COVID-19 infection data encode a dynamic reproduction number in response to policy decisions with secondary wave implications

Supplemental Material

Michael A. Rowland, Todd Swannack, Michael L. Mayo, Matthew Parno, Matthew Farthing, Ian Dettwiller, Glover George, Molly Reif, Jeffrey Cegan, Benjamin Trump, Igor Linkov, Brandon Lafferty, Todd Bridges

Model overview

The ERDC SEIR model simulates the number of infections in a state or metropolitan area. The model is a modification of classic SEIR models that distinguishes between reported and unreported cases, much like the approach employed by the Columbia University team (Li et al, 2020). The ERDC SEIR model however, also employees another compartment to represent people that are isolated from larger population and will not be exposed to the virus. The model parameters are fit independently for each state and metropolitan area to match time-series data of confirmed infection reports. This calibration process is repeated daily as new data becomes available and is performed on ERDC's high performance computing cluster. More details are provided below.

Model formulation

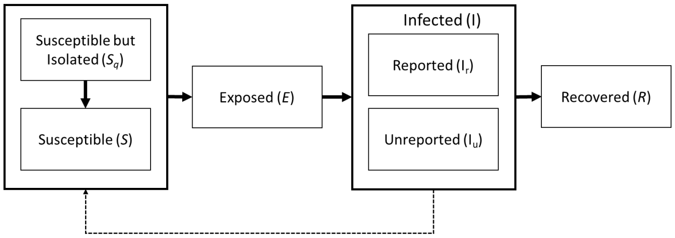

Figure S1. Conceptual model of the disease states. Healthy individuals are exposed to COVID-19 through infected individuals, and only a fraction of symptomatic individuals receive a test. All infected individuals "recover," which we use to account for those who become immune from further infection or die from the symptoms.

The progression of the disease by the transition of individuals in a population through 5 states, as illustrated in Figure S1. These states are: Susceptible (S), Exposed (E), Infected (I) or Recovered/Removed (R). The model makes a number of assumptions. First, it assumes that all individuals of a population can be treated identically; that is, there is no population variation in the transition rates between infection states. Another assumption of this type of model is that individuals are "well mixed," by which we assume that rates of disease progression do not depend upon the positions of any individuals within the subpopulations. Therefore, we do not account for any spatial heterogeneity in population density (e.g., cities vs. rural), which would otherwise directly affect the frequency of individual interactions that drive disease spread. Another important assumption we make is that populations are large enough that fluctuations driven by some of the individual variation in disease progression are small and can be ignored, which is common in population and epidemiological modeling. As a consequence, the individuals move between disease subpopulations (e.g., susceptible transitions to exposed) at deterministic rates. Finally, we have imposed a constraint in which individuals cannot return to a previous disease state, which reflects our presumption that those who recover are immune from further infection. Despite these various approximations and simplifying assumptions, the model is flexible enough for adaption to future scenarios. For example, population density can be accounted for using this conceptual approach by creating additional compartments to describe subpopulations at a scale under which further variation is ignored. However, such changes come at a cost of introducing additional model parameters, which is typically undesirable due to concerns of overfitting and non-uniqueness.

More specifically, these assumptions lead to four ordinary differential equations (ODEs), each of which define evolution of the four disease states:

$$
\begin{aligned}
\frac{dS_q}{dt} &= \left[\sum_{m=1}^{M} \gamma_m S \delta(t, t_m)\right] \\
\frac{dS}{dt} &= -\frac{\beta S I_r}{N} - \frac{\mu \beta S I_u}{N} - \left[\sum_{m=1}^{M} \gamma_m S \delta(t, t_m)\right] \\
\frac{dE}{dt} &= \frac{\beta S I_r}{N} + \frac{\mu \beta S I_u}{N} - \frac{E}{Z} \\
\frac{dI_r}{dt} &= \alpha \frac{E}{Z} - \frac{I_r}{D} \\
\frac{dI_u}{dt} &= (1-\alpha)\frac{E}{Z} - \frac{I_u}{D},
\end{aligned}
$$

Collectively, Eqs. [1]-[4] describe evolution of the number of individuals susceptible to the disease, *S*, exposed to the disease after contact with infected individuals, *E*, those individuals with reported/tested infections, I_r, and individuals with unreported infections, I_u. The overall number of ACTIVE cases being tracked by authorities each day is given by $I_r(t - T_{delay})$; i.e., the number of infected individuals being tracked (Eq. [3]) is shifted deterministically by an amount equal to the delay time, T_{delay}. The current model is entirely deterministic, which may be stochastically adjusted in the future to account for uncertainties in the values for some of the fitted parameters (e.g., number of exposed individuals at t=0, which is not measured).

Model Calibration and Uncertainty Quantification

There are 9 unknown quantities in Eqs. [1]-[4]: the initial conditions $[S_0, E_0, I_{r0}, I_{u0}]$ and the parameters $[\alpha, \beta, \mu, Z, D]$. To characterize these 9 variables, we use daily observations of the number of active cases, denoted here by $\{d_1, d_2, \ldots, d_T\}$, where T is the total number of days with observations. These observations are related to the value of reported infections $I_r(t)$ in the model equations. We employ a Bayesian formulation of this model calibration problem and define a posterior probability distribution over the 9 model variables given the observations. Maximizing the density of this posterior distribution is akin to nonlinear least squares and results in a single point estimate of the most likely model variables. These parameters are called the Maximum aposteriori (MAP) point. The posterior distribution however, also contains information about uncertainty in parameters that can result from observation noise or ill-posedness.[1] The posterior distribution depends on the nonlinear model in Eqs. [1]-[4] and does not therefore fall into a canonical family of distributions that can be sampled directly (e.g., Gaussian). We therefore

[1] Ill-posedness in this context means that multiple parameter settings can match the data equally well.

employ Markov chain Monte Carlo (MCMC) to generate samples of the posterior.[2] These samples are then propagated through the system of differential equations in Eqs. [1]-[4] to characterize uncertainty in our predictions.

Defining the posterior distribution requires us to define two things: (1) a prior probability distribution over the model variables that describes knowledge we may have about the variables before observing the data (e.g., known bounds, positivity, etc…) and (2) a statistical error model for the difference between the model predictions and observations. For the parameters $[\alpha, \beta, \mu, Z, D]$ we adapt the prior distribution used in the Columbia model. For the initial conditions $[S_0, E_0, I_{r0}, I_{u0}]$ we employ a combination of log-normal and uniform distributions to represent our prior knowledge. Notice that even though the prior distribution is constructed from canonical distribution families, the posterior will not be of a standard form because it depends on the model in Eqs. [1]-[4]. For the statistical error model, we make the common assumption that at any time t_i, the errors between the model predictions and observations, $e_i = I_r(t_i) - d_i$, are normally distributed with some constant variance σ^2. This variance accounts for both observation noise and the difference between our model and reality. We employ a hierarchical formulation and estimate σ^2 along with the model variables. An inverse-Gamma hyper prior is used.

[2] We employ a simple random walk Metropolis MCMC algorithm with a proposal based on the Laplace approximation of the posterior density at the MAP point.