



Research article

A stochastic metapopulation state-space approach to modeling and estimating COVID-19 spread

Yukun Tan¹, Durward Cator III¹, Martial Ndeffo-Mbah^{2,3} and Ulisses Braga-Neto^{1,*}

¹ Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843, USA

² Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, 77843, USA

³ Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, College Station, TX, 77843, USA

* **Correspondence:** Email: ulisses@tamu.edu.

Abstract: Mathematical models are widely recognized as an important tool for analyzing and understanding the dynamics of infectious disease outbreaks, predict their future trends, and evaluate public health intervention measures for disease control and elimination. We propose a novel stochastic metapopulation state-space model for COVID-19 transmission, which is based on a discrete-time spatio-temporal susceptible, exposed, infected, recovered, and deceased (SEIRD) model. The proposed framework allows the hidden SEIRD states and unknown transmission parameters to be estimated from noisy, incomplete time series of reported epidemiological data, by application of unscented Kalman filtering (UKF), maximum-likelihood adaptive filtering, and metaheuristic optimization. Experiments using both synthetic data and real data from the Fall 2020 COVID-19 wave in the state of Texas demonstrate the effectiveness of the proposed model.

Keywords: epidemic model; SEIRD model; nonlinear stochastic model; unscented Kalman filter; maximum likelihood; adaptive filtering; parameter estimation

1. Introduction

Infectious disease outbreaks remain a major threat to global health. This is especially the case for highly pathogenic and transmissible diseases with pandemic potential. These global threats were recently exemplified by the 2009 swine flu outbreak and the ongoing COVID-19 pandemic caused by the novel Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2). To effectively mitigate and control the spread of an epidemic, it is paramount for public health decision-making to be informed

by an accurate understanding of the dynamics of disease transmission and the potential impact of intervention measures.

To this end, epidemic models have become an important tool to help better understand epidemic dynamics, predict its future trends, evaluate the effectiveness of intervention measures, such as lock-down or vaccination, and ultimately control the epidemic.

Mathematical epidemiology models can be broadly divided into two main types: compartmental models and agent-based models. An agent-based model is a very detailed stochastic model where the agent represent single individuals within a population [1–8]. Such models are generally complex and computationally expensive. In compartmental models, the population is subdivided into epidemiological compartments with each compartmental generally representing a health or disease progression stage [9–16]. Thus compartmental models are computationally simple, scalable, and capable of describing the dynamics of the number of people in each compartment throughout the course of an epidemic. In this work, we propose a nonlinear state-space compartmental model.

Compartmental model date back to the early twentieth century, most notably to the work by [17], whose susceptible, infected, and recovered (SIR) model was used for modeling the plague (London 1665–1666, Bombay 1906) and cholera (London 1865) epidemics [10]. More specifically, it is a weighted directed graph representation of a dynamic system. The susceptible refers to those healthy people who are susceptible to the disease and may get infected; the infected refers to those under infection; the recovered refers to those who recovered from infection and will be temporarily or permanently immune to the disease. However, there are several drawbacks in the original SIR model:

- 1) It is a deterministic model, meaning the model always performs the same for a given initial condition, which cannot explain the unknowable randomness in the observations [12, 16, 18];
- 2) It is only a temporal model, which does not consider the spread in geographical regions, e.g., human interaction induced by modern transportation [19];
- 3) It assumes that all the parameters are known, which is not realistic. Parameter estimation from noisy observations is needed for better understanding and forecasting epidemics [18].

Many researchers have proposed variants of the basic SIR model to address some of these issues, such as [20], who considered the age dependence of social structure; [21], who used hyperbolic compartmental models to solve the large uncertainty in the values reported by official sources; [22, 23], who applied multiscale kinetic transport equations to describe the transport dynamic of the commuters and non-commuting population; but accurate state and parameter estimation from partial and noisy observations remains an open problem. To handle this issue, our proposed framework embeds the classical compartmental model within a nonlinear state-space model. The state model is a spatial-temporal stochastic dynamic model that allows hidden states in a given location to change over time and the disease dynamics in one location to affect neighbouring locations through human movements between locations. The proposed framework consists of a multinomial state model based on a variant of the SIR model — the SEIRD model [24–26] — and an observation model to allow the assimilation of publicly available data, including daily testing rate, daily test positivity rate, specificity and sensitivity of the tests. In addition, the model considers the differential testing rate between symptomatic patients and asymptomatic and healthy individuals. The proposed framework is employed to estimate i) the hidden epidemic state vector $X^t = (S^t, E^t, I^t, R^t, D^t)$ and ii) epidemiological parameters, such as infection rate and the average infectious period, using noisy incomplete time series epidemic

data. Accurate estimation of the true epidemic state and parameters is paramount for designing and evaluating the effectiveness of control strategies.

2. Mathematical model

2.1. Classical SEIRD model

The SEIRD model predicts the time evolution of epidemic. It models the dynamic interaction of people between five different compartments, namely, the susceptible (S), the exposed (E), the infected (I), the recovered (R) and the deceased (D). The classic discrete-time SEIRD model can be described by the following equations:

$$\begin{aligned}
 S_{t+1} &= S_t - \lambda_S S_t I_t \\
 E_{t+1} &= E_t + \lambda_S S_t I_t - \lambda_E E_t \\
 I_{t+1} &= I_t + \lambda_E E_t - (\lambda_R + \lambda_D) I_t \\
 R_{t+1} &= R_t + \lambda_R I_t \\
 D_{t+1} &= D_t + \lambda_D I_t
 \end{aligned} \tag{2.1}$$

with $S_t + E_t + I_t + R_t + D_t = 1$, where S_t, E_t, I_t, R_t, D_t are the fraction of the population (the size of which is assumed to be constant over the time interval of interest) at time t that, respectively, S is not yet infected with the disease; E has been exposed to the virus but does not show symptoms yet; I is infective after the virus incubation period; R has been infected and then recovered; and D is deceased due to the epidemic.

The model is governed by the following parameters:

- λ_S is the infection rate, which is the probability that an individual moves from the S to the E compartment. The nonlinear term $\lambda_S S_t I_t$ models the infection speed, which depends not only on the infection rate, but also on the fraction of susceptible and infected people at time t ;
- λ_E is the probability that an individual moves from the E to the I compartment. It can be understood as the inverse of the average incubation time;
- λ_R is the recovery rate, which is the probability than an individual moves from the I to the R compartment. It can be understood as the inverse of the average recovery time;
- λ_D is the mortality rate, which is the probability than an individual moves from the I to the D compartment.

2.2. Covid-19 epidemic metapopulation state-space model

Given the complexity and reality of epidemics, many different implementations of the classical SEIRD model have been proposed [24–27]. Two clear drawbacks of the classical SEIRD model is the inability to model variation over a geographic area, and the fact that that the SEIRD values are assumed to be directly observable. In this paper, we address these drawbacks by means of a novel nonlinear metapopulation state-space framework, where the state process provides a spatial-temporal discrete-time stochastic model of the evolution of the number of individuals in the SEIRD compartments over several geographical regions, while the observation process models noisy time series of reported

epidemiological data, considering the accuracy of tests and different testing rates for symptomatic and asymptomatic people.

2.2.1. State model

Consider a discrete-time state process $\{\mathbf{X}_i^t; i = 1, \dots, G; t = 0, 1, \dots\}$, where $\mathbf{X}_i^t = (S_i^t, E_i^t, I_i^t, R_i^t, D_i^t)$ is a state vector, containing the number of individuals in the SEIRD compartments in geographical area i at time t , where the unit of time is typically day or week. At time t , the total number of individuals in region i is $N_i = S_i^t + E_i^t + I_i^t + R_i^t + D_i^t$. We assume that this number stays constant over the time interval of interest (e.g., no significant amounts of migration is assumed to occur during the time interval of interest; plus, births and non-specific deaths are assumed to approximately balance out). This results in a good approximation provided that the interval of time in consideration consists of weeks or a few months, especially in the case of a full-blown outbreak when, the impact of migration is negligible.

Before proceeding, we recall the definition of the binomial and multinomial probability distributions. A random variable Z has a binomial distribution with parameters $n \geq 1$ and $0 < p < 1$, denoted by $Z \sim \text{Binomial}(n, p)$, if

$$P(Z = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k = 0, \dots, n. \quad (2.2)$$

Variable Z is the number of k occurrences of an outcome among n identical independent trials that can result in the outcome with probability p . The expected value and variance of this variable are $E[Z] = np$ and $\text{Var}(Z) = np(1-p)$, respectively. More generally, we define a random vector (Z_1, \dots, Z_M) to have a multinomial distribution with parameters $n \geq 1$ and $p_1, \dots, p_M \geq 0$ such that $\sum p_i \leq 1$, denoted by $(Z_1, \dots, Z_M) \sim \text{Multinomial}(n, p_1, \dots, p_M)$, if

$$P(Z_1 = k_1, \dots, Z_M = k_M) = \frac{n!}{k_1! \dots k_M! (n - \sum k_i)!} p_1^{k_1} \dots p_M^{k_M} (1 - \sum p_i)^{n - \sum k_i}, \quad k_1, \dots, k_M \geq 0, \sum k_i \leq n. \quad (2.3)$$

Variables Z_1, \dots, Z_M are the number of occurrences of each of M outcomes over n identical independent trials that can result in outcome m with probability p_m . It is clear that $Z_m \sim \text{Binomial}(n, p_m)$, for $m = 1, \dots, M$ (thus $E[Z_m] = np_m$ and $\text{Var}(Z_m) = np_m(1-p_m)$). In addition, the case $M = 1$ results in the binomial distribution.

In our model, the state vector $\mathbf{X}_i^t = (S_i^t, E_i^t, I_i^t, R_i^t, D_i^t)$ is assumed to evolve according to the following nonlinear stochastic model:

$$\begin{aligned} S_i^{t+1} &= S_i^t - \sum_j N_{S_{i,j}}^t \\ E_i^{t+1} &= E_i^t + \sum_j N_{S_{i,j}}^t - N_{E_i}^t \\ I_i^{t+1} &= I_i^t + N_{E_i}^t - N_{R_i}^t - N_{D_i}^t \\ R_i^{t+1} &= R_i^t + N_{R_i}^t \\ D_i^{t+1} &= D_i^t + N_{D_i}^t \end{aligned} \quad (2.4)$$

for $t = 0, 1, \dots$, where $N_{S_{i,j}}^t$ is the number of susceptible individuals at time t in region i who will become exposed at time $t + 1$ due to contact with an infective individual from region j , $N_{E_i}^t$ is the number of exposed people at time t in region i who will become infective at time $t + 1$, while $N_{R_i}^t$ and

$N_{D_i}^t$ are the numbers of infected people at time t in region i who will become recovered or deceased at time $t + 1$, respectively.

We assume that each susceptible individual in region i becomes exposed due to contact with an infective individual from region j at time t independently with a probability $\lambda_S c_{ij} I_j^t / N_j$ (these parameters are explained below). Therefore, the distribution of the numbers of new infections over the regions is multinomial:

$$(N_{S_{i,1}}^t, \dots, N_{S_{i,G}}^t) \sim \text{Multinomial} \left(S_i^t, \lambda_S \frac{c_{i1} I_1^t}{N_1}, \dots, \lambda_S \frac{c_{iG} I_G^t}{N_G} \right). \quad (2.5)$$

The infection rate $\lambda_S > 0$ is disease-specific, but also affected by public policies, such as mask wearing and social distancing. We assume that these factors are constant over the geographical area (e.g., a single political unit, such as a state or country) and time interval in the study, so that λ_S is the same for all regions. If $i = j$, then the contact is internal to region i , and $c_{ii} = 1$. If $i \neq j$, then $0 \leq c_{ij} \leq 1$ models the relative amount of transient interchange of individuals between regions i and j , due to commuting, tourism, and so on. If regions i and j are far apart, or one of them is not economically important, then c_{ij} is close to zero. Note that $c_{ij} = c_{ji}$. The values of c_{ij} are selected in our study based on the *gravity model*, which is a model of population movement between discrete locations, based on the populations of the origin and destination and the distance between them [28–30]. Namely, the c_{ij} values are determined by a proportional factor (the so-called “gravitational constant”) multiplied by the product of the populations of both regions and divided by the square of their distance from one another. Finally, the ratio I_j^t / N_j indicates that if there are more infective individuals in region j , they will spread the disease with higher probability. Note that the expected value of the (normalized) j -th flux is $E[N_{S_{i,j}}^t / N_i] = \lambda_S c_{ij} (S_i^t / N_i) (I_j^t / N_j)$. This may be contrasted to the flux $\lambda_S S(t) I(t)$ in the classical case.

In our model, an exposed individual in region i becomes infective at time t , independently of the other regions, with probability λ_E . Hence,

$$N_{E_i}^t \sim \text{Binomial}(E_i^t, \lambda_E). \quad (2.6)$$

The expected number of exposed individuals who become infective at time t is thus $E[N_{E_i}^t] = \lambda_E E_i^t$. Note also that the distribution of the total time until an exposed individual becomes infective is *geometric* with parameter λ_E , so that the expected number of time units until an exposed individual becomes infective is λ_E^{-1} .

Finally, each infective individual in region i becomes recovered or deceased at time t , independently of the other regions, with probabilities λ_R and λ_D , respectively. Hence,

$$(N_{R_i}^t, N_{D_i}^t) \sim \text{Multinomial}(I_i^t, \lambda_R, \lambda_D). \quad (2.7)$$

The expected numbers of infective individuals who become recovered or deceased at time t are $E[N_{R_i}^t] = \lambda_R I_i^t$ and $E[N_{D_i}^t] = \lambda_D I_i^t$, respectively. The expected numbers of time units until an infective individual becomes recovered or deceased are λ_R^{-1} and λ_D^{-1} , respectively.

2.2.2. Observation model

The observation model is a major contribution of this work. We model reported epidemiological data, namely confirmed new cases and cumulative recorded deaths, as noisy observations on the true

state of the pandemic, as defined in the previous section. The proposed observation model addresses the uncertainty introduced by imperfect testing and the imbalance in testing of symptomatic and asymptomatic individuals (the former cohort is tested more [31]).

We model the reported data as a time series $\{\mathbf{Y}_i^t; i = 1, \dots, G; t = 0, 1, \dots\}$ where $\mathbf{Y}_i^t = (P_i^t, Q_i^t)$ contains the numbers P_i^t and Q_i^t of new confirmed cases and deaths, respectively, in geographical area i at time t . The number of reported new cases contain both false and true positives,

$$P_i^t = N_{TP_i}^t + N_{FP_i}^t, \quad (2.8)$$

for $t = 0, 1, \dots$, where $N_{TP_i}^t$ and $N_{FP_i}^t$ are the number of false and true testing positives, respectively. Let α and β be the false positive and true positive rates, respectively, of the COVID-19 test under consideration (multiple tests of different accuracy can be introduced by splitting the populations according to the test received). True positives come from the exposed and infective populations. In addition, a percentage of the infective population is asymptomatic; we assume that those are tested at a smaller rate than the symptomatic ones. Accordingly, we split the number of new true positives at time t as the sum of the number of positive-tested symptomatic infective people and the number of positive-tested asymptomatic infective and exposed individuals:

$$N_{TP_i}^t \sim \text{Binomial}(\varepsilon_2^t(1 - \varepsilon_4^t)I_i^t, \beta) + \text{Binomial}(\varepsilon_1^t(\varepsilon_4^t I_i^t + E_i^t), \beta), \quad (2.9)$$

where ε_1^t and ε_2^t are the testing rates of asymptomatic and symptomatic individuals at time t , respectively, while ε_4^t is the percentage of infective individuals who are asymptomatic at time t . These parameters are assumed to be the same over all regions. As shown later, these parameters can be calculated from the overall testing and positivity rates, which are available from public records.

Similarly, the false positives come from the susceptible and recovered populations, and here we have to distinguish between symptomatic individuals (who are not infected by COVID-19 but instead by other similar viruses, such as Influenza) and asymptomatic individuals:

$$N_{FP_i}^t \sim \text{Binomial}(\varepsilon_1^t(1 - \varepsilon_3^t)(S_i^t + R_i^t), \alpha) + \text{Binomial}(\varepsilon_2^t \varepsilon_3^t (S_i^t + R_i^t), \alpha), \quad (2.10)$$

where ε_3^t is the percentage of non-specific symptomatic individuals at time t .

Finally, we assume that the reported cumulative recorded deaths in each region is a fraction of the true number D_i^t , according to

$$Q_i^t \sim \text{Binomial}(D_i^t, \beta). \quad (2.11)$$

Even though there might be a number of non-COVID-19 deaths falsely reported as COVID-19, we are assuming that this number is small and can be ignored. Notice that (2.11) implies that COVID-19 deaths are underreported, which is widely assumed to be the case.

Let R_T^t be the testing rate, i.e., the percentage of the population that is tested at time t , and let R_p^t the positivity rate at time t . Both these parameters are reported by health authorities. Next, we show how the parameters ε_1^t , ε_2^t , ε_3^t , and ε_4^t can be calculated from R_T^t and R_p^t .

The test positivity rate (R_p^t) is the percentage of positive tests P^t over the total number of tests N_T^t in a given region; it can be written as

$$R_p^t = \frac{P^t}{N_T^t} = \frac{\varepsilon_1^t A \alpha + \varepsilon_2^t B \alpha + \varepsilon_2^t C \beta + \varepsilon_1^t D \beta}{\varepsilon_1^t A + \varepsilon_2^t B + \varepsilon_2^t C + \varepsilon_1^t D} \quad (2.12)$$

where $A = (1 - \varepsilon_3^t)(S^t + R^t)$, $B = \varepsilon_3^t(S^t + R^t)$, $C = (1 - \varepsilon_4^t)I^t$, and $D = \varepsilon_4^t I^t + E^t$.

Similarly, the testing rate R_T^t is the percentage of total tests over the total population, which can be written as

$$R_T^t = \frac{N_T^t}{N^t} = \frac{\varepsilon_1^t A + \varepsilon_2^t B + \varepsilon_2^t C + \varepsilon_1^t D}{A + B + C + D} \quad (2.13)$$

Using the previous two equations, we can solve for ε_1^t and ε_2^t in terms of ε_3^t and ε_4^t :

$$\begin{aligned} \varepsilon_1^t &= \frac{(S^t + E^t + I^t + R^t)R_T^t(R_p^t(\varepsilon_3^t(S^t + R^t) + (1 - \varepsilon_4^t)I^t) - \alpha\varepsilon_3^t(S^t + R^t) - \beta(1 - \varepsilon_4^t)I^t)}{(\beta - \alpha)(\varepsilon_3^t(S^t + R^t)(\varepsilon_4^t I^t + E^t) - (1 - \varepsilon_3^t)(S^t + R^t)(1 - \varepsilon_4^t)I^t)} \\ \varepsilon_2^t &= \frac{(S^t + E^t + I^t + R^t)R_T^t}{\beta - \alpha} \times \\ &\quad \frac{\alpha(1 - \varepsilon_3^t)(S^t + R^t) + \beta(\varepsilon_4^t I^t + E^t) - R_p^t((1 - \varepsilon_3^t)(S^t + R^t) + (\varepsilon_4^t I^t + E^t))}{\varepsilon_3^t(S^t + R^t)(\varepsilon_4^t I^t + E^t) - (1 - \varepsilon_3^t)(S^t + R^t)(1 - \varepsilon_4^t)I^t} \end{aligned} \quad (2.14)$$

Now, ε_1^t and ε_2^t are the testing rates of asymptomatic and symptomatic individuals at time t , and can be related to each other. According to [31], we have

$$\varepsilon_2^t = \zeta \varepsilon_1^t, \quad (2.15)$$

where ζ is uniformly distributed in the interval [2, 4.3] to account for various non-specific symptoms (e.g., fever, cough, loss of taste, smell) that are shared by COVID-19 and other common viral diseases. Substituting (2.15) into (2.14) results, after some algebra, in the following expression for ε_3^t

$$\varepsilon_3^t = \frac{(\beta - R_p^t)(\varepsilon_4^t I^t + E^t) + \zeta(1 - \varepsilon_4^t)I^t(\beta - R_p^t) - (S^t + R^t)(R_p^t - \alpha)}{(\zeta - 1)(S^t + R^t)(R_p^t - \alpha)} \quad (2.16)$$

In addition, according to [32], the percentage of asymptomatic infective people ε_4^t is around 20% and from this the other parameters ε_1^t , ε_2^t , ε_3^t can be calculated.

3. Covid-19 epidemic state and parameter estimation

In this section, we describe state-of-the-art estimators for the state and parameters of the metapopulation state-space model introduced in the previous section. These estimators use as input the noisy time series $\mathbf{Y}_i^t = (P_i^t, Q_i^t)$ of reported new cases and deaths in the several geographical regions in the study. The state vector $\mathbf{X}_i^t = (S_i^t, E_i^t, I_i^t, R_i^t, D_i^t)$ is estimated at each time t by using the Unscented Kalman Filter (UKF) [33], which is the state-of-the-art estimator for stochastic nonlinear state-space models, while the parameters $\lambda_S, \lambda_R, \lambda_D$ are estimated by maximum-likelihood, by using the metaheuristic optimization - fish school search algorithm [34–36].

First, we write the state-space model in the previous section in the standard form

$$\begin{aligned} \mathbf{X}_i^{t+1} &= f(\mathbf{X}_i^t) + \mathbf{n}_i^t \\ \mathbf{Y}_i^t &= h(\mathbf{X}_i^t) + \mathbf{v}_i^t \end{aligned} \quad (3.1)$$

where f and h are nonlinear mappings, and $\{\mathbf{n}_i^t; i = 1, \dots, G; t = 0, 1, \dots\}$ and $\{\mathbf{v}_i^t; i = 1, \dots, G; t = 0, 1, \dots\}$ are white-noise (i.e., uncorrelated in time) transition and observation noise processes, respectively, which are independent of each other and independent of the initial state $\{\mathbf{X}_i^0; i = 1, \dots, G\}$.

The proposed state-space model can be put in the standard form (3.1) by rewriting the state equations as:

$$\begin{aligned}
 S_i^{t+1} &= \left(1 - \lambda_S \sum_j c_{ij} I_j^t / N_j\right) \times S_i^t + n_{S_i}^t \\
 E_i^{t+1} &= (1 - \lambda_E) \times E_i^t + \left(\lambda_S \sum_j c_{ij} I_j^t / N_j\right) \times S_i^t + n_{E_i}^t \\
 I_i^{t+1} &= (1 - \lambda_R - \lambda_D) \times I_i^t + \lambda_E \times E_i^t + n_{I_i}^t \\
 R_i^{t+1} &= R_i^t + \lambda_R \times I_i^t + n_{R_i}^t \\
 D_i^{t+1} &= D_i^t + \lambda_D \times I_i^t + n_{D_i}^t
 \end{aligned} \tag{3.2}$$

where the transition noise terms are:

$$\begin{aligned}
 n_{S_i}^t &= -\sum_j \left(N_{S_{i,j}}^t - E[N_{S_{i,j}}^t]\right) \\
 &= -\sum_j \left(N_{S_{i,j}}^t - \lambda_S c_{ij} S_i^t I_j^t / N_j\right) \\
 n_{E_i}^t &= -\left(N_{E_i}^t - E[N_{E_i}^t]\right) + \sum_j \left(N_{S_{i,j}}^t - E[N_{S_{i,j}}^t]\right) \\
 &= -\left(N_{E_i}^t - \lambda_E E_i^t\right) + \sum_j \left(N_{S_{i,j}}^t - \lambda_S c_{ij} S_i^t I_j^t / N_j\right) \\
 n_{I_i}^t &= \left(N_{E_i}^t - E[N_{E_i}^t]\right) - \left(N_{R_i}^t - E[N_{R_i}^t]\right) - \left(N_{D_i}^t - E[N_{D_i}^t]\right) \\
 &= \left(N_{E_i}^t - \lambda_E E_i^t\right) - \left(N_{R_i}^t - \lambda_R I_i^t\right) - \left(N_{D_i}^t - \lambda_D I_i^t\right) \\
 n_{R_i}^t &= \left(N_{R_i}^t - E[N_{R_i}^t]\right) = \left(N_{R_i}^t - \lambda_R I_i^t\right) \\
 n_{D_i}^t &= \left(N_{D_i}^t - E[N_{D_i}^t]\right) = \left(N_{D_i}^t - \lambda_D I_i^t\right)
 \end{aligned} \tag{3.3}$$

Similarly, the observation model can be rewritten in the standard form (3.1):

$$\begin{aligned}
 P_i^t &= \beta \varepsilon_2^t \left(1 - \varepsilon_4^t\right) I_i^t + \beta \varepsilon_1^t \left(\varepsilon_4^t I_i^t + E_i^t\right) \\
 &\quad + \alpha \varepsilon_1^t \left(1 - \varepsilon_3^t\right) \left(S_i^t + R_i^t\right) + \alpha \varepsilon_2^t \varepsilon_3^t \left(S_i^t + R_i^t\right) + v_{P_i}^t \\
 Q_i^t &= \beta D_i^t + v_{Q_i}^t
 \end{aligned} \tag{3.4}$$

where the observation noise terms are:

$$\begin{aligned}
 v_{P_i}^t &= N_{TP_i}^t - E[N_{TP_i}^t] + N_{FP_i}^t - E[N_{FP_i}^t] \\
 &= N_{TP_i}^t - \beta \left(\varepsilon_2^t (1 - \varepsilon_4^t) I_i^t + \varepsilon_1^t (\varepsilon_4^t I_i^t + E_i^t)\right) \\
 &\quad + N_{FP_i}^t - \alpha \left(\varepsilon_1^t (1 - \varepsilon_3^t) (S_i^t + R_i^t) + \varepsilon_2^t \varepsilon_3^t (S_i^t + R_i^t)\right) \\
 v_{Q_i}^t &= D_{TP_i}^t - E[D_{TP_i}^t] = D_{TP_i}^t - \beta D_i^t
 \end{aligned} \tag{3.5}$$

3.1. Unscented Kalman filter

With the state-space model in the standard format (3.2), one can apply the *Unscented Kalman Filter* (UKF) algorithm [33, 37, 38] to estimate the state variables $(S_i^t, E_i^t, I_i^t, R_i^t, D_i^t)$ from the noisy data (P_i^t, Q_i^t) of new cases and cumulative deaths in geographical area i at time t . The UKF assumes that the statistics of the state variables at time $t = 0$ are known (in practice, these values need to be only very roughly guessed since, as the time t increases, the UKF generally “forgets” the information in the initial state). It also assumes that all the parameters are known; however, in the next section we describe a methodology to estimate the parameters from the data as well.

In the UKF algorithm, $\mathbf{m}_{t|t}$ and $\mathbf{P}_{t|t}$ denote the estimates at time t of the mean and error covariance matrix, respectively, of the state vector $\mathbf{X}^t = (S^t, E^t, I^t, R, D^t)$ using all observed data up to time t (for brevity, the subscript i used previously to discriminate the region is omitted throughout this section, since the same process is applied to all regions separately). We initialize the mean vector via $E[I^0] = E[E^0] = P^0$, $E[R^0] = 0$, $E[D^0] = Q^0/\beta$, and $E[S^0] = N - E[I^0 + E^0 + R^0 + D^0]$, while the covariance matrix is initialized to the identity matrix. The UKF estimate of the state \mathbf{X}^t is $\mathbf{m}_{t|t}$, with uncertainty given by $\mathbf{P}_{t|t}$. These estimates are computed by the following iteration:

Initialization: Given the initial observation $\mathbf{Y}^0 = (P^0, Q^0)$, let

$$\mathbf{m}_{0|0} = (N - 2P^0 - D^0/\beta, P^0, P^0, 0, Q^0/\beta), \quad \mathbf{P}_{0|0} = \mathbf{I}. \quad (3.6)$$

For $t = 1, 2, \dots$ **repeat:**

Prediction:

1) Generate $(2n+1)$ sigma points,

$$\begin{aligned} \mathbf{z}_{t-1|t-1}^0 &= \mathbf{m}_{t-1|t-1}, \\ \mathbf{z}_{t-1|t-1}^i &= \mathbf{m}_{t-1|t-1} + \left[\sqrt{n\mathbf{P}_{t-1|t-1}} \right]_i, \quad i = 1, \dots, n, \\ \mathbf{z}_{t-1|t-1}^{i+n} &= \mathbf{m}_{t-1|t-1} - \left[\sqrt{n\mathbf{P}_{t-1|t-1}} \right]_{i-n}, \quad i = n+1, \dots, 2n, \end{aligned} \quad (3.7)$$

where $\left[\sqrt{n\mathbf{P}_{t-1|t-1}} \right]_i$ is the i -th column of the matrix square root.

2) Propagate the sigma points through the state equation:

$$\mathbf{x}_t^i = \mathbf{f}(\mathbf{z}_{t-1|t-1}^i), \quad i = 0, \dots, 2n \quad (3.8)$$

3) Compute predicted mean and predicted error covariance:

$$\begin{aligned} \mathbf{m}_{t|t-1} &= \frac{1}{2n} \sum_{i=0}^{2n} \hat{\mathbf{x}}_t^i, \\ \mathbf{P}_{t|t-1} &= \frac{1}{2n} \sum_{i=0}^{2n} (\hat{\mathbf{x}}_t^i - \mathbf{m}_{t|t-1}) \times (\hat{\mathbf{x}}_t^i - \mathbf{m}_{t|t-1})^T + \mathbf{Q}_{t-1}, \end{aligned} \quad (3.9)$$

where \mathbf{Q}_{t-1} is the covariance matrix of the transition noise (see the Supplementary Information for its derivation).

Update:

1) Update sigma points based on the predicted mean and error covariance:

$$\begin{aligned} \mathbf{z}_{t|t-1}^0 &= \mathbf{m}_{t|t-1}, \\ \mathbf{z}_{t|t-1}^i &= \mathbf{m}_{t|t-1} + \left[\sqrt{n\mathbf{P}_{t|t-1}} \right]_i, \quad i = 1, \dots, n, \\ \mathbf{z}_{t|t-1}^{i+n} &= \mathbf{m}_{t|t-1} - \left[\sqrt{n\mathbf{P}_{t|t-1}} \right]_{i-n}, \quad i = n+1, \dots, 2n. \end{aligned} \quad (3.10)$$

2) Propagate the sigma point through the observation equation:

$$\hat{\mathbf{y}}_t^i = \mathbf{h}(\mathbf{z}_{t|t-1}^i), \quad i = 0, \dots, 2n. \quad (3.11)$$

3) Compute predicted measurement mean, measurement covariance matrix, and cross-covariance matrix:

$$\begin{aligned} \boldsymbol{\mu}_t &= \frac{1}{2n} \sum_{i=0}^{2n} \hat{\mathbf{y}}_t^i, \\ \mathbf{S}_t &= \frac{1}{2n} \sum_{i=0}^{2n} (\hat{\mathbf{y}}_t^i - \boldsymbol{\mu}_t)(\hat{\mathbf{y}}_t^i - \boldsymbol{\mu}_t)^T + \mathbf{R}_t, \\ \mathbf{C}_t &= \frac{1}{2n} \sum_{i=0}^{2n} (\mathbf{z}_{t|t-1}^i - \mathbf{m}_{t|t-1})(\hat{\mathbf{y}}_t^i - \boldsymbol{\mu}_t)^T, \end{aligned} \quad (3.12)$$

where \mathbf{R}_{t-1} is the covariance matrix of the observation noise (see the Supplementary Information for its derivation).

4) Compute the filter gain and new state error covariance matrix. Assimilate the observation \mathbf{Y}^t at time t to find new state mean vector $\mathbf{m}_{t|t}$:

$$\begin{aligned} \mathbf{K}_t &= \mathbf{C}_t \mathbf{S}_t^{-1}, \\ \mathbf{m}_{t|t} &= \mathbf{m}_{t|t-1} + \mathbf{K}_t (\mathbf{Y}^t - \boldsymbol{\mu}_t), \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{S}_t \mathbf{K}_t^T. \end{aligned} \quad (3.13)$$

3.2. Maximum-likelihood adaptive filtering

The UKF algorithm requires that all parameter values be known. The parameters that govern the proposed model in (3.2) and (2.8) are λ_S , λ_E , λ_R , λ_D , α , β , ε_1 , ε_2 , ε_3 , and ε_4 . Of these, α and β are known, while ε_1 through ε_4 can be obtained using the procedure described in Section 2.2.2. Among the λ parameters, some may be known a priori, but others may be unknown. Let $\boldsymbol{\theta}$ be a vector containing these parameters. Accurate estimation of $\boldsymbol{\theta}$ from the observed data is key to make the proposed methodology useful in practice. We estimate $\boldsymbol{\theta}$ using maximum-likelihood method in combination with the UKF, which is known as maximum-likelihood adaptive filtering [39–41].

Let $\mathbf{Y}^{0:t} = \{\mathbf{Y}^0, \mathbf{Y}^1, \dots, \mathbf{Y}^t\}$ denote the observed data up to time t . The log-likelihood of the model parameters $\boldsymbol{\theta}$ at time t is given by:

$$\begin{aligned} L_t(\boldsymbol{\theta}) &= \log p_{\boldsymbol{\theta}}(\mathbf{Y}^{0:t}) = \log \left[p_{\boldsymbol{\theta}}(\mathbf{Y}_t | \mathbf{Y}^{0:t-1}) p_{\boldsymbol{\theta}}(\mathbf{Y}^{t-1} | \mathbf{Y}^{0:t-2}) \cdots p_{\boldsymbol{\theta}}(\mathbf{Y}^1 | \mathbf{Y}^0) p_{\boldsymbol{\theta}}(\mathbf{Y}^0) \right] \\ &= L_{t-1}(\boldsymbol{\theta}) + \log p_{\boldsymbol{\theta}}(\mathbf{Y}^t | \mathbf{Y}^{0:t-1}), \end{aligned} \quad (3.14)$$

where

$$p_{\theta}(\mathbf{Y}^t | \mathbf{Y}^{0:t-1}) = -\frac{1}{2} \log |2\pi \mathbf{S}_t(\theta)| - \frac{1}{2} \mathbf{V}_t^T(\theta) \mathbf{S}_t^{-1}(\theta) \mathbf{V}_t(\theta), \quad (3.15)$$

with $\mathbf{V}_t = \mathbf{Y}^t - \boldsymbol{\mu}_t$. The quantities $\boldsymbol{\mu}_t$ and \mathbf{S}_t are calculated in the UKF recursion. Then, the target is to maximize the log-likelihood $L_t(\theta)$,

$$\hat{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta}} L_t(\boldsymbol{\theta}). \quad (3.16)$$

There are several possible ways to address this optimization problem. For example, one can apply the Expectation-Maximization (EM) algorithm, which is especially effective when there is a closed-form solution for the ‘‘M’’ maximization step, which avoids recursive gradient calculation. However, there is no such closed-form solution in our case. Other gradient-based optimization methods did not produce good results. Instead, we used biology-inspired metaheuristic optimization, in this case, the *Fish School Search* (FSS) algorithm [34, 35], which we had already applied in our previous work [36]. Metaheuristic optimization algorithms, such as FSS, can find good solutions, in a reasonable amount of time, of complex, non-linear, and multi-dimensional problems, where traditional gradient-descent optimization is not effective [42–44]. A brief description of the FSS algorithm is found in the Supplementary Information. For more details about the FSS algorithm, please see [34–36].

4. Numerical experiments

In this section, we present the results of numerical experiments, using synthetic and real data from the state of Texas in the United States. We divided Texas into 11 regions, based on the Texas Health and Human Services (HHS) regional map.

Sections 4.1 and 4.2 assume the parameter values displayed in Table 1 (unless stated otherwise). We initialize 100 individuals in the exposed state in region 1. The COVID-19 test false positive rate α and false negative rate $1 - \beta$ are based on [45]; the infection rate λ_S is an assumption; mean incubation period $1/\lambda_E$, and mean recovery time $1/\lambda_R$ are consistent with the Center for Disease Control and Prevention (CDC) data [46, 47]; while the mortality rate λ_D is based on publicly reported data [48]. The values for ε_1^t through ε_4^t are computed as described in Section 2.2.2. On the other hand, in Section 4.3, the methodology was applied to real COVID-19 epidemic data.

Table 1. Parameter values for numerical experiments (section 4.1, 4.2).

| Parameter | Value | Reference |
|--|---------------------|---------------|
| Test false positive rate (α) | 0.01 | [45] |
| Test false negative rate ($1 - \beta$) | 0.15 | [45] |
| Infection rate (λ_S) | 0.4 | Assumption |
| Mean incubation period ($1/\lambda_E$) | 10 (days) | [46] |
| Mean recovery time ($1/\lambda_R$) | 14 (days) | [47] |
| Mortality rate (λ_D) | 0.01 | [48] |
| $\varepsilon_1^t, \varepsilon_2^t, \varepsilon_3^t, \varepsilon_4^t$ | 0.03, 0.1, 0.1, 0.2 | Section 2.2.2 |

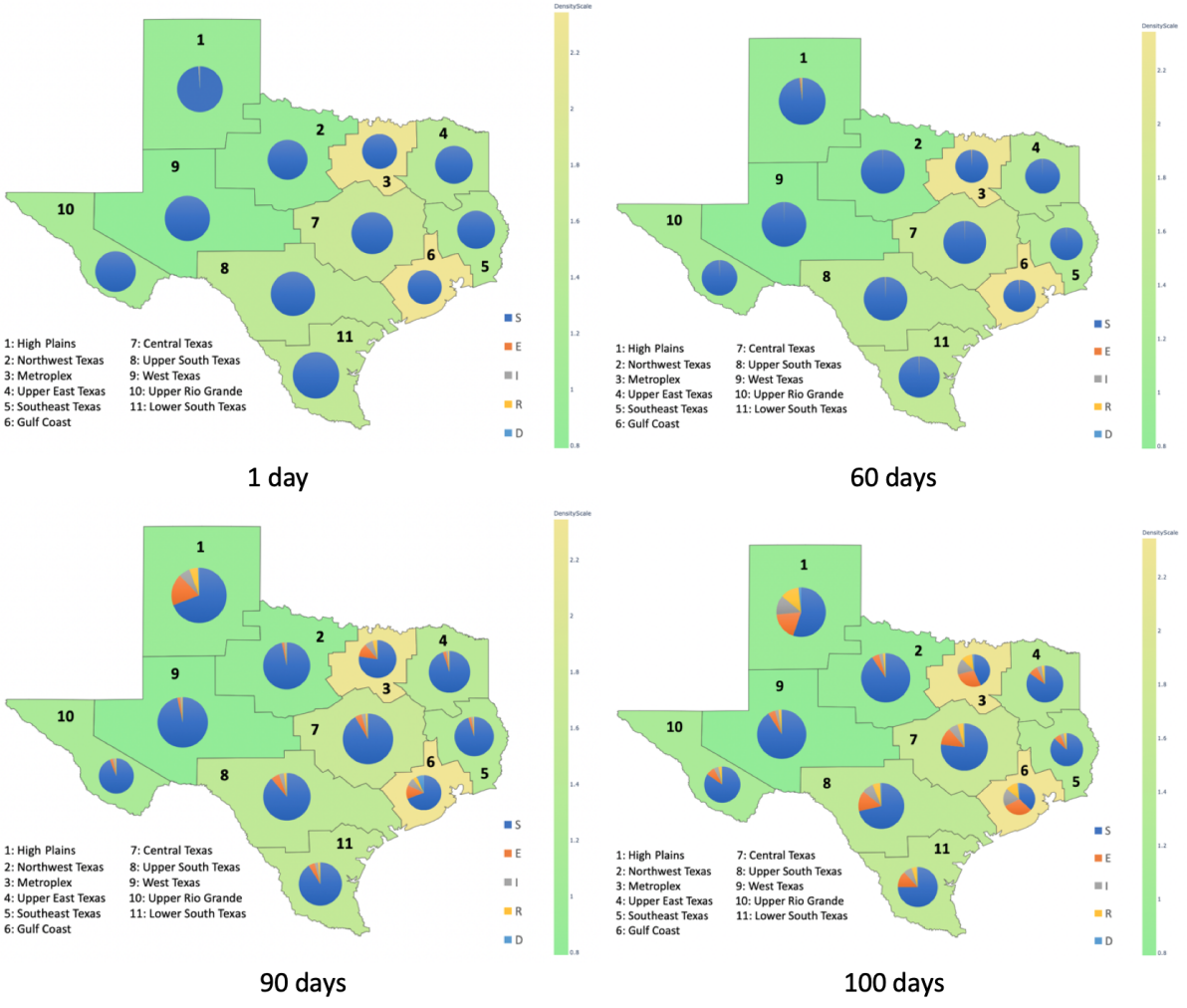


Figure 1. Prediction of COVID-19 spread over the state of Texas after 1, 60, 90, and 100 days from initial infection.

4.1. Forward prediction of epidemic dynamics

Here we illustrate the proposed model's ability to predict the spatial and temporal dynamics of epidemic, assuming the parameter values displayed in Table 1. We ran the model for a period of 200 days. Figure 1 displays the results at day 1, day 60, day 90, and day 100. In this simulation, the epidemic is assumed to have originated in region 1 with only a few infective individuals, as can be seen in the upper left diagram of Figure 1. After 60 days, the pandemic has spread to other regions, but it is still very limited. However, after 90 days, i.e., only 30 days after the previous snapshot, the epidemic has spread much more widely, especially in the originating region 1, as well as regions with large populational density, e.g., region 3 and region 6, which is where Dallas and Houston are located, respectively. Finally, after only another 10 days, the number of infected individuals has almost doubled. These results are in agreement with the fact that an epidemic is easier to control at an early stage, and will spread at an exponential rate if not controlled early [49–51]. These predictions underscore the need for early-stage public health interventions, such as social distancing, mask wearing, or a limited lockdown in the critical originating region 1.

4.2. State and parameter estimation from synthetic data

In this section, we investigate the ability of the proposed maximum-likelihood adaptive filtering methodology in recovering both the hidden state and all unknown parameters of the pandemic from a synthetic time series of reported new cases and deaths. The synthetic data allow us to evaluate the performance of the methodology against the simulated ground truth.

In the first experiment, we assume that the parameter values are known (see Table 1) and evaluate the ability of the filtering methodology to recover the hidden SEIRD state from the synthetic times series. Although in practice not all parameters would be known, this experiment allows us to evaluate the pure state estimation capabilities of the algorithm. Figure 2 displays the results over region 1 (the behavior was similar over all other regions). We can see that the maximum-likelihood adaptive filter can track the state evolution remarkably well. This is confirmed quantitatively by the relative error, defined as

$$\text{relative error} = \frac{|\text{estimated state} - \text{true state}|}{\text{true state}},$$

which is displayed in Table 2 for a few time snapshots. We can see that the errors are small, with the exception of early in the epidemic, when not enough data has been accumulated yet.

Table 2. Relative error with all parameter values known.

| Parameter | day 10 | day 25 | day 40 | day 55 | day 70 | day 85 | day 100 | day 115 | day 130 |
|---------------------|--------|--------|--------|--------|--------|--------|---------|---------|---------|
| Susceptible (S) | 0.001 | 0.003 | 0.010 | 0.038 | 0.104 | 0.139 | 0.096 | 0.068 | 0.052 |
| Exposed (E) | 0.169 | 0.175 | 0.137 | 0.089 | 0.004 | 0.071 | 0.099 | 0.104 | 0.107 |
| Infected (I) | 0.311 | 0.137 | 0.140 | 0.107 | 0.050 | 0.022 | 0.061 | 0.079 | 0.091 |
| Recovered (R) | 0.53 | 0.193 | 0.154 | 0.134 | 0.092 | 0.046 | 0.017 | 0.004 | 0.001 |
| Deceased (D) | 0.711 | 0.261 | 0.029 | 0.095 | 0.077 | 0.041 | 0.015 | 0.001 | 0.005 |

In the second experiment, we remove the assumption that the infection rate, mean recovery time, and mortality rate parameters are known, and evaluate the performance of the methodology in recovering

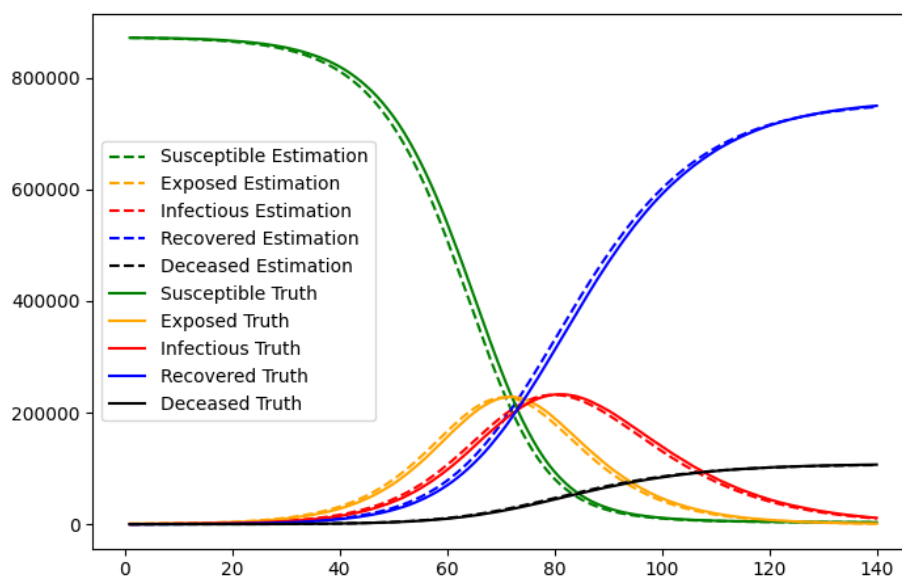


Figure 2. State estimation in region 1 with all parameter values known.

the values of these parameters. This is a difficult problem, since the states are also unknown, and the algorithm must perform simultaneous state and parameter estimation. Table 3 shows that the algorithm produced estimated parameter values that are close to the groundtruth values. Figure 3 displays state estimation results when the estimated parameters are plugged in for the unknown parameters. We can see that the results are not quite as good as the case where all parameters are known in Figure 2, but the proposed methodology is still able to track the state evolution relatively well. The relative errors in Table 4 is still very small in many cases. Once again, the error tends to be larger early in the epidemic because not enough data has been accumulated. In addition, some of the states, such as number of susceptible and exposed individuals, become very small near the end of the time interval, which makes the relative error increase.

Table 3. Parameter estimation with synthetic data.

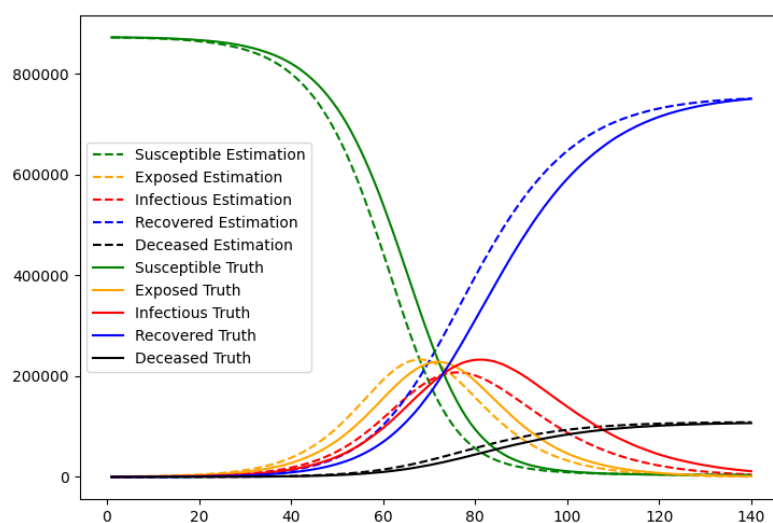
| Parameter | Groundtruth | Estimated Value |
|--------------------------------------|-------------|-----------------|
| Infection rate (λ_S) | 0.4 | 0.451 |
| Mean recovery time ($1/\lambda_R$) | 14 (days) | 12.05 (days) |
| Mortality rate (λ_D) | 0.01 | 0.0121 |

4.3. State and parameter estimation using Johns Hopkins University COVID-19 data

In this experiment, we demonstrate the performance of the proposed algorithm on Texas COVID-19 data from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) [52]. Due to the presence of delayed reporting, there is often incorrect data in the early part of

Table 4. Relative error analysis with all parameter values unknown.

| Parameter | day 10 | day 25 | day 40 | day 55 | day 70 | day 85 | day 100 | day 115 | day 130 |
|---------------------|--------|--------|--------|--------|--------|--------|---------|---------|---------|
| Susceptible (S) | 0.001 | 0.005 | 0.024 | 0.115 | 0.325 | 0.386 | 0.191 | 0.038 | 0.181 |
| Exposed (E) | 0.280 | 0.357 | 0.379 | 0.309 | 0.019 | 0.228 | 0.303 | 0.320 | 0.332 |
| Infected (I) | 0.299 | 0.191 | 0.261 | 0.241 | 0.051 | 0.209 | 0.369 | 0.469 | 0.546 |
| Recovered (R) | 0.775 | 0.449 | 0.470 | 0.495 | 0.400 | 0.217 | 0.094 | 0.034 | 0.009 |
| Deceased (D) | 0.684 | 0.013 | 0.369 | 0.483 | 0.409 | 0.236 | 0.112 | 0.050 | 0.024 |

**Figure 3.** State estimation in region 1 using estimated parameter values.

the week and corrections at the end of the week. To address this, we consider the time unit to be week, not day, and average the daily reported data (from Sunday to Saturday) to produce one data point for the week. The data used in this experiment is from the week of Oct 4th, 2020 to the week of Feb 1st, 2021, which is from the third (Fall 2020) wave of the COVID-19 epidemic in the United States.

In this experiment, we only use data from four large regions (3, 6, 7, 8), which contain Dallas, Houston, Austin and San Antonio, which are the most reliable data available. The population movement parameters c_{ij} of those four areas were calculated using the gravity model, as explained in Section 2.2.1, where the “gravitational constant” is set such that the probability of an individual moving between regions on a given day is equal to 1%. This produced the following calculated values for c_{ij} :

$$c = \begin{pmatrix} 1 & 0.0018 & 0.0032 & 0.0006 \\ 0.0018 & 1 & 0.003 & 0.0007 \\ 0.0032 & 0.003 & 1 & 0.0007 \\ 0.0006 & 0.0007 & 0.0007 & 1 \end{pmatrix}$$

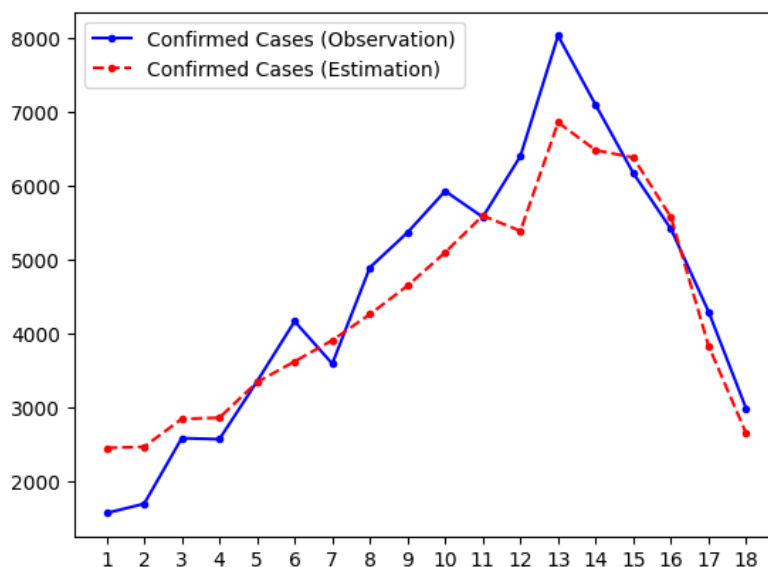


Figure 4. Prediction of new cases using JHU COVID-19 data over region 3. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.

We assume that only the test false positive rate, false negative rate and the mean incubation time are known. However, the value of the latter in Table 1 is based on a daily time unit, so here the value used is $\lambda_E = 10/7 = 0.7$. If desired, all estimated rates obtained by the algorithm in this experiment can be divided by 7 in order to obtain daily rates.

Table 5. Estimated parameters for Texas Fall 2020 wave using JHU COVID-19 data.

| Parameter | Region 3 | Region 6 | Region 7 | Region 8 |
|--------------------------------------|-----------|------------|------------|------------|
| Infection rate (λ_S) | 0.16 | 0.16 | 0.16 | 0.16 |
| Mean recovery time ($1/\lambda_R$) | 9.92 days | 10.52 days | 10.85 days | 10.95 days |
| Mortality rate (λ_D) | 0.0111 | 0.00992 | 0.0135 | 0.0112 |

In addition, unlike in the experiment with synthetic data, we have no groundtruth for the states or parameters. In order to evaluate the performance of the algorithm, we use the estimated values of states and parameters to predict the time series of observed data and check its agreement with the actual JHU data. These plots are displayed in Figures 4 and 5 for region 3 (the plots for other regions are in the Supplementary Information). In all cases, the state vector was initialized using the procedure described in Section 3.1, which produced the value $\mathbf{m}_{0|0} = (7717644, 1574, 1574, 0, 3591)$. Table 5 displays the estimated values of the parameters. These results indicate that our model can accurately estimate the hidden states and unknown parameters. The value for the infection rate is the same over all regions, by assumption; however, Figures 4 and 5 provide evidence that this is a reasonable assumption.

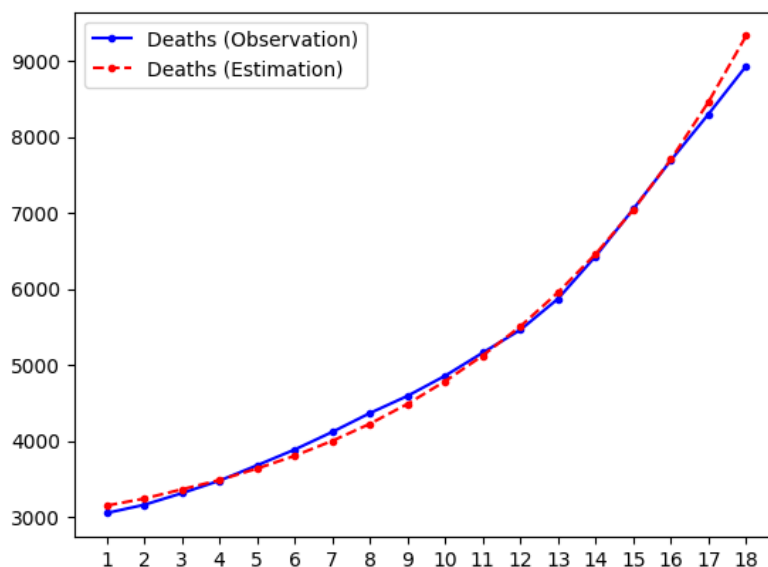


Figure 5. Prediction of cumulative deaths using JHU COVID-19 data over region 3. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.

4.4. Discussion

Compared with the standard SEIRD model, which assumes that the states of the epidemic can be observed directly, our stochastic state-space model realistically treats these states as hidden and only indirectly observed from the reported data. The proposed methodology is also able to incorporate detailed information about the testing procedure, such as false positive and false negative test rates and differential testing rate between symptomatic and asymptomatic patients. Contrary to standard models which consider reported cases as the true state of the epidemic, our model accounts for the impact of unobserved, undiagnosed cases on epidemic prediction. In so doing, our model provides a more accurate understanding of epidemic dynamics than models that ignore the effect of unobserved variables.

Our model has limitations, which can be addressed by future work. First, our model does not consider vital dynamics such as birth and death from other causes. Provided that the model is run over a short interval of time, such as an isolated wave in the epidemic, the results will not be affected too much; in addition, new born play an insignificant role in COVID-19 spread. Second, we assume that individuals infected with COVID-19 will develop life-long immunity. Though some epidemiological data have indicated evidence of COVID-19 reinfections, in most cases immunity persist for at least eight months. Therefore, we believe that this assumption has a marginal impact on our results, provided once again that the model is run over a short enough time interval. Third, our model ignored the impact of age on COVID-19 epidemiological parameters such as infection rate, disease severity (symptomatic vs asymptomatic), mortality rate, recovery rate, and testing rate. Age-specific infection rate and disease-induced complications have been observed during the ongoing COVID-19 pandemic

and considered to play an import role in disease transmission and the design of public health policies. Fourth, our model in its current form assumes that connection between the different regions can be modeled with a simple gravity commuting model [28–30]. This limitation can be addressed by using empirical population mobility data. Finally, Our model did not consider potential temporal changes of parameters over time, e.g., the mortality rate may be high and recovery rate low at the beginning of an outbreak because of lack of effective treatments, or public behavior, such as social distancing and mask wearing, may vary over time. Also, in the case of the Fall 2020 wave, it is known that mass vaccination began at the end of this wave (December 2020), which very likely had a major impact in the disease transmission parameters. Hence, the estimated parameter values reflect average rates over the time interval in consideration.

5. Conclusions

We proposed a novel stochastic metapopulation state-space model for COVID-19 transmission, based on a discrete-time SEIRD model, which is able to estimate the hidden epidemic states and transmission parameters from a noisy time series of reported epidemiological data, by applying unscented Kalman filtering (UKF), Maximum Likelihood (ML) adaptive filtering, and metaheuristic optimization. We reported results from a comprehensive set of experiments, using synthetic data and real epidemic data to demonstrate that our model can estimate parameters and predict future trends accurately and effectively. The proposed framework was applied to the state of Texas in the United States, using data from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. It can also be applied to any city, state, and nation by providing the necessary time series data (e.g., reported cases, deaths, and testing rates). The proposed methodology may provide a valuable tool for revealing the current epidemic status through accurate estimated states and disease transmission parameters, improving the ability of authorities to make informed decisions on public health measures to contain the disease.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

Martial Ndeffo-Mbah acknowledges funding from the National Science Foundation RAPID Award [grant number DEB-2028632].

References

1. M. L. C. Degli Atti, S. Merler, C. Rizzo, M. Ajelli, M. Massari, P. Manfredi, et al., Mitigation measures for pandemic influenza in italy: an individual based model considering different scenarios, *PloS one*, **3** (2008), e1790.
2. L. Perez, S. Dragicevic, An agent-based approach for modeling dynamics of contagious disease spread, *Int. J. Health Geogr.*, **8** (2009), 1–17.

3. E. Hunter, B. Mac Namee, J. Kelleher, An open-data-driven agent-based model to simulate infectious disease outbreaks, *PloS One*, **13** (2018), e0208775.
4. S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, M. Prokopenko, Modelling transmission and control of the COVID-19 pandemic in australia, *Nat. Commun.*, **11** (2020), 1–13.
5. D. L. Chao, A. P. Oron, D. Srikrishna, M. Famulare, Modeling layered non-pharmaceutical interventions against sars-cov-2 in the united states with corvid, *medRxiv*, 2020.
6. J. R. Koo, A. R. Cook, M. Park, Y. Sun, H. Sun, J. T. Lim, et al., Interventions to mitigate early spread of sars-cov-2 in singapore: a modelling study, *Lancet Infect. Dis.*, **20** (2020), 678–688.
7. M. Kretzschmar, G. Rozhnova, M. van Boven, Isolation and contact tracing can tip the scale to containment of covid-19 in populations with social distancing, *Available at SSRN 3562458*, 2020.
8. C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, G. Hart, K. Rosenfeld, et al., Covasim: an agent-based model of COVID-19 dynamics and interventions, *medRxiv*, 2020.
9. D. Balcan, B. Goncalves, H. Hu, J. J. Ramasco, V. Colizza, A. Vespignani, Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model, *J. Comput. Sci.*, **1** (2010), 132–145.
10. V. Dukic, H. F. Lopes, N. G. Polson, Tracking epidemics with state-space seir and google flu trends, *Unpublished manuscript*, 2012.
11. D. Osthus, K. S. Hickmann, P. C. Caragea, D. Higdon, S. Y. Del Valle, Forecasting seasonal influenza with a state-space sir model, *Ann. Appl. Stat.*, **11** (2017), 202.
12. E. Sebastian, P. Victor, A state space approach for sir epidemic model, *Int. J. Differ. Equ.*, **12** (2017), 79–87.
13. M. J. Keeling, T. D. Hollingsworth, J. M. Read, Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19), *J. Epidemiol. Community Health*, **74** (2020), 861–866.
14. R. Sameni, Mathematical modeling of epidemic diseases; a case study of the COVID-19 coronavirus, *arXiv preprint arXiv:2003.11371*, 2020.
15. A. Godio, F. Pace, A. Vergnano, Seir modeling of the italian epidemic of sars-cov-2 using computational swarm intelligence, *Int. J. Environ. Res. Publ. Health.*, **17** (2020), 3535.
16. G. Kobayashi, S. Sugawara, H. Tamae, T. Ozu, Predicting intervention effect for COVID-19 in japan: state space modeling approach, *BioScience Trends*, 2020.
17. W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, *Proc. Math. Phys. Eng. Sci.*, **115** (1927), 700–721.
18. G. Hooker, S. P. Ellner, L. D. V. Roditi, D. J. Earn, Parameterizing state–space models for infectious disease dynamics by generalized profiling: measles in ontario, *J. R. Soc. Interface*, **8** (2011), 961–974.
19. S. Zhong, Q. Huang, D. Song, Simulation of the spread of infectious diseases in a geographical environment, *Sci. China Earth Sci.*, **52** (2009), 550–561.
20. G. Albi, L. Pareschi, M. Zanella, Control with uncertain data of socially structured compartmental epidemic models, *J. Math. Biol.*, **82** (2021), 1–41.

21. G. Bertaglia, L. Pareschi, Hyperbolic compartmental models for epidemic spread on networks with uncertain data: application to the emergence of COVID-19 in italy, *arXiv preprint arXiv:2105.14258*, 2021.
22. G. Bertaglia, W. Boscheri, G. Dimarco, L. Pareschi, Spatial spread of COVID-19 outbreak in italy using multiscale kinetic transport equations with uncertainty, *arXiv preprint arXiv:2106.07262*, 2021.
23. W. Boscheri, G. Dimarco, L. Pareschi, Modeling and simulating the spatial spread of an epidemic through multiscale kinetic transport equations, *Math. Mod. Methods Appl. Sci.*, (2021), 1–39.
24. T. Rapolu, B. Nutakki, T. S. Rani, S. D. Bhavani, A time-dependent seird model for forecasting the COVID-19 transmission dynamics, *medRxiv*, 2020.
25. E. Loli Piccolomini, F. Zama, Monitoring italian COVID-19 spread by a forced seird model, *PLoS One*, **15** (2020), e0237417.
26. I. Korolev, Identification and estimation of the seird epidemic model for COVID-19, *J. Econom.*, **220** (2021), 63–85.
27. V. Tiwari, N. Bisht, N. Deyal, Mathematical modelling based study and prediction of COVID-19 epidemic dissemination under the impact of lockdown in india, *medRxiv*, 2020.
28. G. K. Zipf, The $p \propto 1/d$ hypothesis: on the intercity movement of persons, *Am. sociol. Rev.*, **11** (1946), 677–686.
29. J. Truscott, N. M. Ferguson, Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling, *PLoS Comput. Biol.*, **8** (2012), e1002699.
30. Q. Chen, J. Yan, H. Huang, X. Zhang, Correlation of the epidemic spread of COVID-19 and urban population migration in the major cities of hubei province, china, *Transp. Safety Environ.*, **3** (2021), 21–35.
31. W. E. Allen, H. Altae-Tran, J. Briggs, X. Jin, G. McGee, A. Shi, et al., Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing, *Nat. Hum. Behav.*, **4** (2020), 972–982.
32. D. Buitrago-Garcia, D. Egli-Gany, M. J. Counotte, S. Hossmann, H. Imeri, A. M. Ipekci, et al., Occurrence and transmission potential of asymptomatic and presymptomatic sars-cov-2 infections: A living systematic review and meta-analysis, *PLoS Med.*, **17** (2020), e1003346.
33. E. A. Wan, R. Van Der Merwe, The unscented kalman filter for nonlinear estimation, in *Proc. IEEE 2000 Adaptive Syst. Signal Process., Commun. Control Symposium (Cat. No. 00EX373)*, Ieee, 2000, 153–158.
34. C. J. Bastos Filho, F. B. de Lima Neto, A. J. Lins, A. I. Nascimento, M. P. Lima, A novel search algorithm based on fish school behavior, in *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, IEEE, 2008, 2646–2651.
35. C. Bastos-Filho, D. Nascimento, An enhanced fish school search algorithm, in *Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI & CBIC), 2013 BRICS Congress on*, IEEE, 2013, 152–157.

36. Y. Tan, F. L. Neto, U. Braga-Neto, Pallas: Penalized maximum likelihood and particle swarms for inference of gene regulatory networks from time series data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.
37. D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*, John Wiley & Sons, 2006.
38. S. Särkkä, *Bayesian filtering and smoothing*, Number 3, Cambridge University Press, 2013.
39. K. Ito, K. Xiong, Gaussian filters for nonlinear filtering problems, *IEEE Trans. Automat. Contr.*, **45** (2000), 910–927.
40. Y. Wu, D. Hu, M. Wu, X. Hu, A numerical-integration perspective on gaussian filters, *IEEE Trans. Signal Process.*, **54** (2006), 2910–2921.
41. J. Kokkala, A. Solin, S. Särkkä, Sigma-point filtering and smoothing based parameter estimation in nonlinear dynamic systems, *arXiv preprint arXiv:1504.06173*, 2015.
42. A. R. Yıldız, A novel particle swarm optimization approach for product design and manufacturing, *Int. J. Adv. Manuf. Technol.*, **40** (2009), 617–628.
43. I. Mukherjee, P. K. Ray, A review of optimization techniques in metal cutting processes, *Comput. Ind. Eng.*, **50** (2006), 15–34.
44. M. Madić, D. Marković, M. Radovanović, Comparison of meta-heuristic algorithms for solving machining optimization problems, *Facta universitatis-series: Mech. Eng.*, **11** (2013), 29–44.
45. T. Asai, COVID-19: accurate interpretation of diagnostic tests—a statistical point of view, 2020.
46. Centers for Disease Control and Prevention, Interim clinical guidance for management of patients with confirmed coronavirus disease (COVID-19), <https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html>, 2021, (accessed 22-July-2021).
47. Centers for Disease Control and Prevention, Interim guidance on ending isolation and precautions for adults with COVID-19, <https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html>, 2021,(accessed 22-July-2021).
48. World Health Organization. Estimating mortality from COVID-19, <https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19>, 2020, (accessed 22-July-2021).
49. C. Fraser, S. Riley, R. M. Anderson, N. M. Ferguson, Factors that make an infectious disease outbreak controllable, *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004), 6146–6151.
50. J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, et al., Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts, *Lancet Glob. Health*, **8** (2020), e488–e496.
51. M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, et al., The effect of human mobility and control measures on the COVID-19 epidemic in china, *Science*, **368** (2020), 493–497.
52. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, *Lancet Infect. Dis.*, **20** (2020), 533–534.

Supplementary Information

1. Calculation of the covariance matrix \mathbf{Q} of the process noise.

Covariance matrix \mathbf{Q}_t is a 5 by 5 square matrix giving the covariance between each pair of elements of the state process noise vector $(n_{S_i}^t, n_{E_i}^t, n_{I_i}^t, n_{R_i}^t, n_{D_i}^t)$ for a specific region i at time t .

Diagonal elements of the covariance matrix contain the variances of each variable, which are calculated as:

$$\begin{aligned} \text{Var}(n_{S_i}^t) &= \text{Var}\left(\sum_{j=1}^G N_{S_{i,j}}^t\right) \\ &= \sum_j S_i^t \frac{c_{ij} \lambda_S I_j^t}{N_j} \left(1 - \frac{c_{ij} \lambda_S I_j^t}{N_j}\right) - \sum_{j_1} \sum_{j_2} S_i^t \frac{c_{ij_1} \lambda_S I_{j_1}^t}{N_{j_1}} \frac{c_{ij_2} \lambda_S I_{j_2}^t}{N_{j_2}}, \quad j_1 \neq j_2 \\ &= \sum_j S_i^t \frac{c_{ij} \lambda_S I_j^t}{N_j} - \sum_{j_1} \sum_{j_2} S_i^t \frac{c_{ij_1} \lambda_S I_{j_1}^t}{N_{j_1}} \frac{c_{ij_2} \lambda_S I_{j_2}^t}{N_{j_2}} \end{aligned} \quad (5.1)$$

$$\begin{aligned} \text{Var}(n_{E_i}^t) &= \text{Var}\left(\sum_{j=1}^G N_{S_{i,j}}^t\right) + \text{Var}(N_{E_i}^t) \\ &= \sum_j S_i^t \frac{c_{ij} \lambda_S I_j^t}{N_j} - \sum_{j_1} \sum_{j_2} S_i^t \frac{c_{ij_1} \lambda_S I_{j_1}^t}{N_{j_1}} \frac{c_{ij_2} \lambda_S I_{j_2}^t}{N_{j_2}} + E_i^t \lambda_E (1 - \lambda_E) \end{aligned} \quad (5.2)$$

$$\begin{aligned} \text{Var}(n_{I_i}^t) &= \text{Var}(N_{E_i}^t) + \text{Var}(N_{R_i}^t) + \text{Var}(N_{D_i}^t) + 2\text{Cov}(N_{R_i}^t, N_{D_i}^t) \\ &= E_i^t \lambda_E (1 - \lambda_E) + I_i^t \lambda_R (1 - \lambda_R) + I_i^t \lambda_D (1 - \lambda_D) - 2I_i^t \lambda_R \lambda_D \\ &= E_i^t \lambda_E (1 - \lambda_E) + I_i^t (\lambda_R + \lambda_D) (1 - (\lambda_R + \lambda_D)) \end{aligned} \quad (5.3)$$

$$\text{Var}(n_{R_i}^t) = \text{Var}(N_{R_i}^t) = I_i^t \lambda_R (1 - \lambda_R) \quad (5.4)$$

$$\text{Var}(n_{D_i}^t) = \text{Var}(N_{D_i}^t) = I_i^t \lambda_D (1 - \lambda_D) \quad (5.5)$$

The off-diagonal elements contain the covariances of each pair of variables, which are calculated as:

$$\text{Cov}(n_{S_i}^t, n_{E_i}^t) = -\text{Var}(n_{S_i}^t) \quad (5.6)$$

$$\text{Cov}(n_{E_i}^t, n_{I_i}^t) = -\text{Var}(N_{E_i}^t) \quad (5.7)$$

$$\begin{aligned} \text{Cov}(n_{I_i}^t, n_{R_i}^t) &= -\text{Var}(N_{R_i}^t) - \text{Cov}(N_{R_i}^t, N_{D_i}^t) \\ &= -I_i^t \lambda_R (1 - \lambda_R) + I_i^t \lambda_R \lambda_D \\ &= -I_i^t \lambda_R (1 - (\lambda_R + \lambda_D)) \end{aligned} \quad (5.8)$$

$$\begin{aligned} \text{Cov}(n_{I_i}^t, n_{D_i}^t) &= -\text{Var}(N_{D_i}^t) - \text{Cov}(N_{R_i}^t, N_{D_i}^t) \\ &= -I_i^t \lambda_D (1 - \lambda_D) + I_i^t \lambda_R \lambda_D \\ &= -I_i^t \lambda_D (1 - (\lambda_R + \lambda_D)) \end{aligned} \quad (5.9)$$

$$\text{Cov}(n_{R_i}^t, n_{D_i}^t) = -I_i^t \lambda_R \lambda_D \quad (5.10)$$

The remaining elements will be zero.

2. Covariance of the observation noise (\mathbf{R})

Covariance matrix R_t is a 2 by 2 square matrix giving the covariance between each pair of elements of the observation noise vector $(v_{p_i}^t, v_{q_i}^t)$ for a specific region i at time t .

Similarly, diagonal elements of the covariance matrix contain the variances of each variable, but all the off-diagonal elements will be zero.

$$\begin{aligned} \text{var}(v_{p_i}^t) &= \varepsilon_1^t(1 - \varepsilon_3^t)(S_i^t + R_i^t)\alpha(1 - \alpha) + \varepsilon_2^t\varepsilon_3^t(S_i^t + R_i^t)\alpha(1 - \alpha) \\ &\quad + \varepsilon_2^t(1 - \varepsilon_4^t)I_i^t\beta(1 - \beta) + \varepsilon_1^t(\varepsilon_4^t I_i^t + E_i^t)\beta(1 - \beta) \\ \text{var}(v_{q_i}^t) &= D_i^t\beta(1 - \beta) \end{aligned} \quad (5.11)$$

3. Fish school search algorithm

Fish school search (FSS) is a population-based continuous optimization algorithm, which is inspired by the collective behavior of natural fish schools that expand and contract while searching for food. In the FSS algorithm, the objective is to find a model that maximizes a given score or fitness. Each candidate model, i.e., each candidate parameter vector θ , corresponds to a particle or “fish”. An ensemble of candidate solutions is a “school,” which is iteratively updated by a series of biologically-inspired moves, namely, (1) an individual movement operator for each fish (a small random perturbation), (2) a feeding operator that updates the weight of all fish based on the fitness improvement from the previous step, (3) a collective instinctive movement operator that makes the fish that had successful individual movements influence the collective direction of movement of the school, and (4) a collective volitive movement operator when the fish move in concert, depending on whether the fish school is successful after the previous steps, i.e., its total weight increases, or not. If the fish school is successful, then it should contract, changing from exploration to exploitation mode. Otherwise, it should expand in order to explore the space more. This process is repeated for either a fixed number of iterations or until there is no significant improvement in the score.

4. State and parameter estimation using Johns Hopkins University COVID-19 data

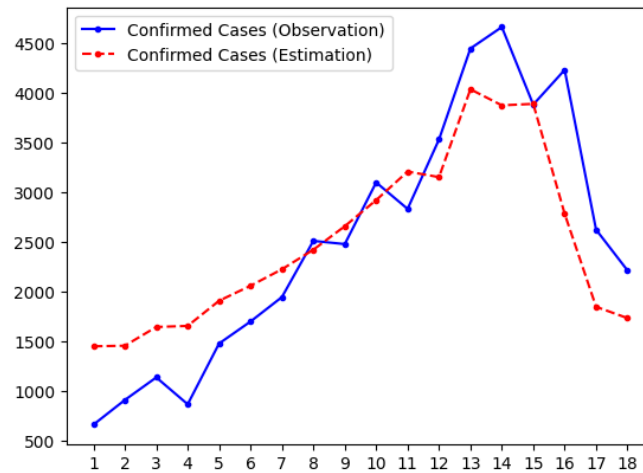


Figure 6. Prediction of new cases using JHU COVID-19 data over region 6. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.

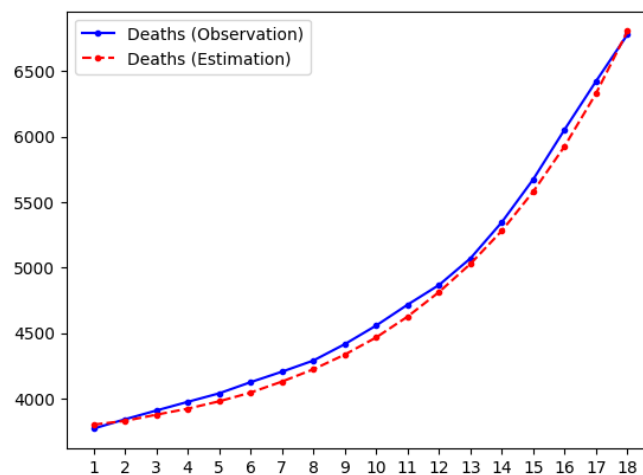


Figure 7. Prediction of cumulative deaths using JHU COVID-19 data over region 6. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.

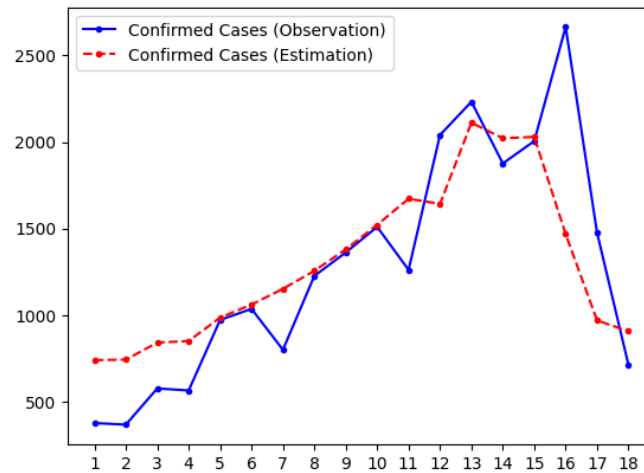


Figure 8. Prediction of new cases using JHU COVID-19 data over region 7. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.

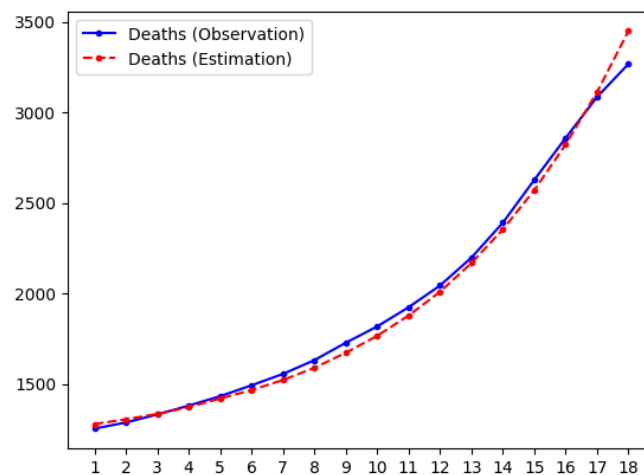


Figure 9. Prediction of cumulative deaths using JHU COVID-19 data over region 7. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.

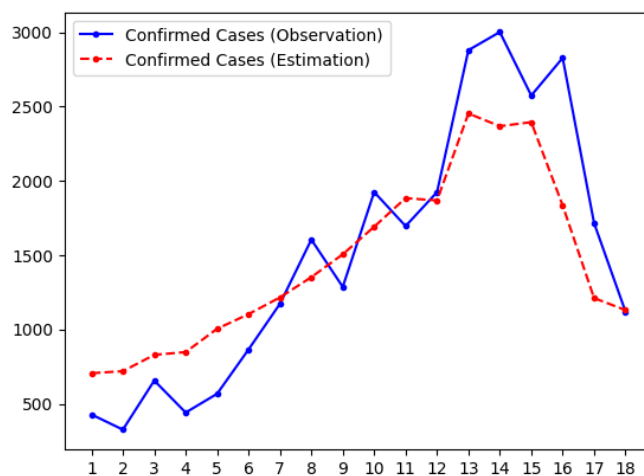


Figure 10. Prediction of new cases using JHU COVID-19 data over region 8. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.

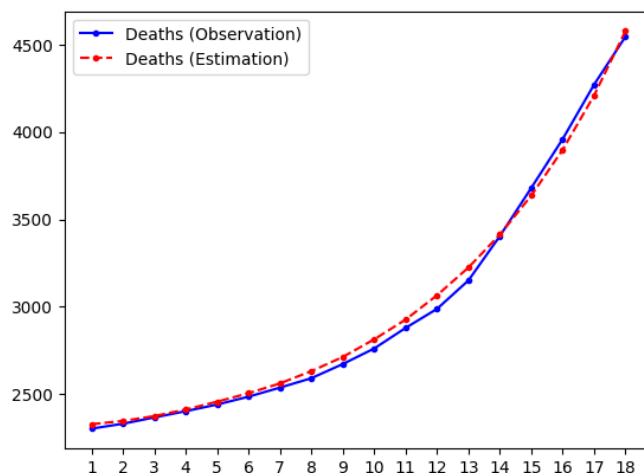


Figure 11. Prediction of cumulative deaths using JHU COVID-19 data over region 8. The blue line represents the actual observed data, the red dashed line are the predicted values in the time interval where the observed data is known.



AIMS Press

© 2021 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)