tau<t} y_v(t).$$
4. The number of infections at day $t = 1$ was initialized to be equal to the total number of infections reported before that date:
$$I(1) = \sum_{\tau<1} y_v(\tau).$$



(a) São Paulo State, Brazil (b) New York, USA

(c) California, USA (d) Texas, USA

(e) Illinois, USA (f) Montana, USA

Fig. 2. Continuation from Fig. 1.

Subsequent values for the initialization of $I(t)$ can be obtained by neglecting noise and disturbances from (3a)–(3c), which leads to

$$I(t+1) = -\gamma I(t) + y_v(t).$$

The initialization for $R(t)$ can then be obtained from $R(t) = U(t) - I(t)$, $\forall t \geqslant 1$.

5. Assuming a constant value for the infection rate β and ignoring disturbances, (3b) can be written as

$$U(t+1) = U(t) + \beta I(t) - \frac{\beta}{N_0} I(t)U(t), \qquad \forall t \geqslant 1,$$

from which initializations for β and β/N_0 can be obtained through a least-squares linear fit. Since the precise value for N_0 is not identifiable (see Section 3), we took this value to be "correct".

6. The deaths reporting rate was initialized with a constant value ω determined from (3d) through a least-squares linear fit.

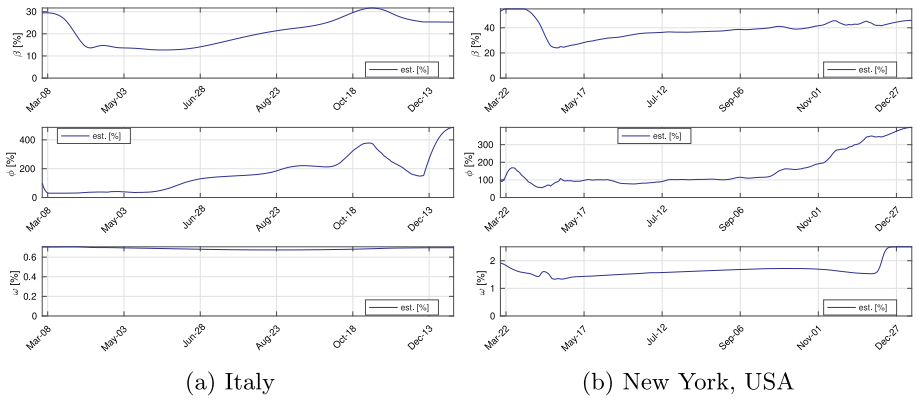
The same initialization process described above was applied to all countries/regions and time intervals in our dataset, including the alternative models discussed in Section 4.4. This initialization resulted in very few failures of the nonlinear solver across the whole dataset.

As noted in Remark 3, to avoid numerical issues we replaced (16) in Algorithm 1 by (18) with $\epsilon = 10^{-4}$. In the interest of time, we have also limited the number of iterations to 30.

4.2. Forecasts based on the entire datasets

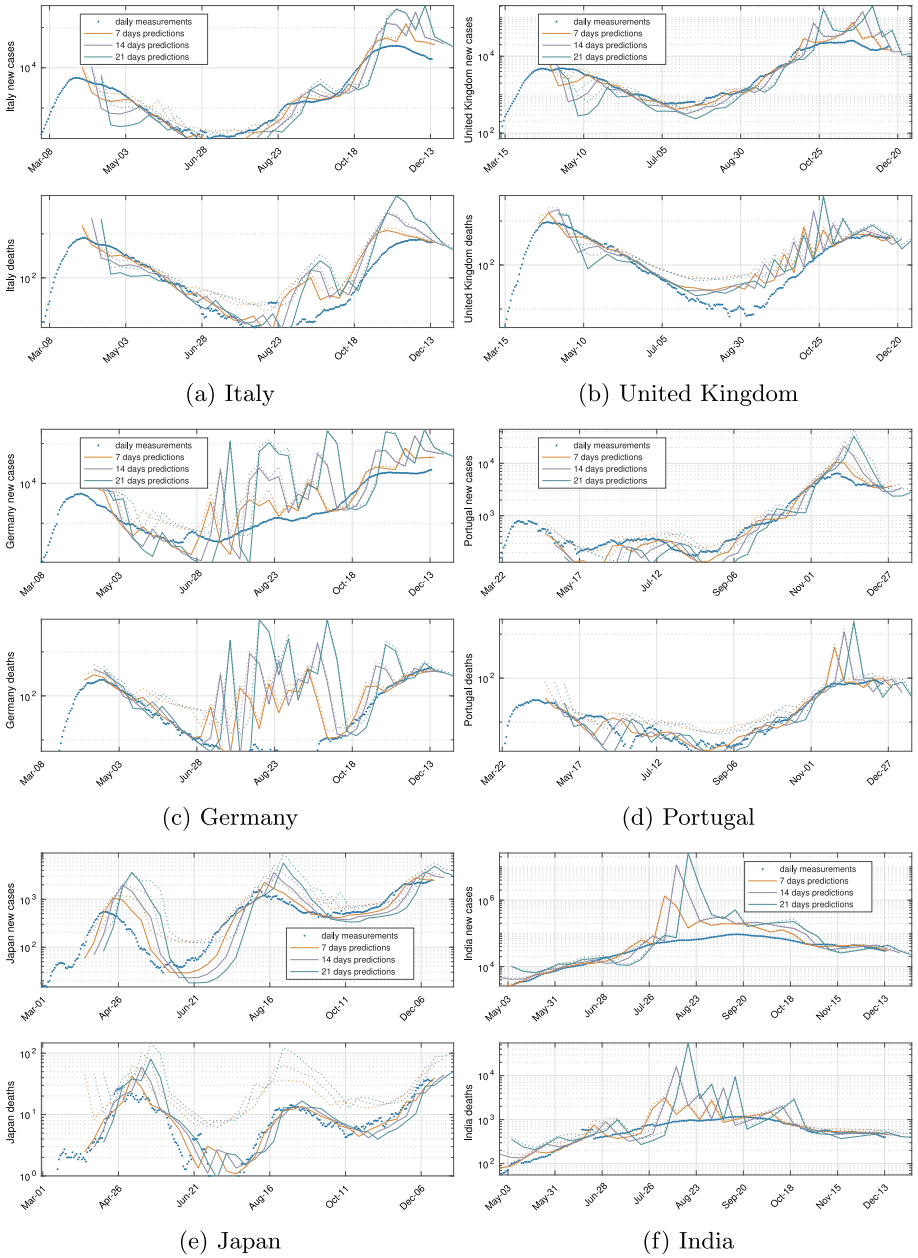
Figs. 1–2 show state estimates and forecasts based on the largest window of daily data available at the writing of this paper for Italy, Portugal, Japan and the US states of New York, California, and Montana. At this time, most of these regions are experiencing a strong resurgence in the number of cases but with fairly distinct progressions since early March: In Italy, the United Kingdom, Germany, Portugal, the state of São Paulo, and the US states of New York, California, Texas, and Illinois we see two clearly defined waves; Japan seems to be in the middle of a third wave; Montana's number of cases seems to evolve in a step-like fashion; and India is not (yet?) showing a clearly defined second wave.

In several countries we see that the second wave exceeds the magnitude of the first wave in terms of the number of new cases, but



(a) Italy (b) New York, USA

Fig. 3. Estimates of the time-varying parameters for the model (3)–(4), based on data available up to mid December. The forecasts extend for 3 weeks past the available data.



(a) Italy (b) United Kingdom

(c) Germany (d) Portugal

(e) Japan (f) India

Fig. 4. Running forecasts for the daily numbers of new cases and deaths for the model (3)–(4), based on data available up to 7, 14, and 21 days prior to the forecast. The solid lines depict the forecasts, whereas the dashed lines of the same color depict the corresponding 95% confidence intervals. To keep the plots less cluttered, only the upper bound of the confidence interval is plotted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

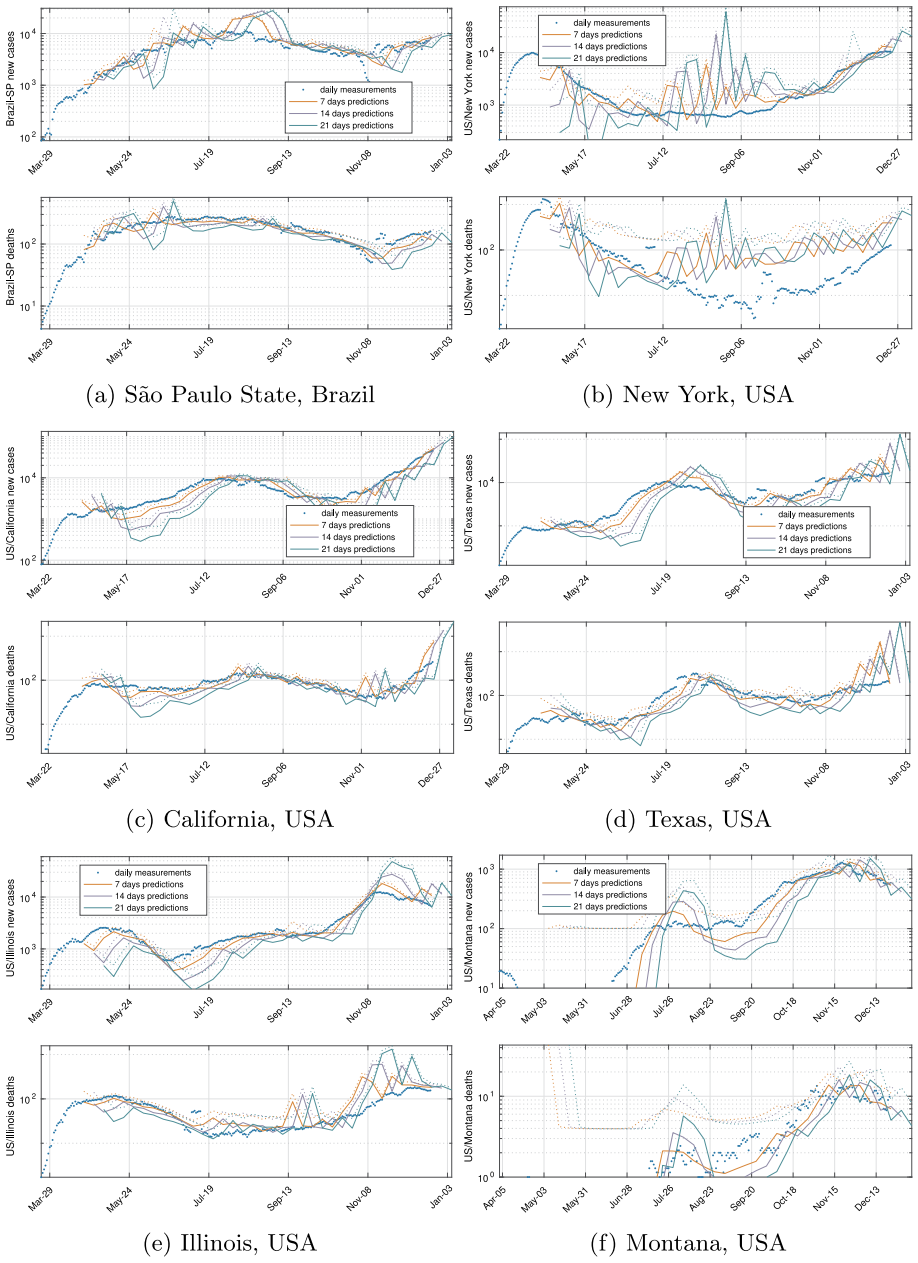


Fig. 5. Continuation from Fig. 4.

not in terms of daily deaths. The discrepancy between the number of new cases and deaths, which is particularly extreme in Italy, the United Kingdom, Germany, Portugal, Japan, New York, and Illinois is "explained" by the model through changes in the new-cases and deaths reporting rates. This can be confirmed by the plots in Fig. 3, which shows the estimates of the different time-varying parameters. However, as noted in Section 2.3, these parameters are fundamentally not identifiable and can, at most be estimated up to the state/input transformation in (5). The plots presented here were obtained by resolving the model ambiguity as discussed in Section 4.1. As noted in Remark 2, by setting the initial new-cases reporting rate ϕ to 100%, an increase in the rate results in values for ϕ larger than 100%, which obviously means that the original rate really started at some value below 100%, but the data available does not permit estimating the precise value of the rate in absolute terms.

Because the time series exhibit large variation, the number of cases in the y-axis are plotted in a logarithmic scale. While this permits a better visualization of the data at multiple scales, it somewhat distorts the confidence intervals, which visually appear much larger when the number of cases is smaller. For example in Italy, the confidence interval for the daily number of deaths in August is roughly [1,40] and in early December it is roughly [695–770], while the latter is twice as wide as the former, it appears far smaller with the logarithmic scale used in Fig. 1.

4.3. Running forecasts

A large effort was devoted to validate the methodology used to create forecasts for the daily numbers of new cases and deaths: Starting from the first Wednesday for which we had 21 days of past data, we performed system identification for the stochastic SIR model and computed forecasts for 7, 14, and 21 days ahead; which can eventually be compared with the actual values. We repeated this procedure for every subsequent Wednesday, resulting in updated forecasts. To obtain true validation, the model identified with data up to a particular Wednesday, does not use any subsequent data, either as measurements

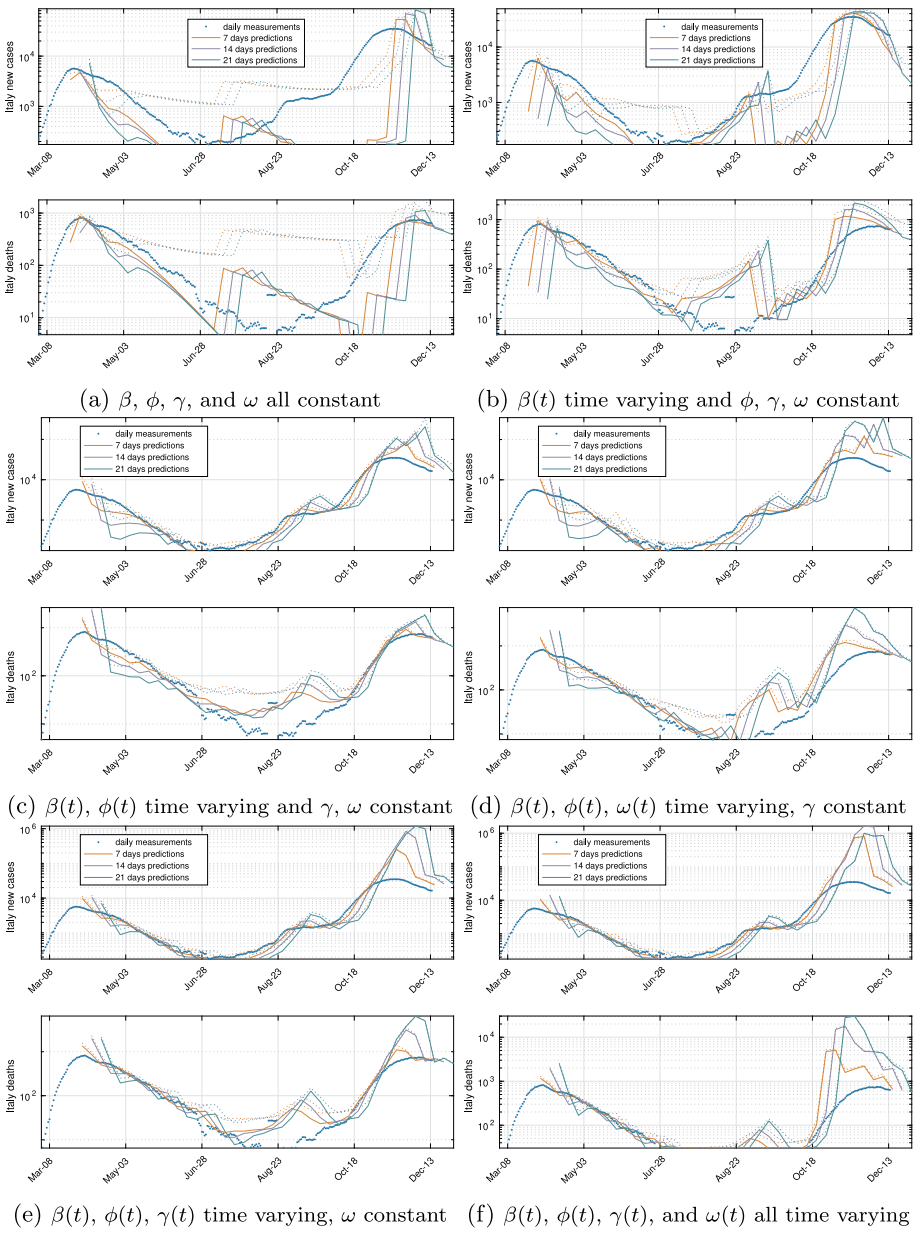


Fig. 6. Italy running forecasts for the daily numbers of new cases and deaths based on data available up to 7, 14, and 21 days prior to the forecast. The different plots corresponds to distinct assumptions regarding which parameters of the model (3)–(4) are allowed to vary with time.

or to help initialize the numerical solver. This procedure was carried out for every country/region in our datasets and for every week for which we had, at least, 21 days of data, resulting in over 4500 identification/forecasting experiments.

Figs. 4–5 shows the results obtained for the same countries/states shown in Figs. 1–2. It is important to emphasize that in Figs. 1–2, forecasts only appear in the 3-weeks at the right-hand side of the plots. Prior to that, the solid and dash lines correspond to estimates and confidence intervals for (past) state/measurements that were computed using the entire datasets. In contrast, every point in the solid lines in Figs. 4–5 corresponds to a forecast that was compute 1, 2, or 3 weeks before. The corresponding dashed lines show the upper bounds of the associated 95% confidence intervals. Several important conclusions can be drawn from these plots:

1. The initial forecasts (computed just with 21 days of measurements) vary greatly in accuracy and often come associated with very wide confidence intervals, reflecting the fact that 21 days of measurements do not provide enough information to obtain reliable estimates. However, by the time 4–5 weeks of data are available, the confidence intervals start to become much tighter.
2. The actual numbers of daily deaths, generally fall inside the 95% confidence intervals computed 7, 14, and 21 days before, showing that the model is especially reliable in predicting disease-related casualties. For the countries/states shown, the main exception can bee seen in California in late May, India in mid-June, Texas in mid-July, and São Paulo in late November.
3. Most of the actual numbers of new cases also fall inside the 95% confidence intervals computed 7, 14, and 21 days before, but we can see more exceptions to this "rule". For example, this can be seen in Italy and the United Kingdom in late April and mid-October; as well as in California and Texas in late May and through most of June.

   The larger difficulty in predicting the daily number of new cases rather than predicting the number of daily deaths is not surprising in view of the fact that the former is highly dependent

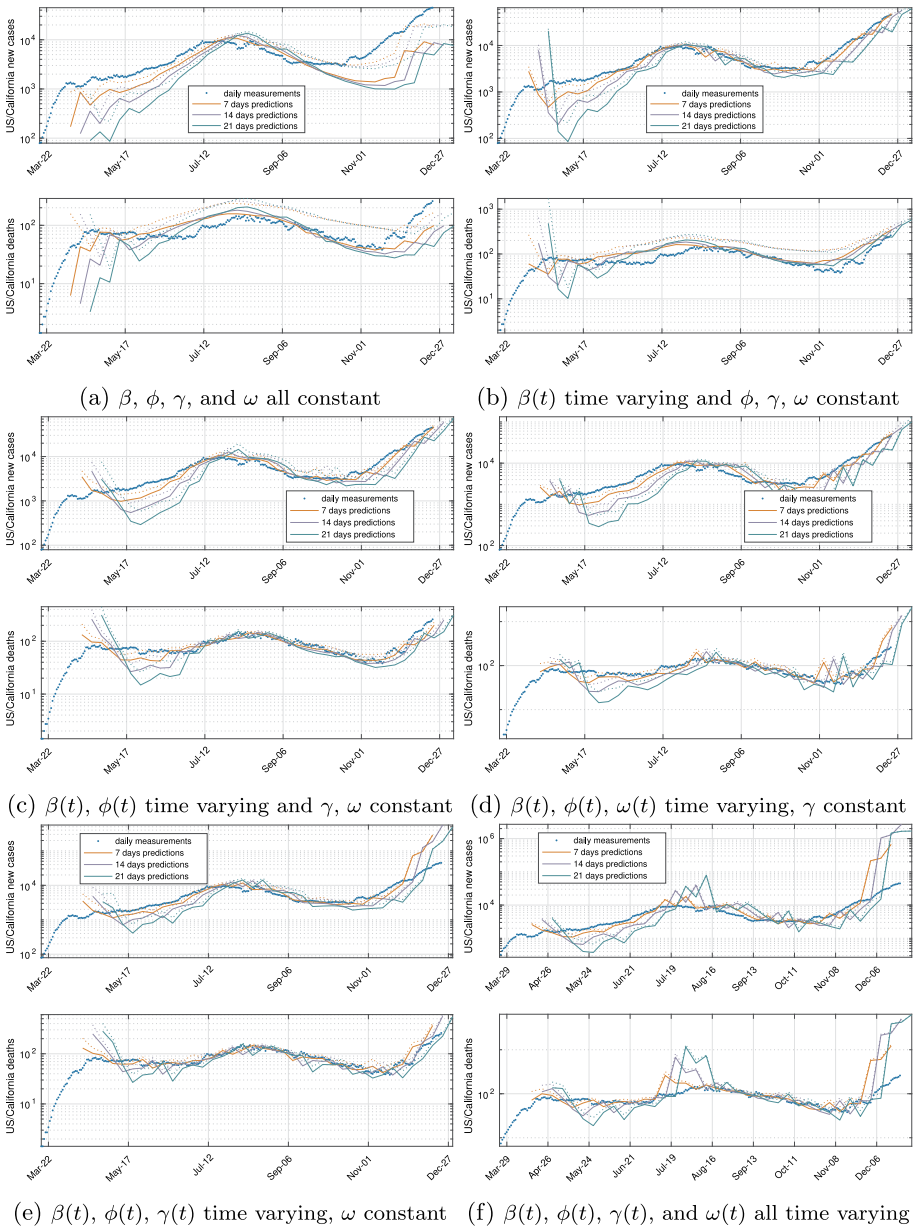


Fig. 7. California, USA running forecasts for the daily numbers of new cases and deaths based on data available up to 7, 14, and 21 days prior to the forecast. The different plots corresponds to distinct assumptions regarding which parameters of the model (3)–(4) are allowed to vary with time.

on the infection and new-cases reporting rates, which can exhibit abrupt changes due to the so-called super-spreader events (medicalxpress.com, 2020) or changes in testing policies and/or test availability.

4. The estimates produced are fairly robust to outliers that appear in essentially all datasets. Due to the use of a 7-day moving average filter, each single-day outlier results in 7 consecutive measurements that stand out from the adjacent data points. Notable examples of such outliers appear in the number of deaths in Italy in August 15–21 and in New York in June 29–July 5 (among others). Typically, such outliers are caused by an agency adding to their daily report a number of past deaths/cases that had occurred over some past period of time, but had been neglected in previous daily reports (see Nov. 4th report, Portuguese Direção-Geral de Saúde, 2020). Even though some outlier are clearly noticeable, we opted not to discard any of these data points because they could potentially be caused by super-spreader events. In practice, the existence of these outliers causes some bias in the estimates, but they generally do not leads to data points outside the 95% confidence intervals.
5. For a few regions/periods we see large fluctuations in the forecasts from one week to the next. This is especially noticeable in the United Kingdom in September and October, in Germany since early July until December, in Portugal in early December, in India in August, in New York from mid-July until mid-December, and in Illinois from mid-September until December. We shall return to these fluctuation in Section 4.4.

4.4. Alternative models

The general model in (3)–(4) allows the infection rate $\beta(t)$, the new-cases reporting rate $\phi(t)$, and the deaths reporting rate $\omega(t)$ to be time varying with increments determined by the zero-mean random processes $d_\beta(t), d_\omega(t), d_\phi(t)$ in (4). A special instance of this model includes the case in which the zero-mean increment processes have zero

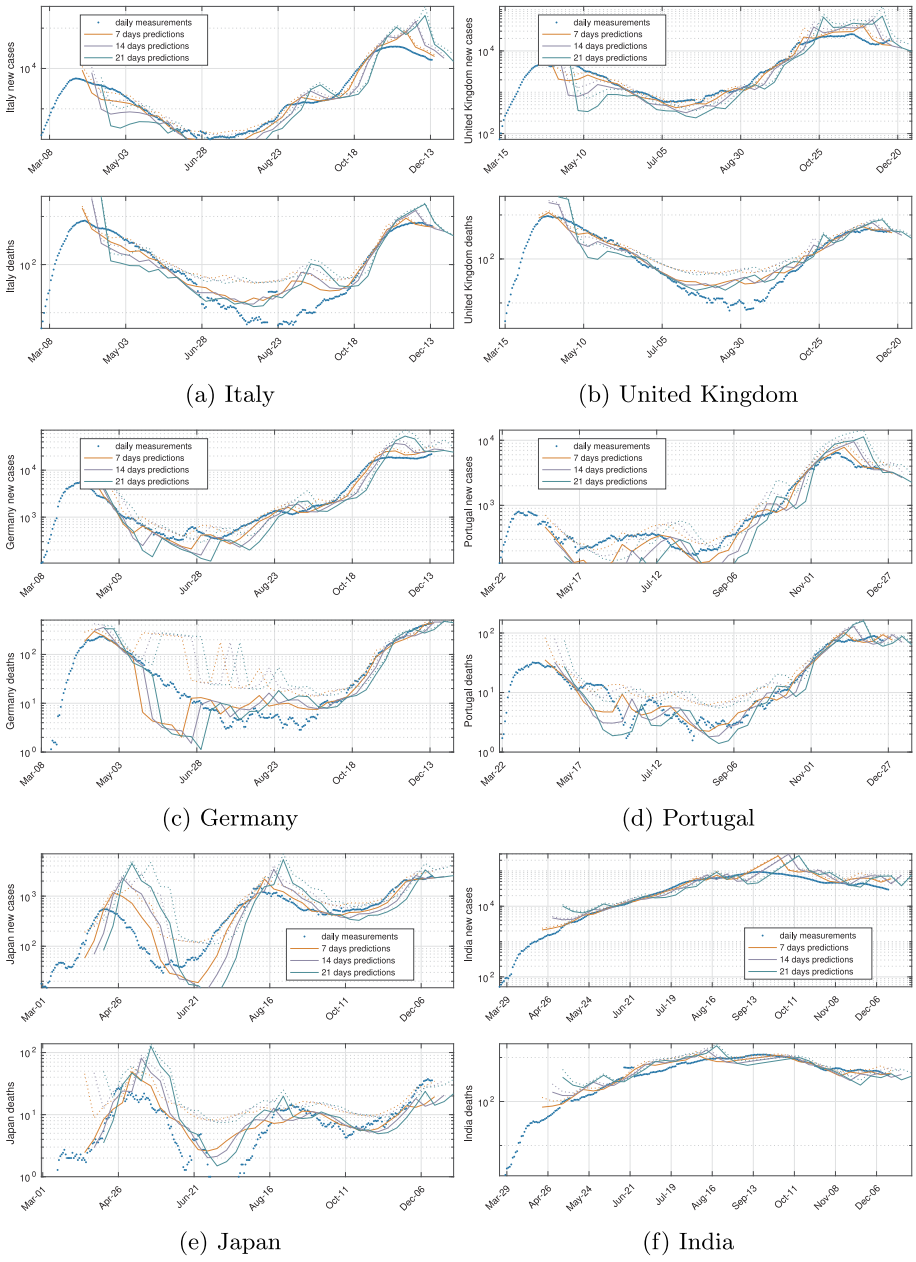


Fig. 8. Running forecasts for the daily numbers of new cases and deaths for the model (3)–(4), based on data available up to 7, 14, and 21 days prior to the forecast. The solid lines depict the forecasts, whereas the dashed lines of the same color depict the corresponding 95% confidence intervals. To keep the plots less cluttered, only the upper bound of the confidence interval is plotted. These plots differ from those in Fig. 4 in that here the death rate ω was assume constant. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

variance and it is instructive to see how the forecasts would change if we were to consider constant rather than time-varying values for these parameters. In addition, we could also question the assumption of keeping the removal rate γ constant, rather than a time varying parameter, also with increments determined by a zero-mean random process.

Different combinations are possible for which parameters are allowed to vary and which should remain constant. Figs. 6–7 show a comparison of forecasts for Italy and the US state of California for the following 6 possibilities:

(a) all parameters constant;
(b) a variable infection rate $\beta(t)$, but constant removal γ, new-cases reporting rate ϕ and deaths reporting rates ω;
(c) variable infection rate $\beta(t)$, new-cases reporting rate $\phi(t)$, deaths reporting rate $\omega(t)$, but constant removal rate γ;
(d) variable infection rate $\beta(t)$ and new-cases reporting rate $\phi(t)$, but constant removal rate γ and deaths reporting rate ω;
(e) variable infection rate $\beta(t)$, removal rate $\gamma(t)$, new-cases reporting rate $\phi(t)$, but constant deaths reporting rate ω; and
(f) all parameters time varying.

The selection of the combinations above is motivated by the widely accepted observation that the infection rate has varied greatly over time, and therefore we take it as time-varying in all but the first option. In addition, it is also well known that the testing rate for asymptomatic patients has varied greatly, so we only take the new-cases reporting rate as constant in the first 2 options. Variability on the remaining parameters is likely, but almost certainly of a smaller magnitude.

A comparison of plots like the ones in Figs. 6–7 for a large number of countries/regions yields the following observations:

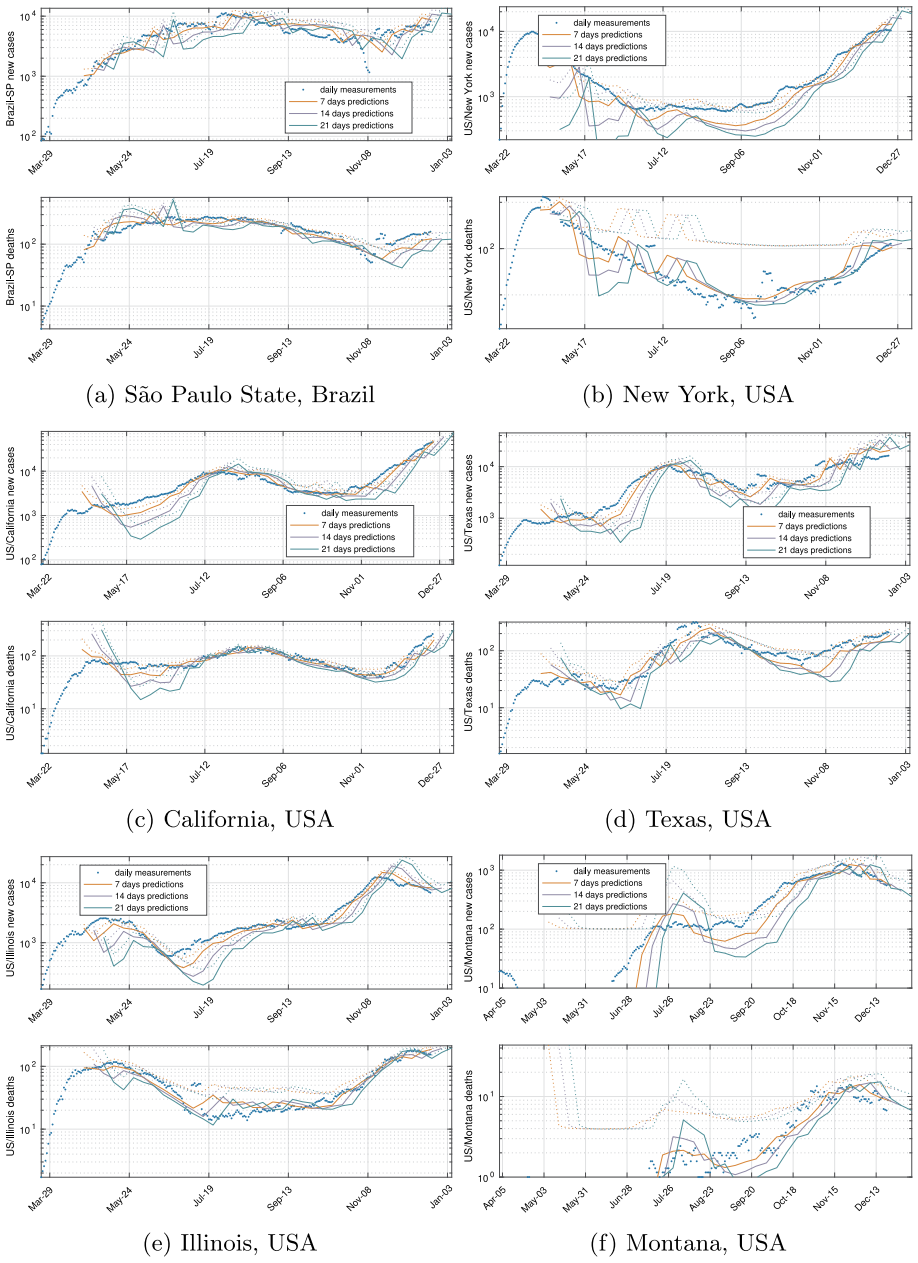


Fig. 9. Continuation from Fig. 8.

1. In general, assuming that all parameters are constant or that all but the infection rate $\beta(t)$ are constant severally compromises the models' ability to produce useful forecasts.
   For the constant parameters model, in every dataset that exhibits a second wave, the numbers of new cases during the second wave fall grossly outside the confidence intervals associated with the forecasts. This is not surprising in view of the fact that a constant-parameter SIR model cannot exhibit multiple waves.
   When only the infection rate is allowed to be time-varying, it is possible to "explain" the measurements, but this typically requires large levels of noise and/or a highly variable infection rate, which typically result in confidence intervals that are larger than those obtained with richer models.
2. For a large number of countries/regions, the results obtained assuming that all parameters are time varying differ little from those assuming that the infection rate $\beta(t)$ and the new-cases reporting rate $\phi(t)$ are the only time-varying parameters. This is generally true both for the point forecasts and the confidence intervals.
   For a more detailed comparison of the original model (3)–(4), with the simplified model that considers a constant removal rate γ and deaths reporting rate ω, but time varying infection $\beta(t)$ and new-cases reporting $\phi(t)$ rates, we present in Figs. 8–9 the same running forecasts we have seen in Figs. 4–5. While for many countries/regions the results are indeed similar, by and large, assuming a constant death rate results in more accurate estimates and tighter confidence intervals, without significantly increase the number of forecasts outside the confidence intervals. In fact, the large fluctuations in the forecasts from one week to the next that we had noticed in Section 4.3, are mostly absent in Figs. 8–9. This indicates that a stochastic SIR model with just two time-varying parameters $\beta(t)$ and $\phi(t)$ is still sufficiently rich to represent the data available and that a time-varying death rate introduces unnecessary model uncertainty.

5. Conclusions and future work

We have shown that it is possible to construct reliable forecasts for the evolution of an epidemic purely from time series of new cases and deaths. This is particularly important in scenarios where social behavior and nonpharmaceutical interventions cause a continuous change in epidemic parameters such as the infection and reporting rates.

Because of inherent model ambiguities, the a posteriori forecasts are not always accurate; see for example the estimates in Fig. 8 for the number of deaths in Germany in June or for the United Kingdom in August. However, those inaccurate estimates are typically associated with large 95% confidence intervals that still contain the actual future measurements. From the perspective of epidemic management, the magnitude of the forecast confidence intervals, and in particular the "pessimistic" upper bounds of these intervals, is probably more important than the estimates themselves, as it should inform decision makers of the expected worst-case stress on the healthcare system.

We have used a Gaussian random walk stochastic model for parameter drift that is completely agnostic to external factors. It should be possible to improve forecasting when we have available a set of "covariates" that can be used to estimate parameter variations, as in IHME COVID-19 (2020). Introducing such covariates is the subject of future research.

We have seen that there are fundamental limitations in identifying the internal state and parameters of an SIR model based on daily counts of new cases and deaths. However, this lack of identifiability can be lifted through the use of additional measurements, such as how many new cases are asymptomatic and/or how many individuals were tested and received a negative result. Incorporating such measurements is also the subject of future research, as well as a more systematic study of the identifiability of SIR models (Miao et al., 2011).

Our SIR model is focused on a specific region and the transfer of infected individuals from other regions was only accounted for through stochastic additive terms. This could be improved by considering more sophisticated models that explicitly take into account external effects. However, we foreseen significant challenges in identifying the associated parameters.

Declaration of competing interest

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

Acknowledgments

This research was supported by the National Science Foundation, USA under Grant No. ECCS-2029985. The authors thank Prof. Francesco Bullo for providing important references and information related to this work and to Dr. Radhakisan Baheti for encouragement and guidance.

References

Al-Salti, N., Al-Musalhi, F., Elmojtaba, I., & Gandhi, V. (2020). SIR model with time-varying contact rate. *International Journal of Biomathematics*, http://dx.doi.org/10.1142/S1793524521500170.

Andersson, H., & Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Springer.

Ball, F., & Neal, P. (2002). A general model for stochastic SIR epidemics with two levels of mixing. *Mathematical Biosciences*, *180*(1–2), 73–102.

Beretta, E., Capasso, V., & Rinaldi, F. (1988). Global stability results for a generalized Lotka-Volterra system with distributed delays. *Journal of Mathematical Biology*, *26*(6), 661–688.

Beretta, E., Kolmanovskii, V., & Shaikhet, L. (1998). Stability of epidemic model with time delays influenced by stochastic perturbations. *Mathematics and Computers in Simulation*, *45*(3–4), 269–277.

Beretta, E., & Takeuchi, Y. (1995). Global stability of an SIR epidemic model with time delays. *Journal of Mathematical Biology*, *33*(3), 250–260.

Bertsekas, D. P. (1999). *Nonlinear Programming* (second ed.). Belmont, MA: Athena Scientific.

Brauer, F., Castillo-Chavez, C., & Feng, Z. (2019). *Mathematical Models in Epidemiology*. Springer.

Calafiore, G. C., Novara, C., & Possieri, C. (2020). A time-varying SIRD model for the COVID-19 contagion in Italy. *Annual Reviews in Control*.

Capasso, V. (2008). *Mathematical Structures of Epidemic Systems*. Springer.

Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (2020). COVID-19 data repository. https://github.com/CSSEGISandData/COVID-19.

Comunian, A., Gaburro, R., & Giudici, M. (2020). Inversion of a SIR-based model: a critical analysis about the application to COVID-19 epidemic. *Physica D: Nonlinear Phenomena*, *413*, Article 132674.

CoronaCidades.org. (0000). Farol Covid, https://data.brasil.io/dataset/covid19.

COVID-19 (2020). COVID-19 forecasts based on a stochastic SIR model. URL http://www.ece.ucsb.edu/~hespanha/covid19.

Davis, T. A., Gilbert, J. R., Larimore, S. I., & Ng, E. G. (2004). A column approximate minimum degree ordering algorithm. *Association for Computing Machinery. Transactions on Mathematical Software*, *30*(3), 353–376.

Della Rossa, F., Salzano, D., Di Meglio, A., De Lellis, F., Coraggio, M., Calabrese, C., et al. (2020). A network model of Italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic. *Nature Commun.*, *11*(1), 1–9.

Efimov, D., & Ushirobira, R. (2020). On interval prediction of COVID-19 development based on a SEIR epidemic model. In *Proc. of the 59th Conf. on Decision and Contr.*.

European Center for Disease Prevention and Control. (0000) Download COVID-19 datasets, https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., et al. (2020). Modeling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*.

Hamer, W. H. (1906). *Epidemic disease in England: the evidence of variability and of persistency of type*. Bedford Press.

Hespanha, J. P. (2017). *TensCalc — A toolbox to generate fast code to solve nonlinear constrained minimizations and compute Nash equilibria*: *Technical report*, Santa Barbara: University of California, Santa Barbara, Available at http://www.ece.ucsb.edu/~hespanha/techrep.html.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, *42*(4), 599–653.

IHME COVID-19 Forecasting Team, R. C. Reiner, et al. (2020). Modeling COVID-19 scenarios for the United States. *Nature Medicine*.

Ji, C., & Jiang, D. (2014). Threshold behaviour of a stochastic SIR model. *Applied Mathematical Modelling*, *38*(21–22), 5067–5079.

Keeling, M. J., & Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

King, A. A., Domenech de Cellès, M., Magpantay, F. M. G., & Rohani, P. (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. In *Proc. of the Royal Society B: Biological Sciences Vol. 282* 1806.

Köhler, J., Schwenkel, L., Koch, A., Berberich, J., Pauli, P., & Allgöwer, F. (2020). Robust and optimal predictive control of the COVID-19 outbreak. arXiv preprint arXiv:2005.03580.

Li, M. L., Bouardi, H. T., Lami, O. S., Trikalinos, T. A., Trichakis, N. K., & Bertsimas, D. (2020). Forecasting Covid-19 and analyzing the effect of government interventions. medRxiv.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

medicalxpress. com (2020). Superspreader events key driver in COVID-19 pandemic. https://medicalxpress.com/news/2020-11-superspreader-events-key-driver-covid-.html.

Mei, W., Mohagheghi, S., Zampieri, S., & Bullo, F. (2017). On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, *44*, 116–128.

Miao, H., Xia, X., Perelson, A. S., & Wu, H. (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Review Society for Industrial and Applied Mathematics*, *53*(1), 3–39.

Morgan, O. (2019). How decision makers can use quantitative approaches to guide outbreak responses. *Philosophical Transactions of the Royal Society B*, *374*(1776), Article 20180365.

Peng, L., Yang, W., Zhang, D., Zhuge, C., & Hong, L. (2020). Epidemic analysis of COVID-19 in China by dynamical modeling. arXiv preprint arXiv:2002.06563.

y Piontti, A. P., Perra, N., Rossi, L., Samay, N., & Vespignani, A. (2019). *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age*. Springer.

Portuguese Direção-Geral de Saúde (2020). COVID-19 Relatório de Situação. https://covid19.min-saude.pt/relatorio-de-situacao/.

Roda, W. C., Varughese, M. B., Han, D., & Li, M. Y. (2020). Why is it difficult to accurately predict the COVID-19 epidemic?. *Infectious Disease Modelling*.

Shearer, F. M., Moss, R., McVernon, J., Ross, J. V., & McCaw, J. M. (2020). Infectious disease pandemic planning and response: Incorporating decision analysis. *PLoS Medicine*, *17*(1), Article e1003018.

Srivastava, A., Xu, T., & Prasanna, V. K. (2020). Fast and accurate forecasting of COVID-19 deaths using the sikjα model. arXiv preprint arXiv:2007.05180.

Stolerman, L. M., Coombs, D., & Boatto, S. (2015). SIR-network model and its application to dengue fever. *SIAM Journal of Applied Mathematics*, *75*(6), 2581–2609.

Tornatore, E., Buccellato, S. M., & Vetro, P. (2005). Stability of a stochastic SIR system. *Physica A: Statistical Mechanics and its Applications*, *354*, 111–126.

Xia, Z.-Q., Zhang, J., Xue, Y.-K., Sun, G.-Q., & Jin, Z. (2015). Modeling the transmission of middle east respirator syndrome corona virus in the Republic of Korea. *PLoS One*, *10*(12).

Youssef, M., & Scoglio, C. (2011). An individual-based approach to SIR epidemics in contact networks. *Journal of Theoretical Biology*, *283*(1), 136–144.

Zou, D., Wang, L., Xu, P., Chen, J., Zhang, W., & Gu, Q. (2020). Epidemic model guided machine learning for COVID-19 forecasts in the United States. medRxiv.