# Modeling the Progression of SARS-CoV-2 in New York City

Research Statement

Jimmy Kelliher

**Abstract**

Below is an excerpt from my ongoing research on the impact of the COVID-19 pandemic on New York City. My work involves estimating the otherwise unobservable infection rate in New York City, that policymakers at City Hall might better understand the effects of mask usage, antigen testing, and lockdowns on infection trends going forward. Using a novel adaptation of the SIR model of epidemiology and a machine learning approach to hyperparameter estimation, I estimate the true path of various health compartments given the sparse data available. In particular, I find that each sick New Yorker would infect an average of 3.36 other New Yorkers, that 365 per 100,000 New Yorkers were incubating the virus as early as February 1, and that 280 per 100,000 infected New Yorkers would be hospitalized at some point in the course of their illness. Moreover, I estimate that more than one quarter of New Yorkers have been infected with COVID-19 as of November.

# The SEIHR Model

In a typical epidemic, a virus spreads through a closed community until it reaches an equilibrium infection rate or until it dies out. In the ongoing SARS-CoV-2 pandemic, we observe frequent changes in behavior and policy that endogenously impact that spread of the virus. These dynamics make it difficult for researchers, healthcare professionals, and policymakers to make sense of key metrics like daily case numbers, seroprevalence data, and infection-fatality rates, to name a few.

In this paper, we devise a simple, parsimonious model that can shed light on the progression of SARS-CoV-2 in spite of these challenges. We begin by adapting a multistage SEIR model to construct the $\mathrm{SE}^l\mathrm{I}^m\mathrm{H}^n\mathrm{R}$ model, which we will henceforth abbreviate as the SEIHR model. As we will discuss more deeply, the stages of each disease compartment are implemented purely for mathematical reasons and are therefore unobservable. For example, when we say that there are seven *stages of infection* in this model, we mean that the infection period is distributed as a sum of seven i.i.d. exponential random variables.

We begin by defining the system of ordinary differential equations that characterizes the model. Let $N \in \mathbb{N}$ denote the population of New York City, and let $t \in \mathbb{R}_+$ denote the number of days since the outbreak of SARS-CoV-2 in New York City. As usual, the susceptible population is given by the function $S : \mathbb{R}_+ \to [0, N]$, such that $S(t)$ denotes the number of individuals who are yet susceptible to the virus on day $t$. The removed population is given by the function $R : \mathbb{R}_+ \to [0, N]$, and it includes those who have recovered from the virus as well as those who have succumbed to it.

In a typical SEIR model, we additionally have functions $E, I : \mathbb{R}_+ \to [0, N]$ that denote those who are currently incubating virus and those who are actively infectious, respectively. We might also consider the number of individuals hospitalized on any given day, given by the function $H : \mathbb{R}_+ \to [0, N]$. In our multistage model, however, for a given disease compartment $X \in \{E, I, H\}$ and for some number $k_X \in \mathbb{N}$, we have sub-compartments $X_i : \mathbb{R}_+ \to [0, N]$ for $i \in \{1, \ldots, k_X\}$. Naturally, we can recover the original disease compartment via

$$X(t) = \sum_{i=1}^{k_X} X_i(t).$$

The reason we decompose these compartments is to generate a more realistic distribution of the time individuals spend in them. In particular, let $T_i$ be the random variable describing the time an individual spends in sub-compartment $X_i$, such that $T$ denotes the total time an individual spends in compartment $X$. It follows that if each $T_i \sim \mathrm{Exp}(\lambda_X)$, then $T \sim \mathrm{Erlang}(k_X, \lambda_X)$. Whereas the typical SEIR model is restricted to exponentially-distributed compartment times, our multistage SEIHR model enables us to accommodate a much larger class of distributions.

Because we are interested in health outcomes at a daily frequency, we do not consider vital dynamics in our model. The annual death rate in New York City hovers around 0.6%, giving us a daily death rate below 0.002%, which we can reasonably ignore. As such, our multistage SEIHR model is characterized by the following system of ordinary differential equations. For all $t \in \mathbb{R}_+$,

$$N = S(t) + E(t) + I(t) + H(t) + R(t), \text{ and}$$

$$\begin{aligned}
\dot{S}(t) &= -\beta \left(\frac{I(t)}{N}\right) C(t) S(t) \\
\dot{E}_1(t) &= \beta \left(\frac{I(t)}{N}\right) C(t) S(t) - \left(\frac{k_E}{\mu_E}\right) E_1(t) \\
\dot{E}_l(t) &= \left(\frac{k_E}{\mu_E}\right) E_{l-1}(t) - \left(\frac{k_E}{\mu_E}\right) E_l(t) && \text{for } l \in \{2, \ldots, k_E\} \\
\dot{I}_1(t) &= \left(\frac{k_E}{\mu_E}\right) E_{k_E}(t) - \left(\frac{k_I}{\mu_I} + \varepsilon\right) I_1(t) \\
\dot{I}_m(t) &= \left(\frac{k_I}{\mu_I}\right) I_{m-1}(t) - \left(\frac{k_I}{\mu_I} + \varepsilon\right) I_m(t) && \text{for } m \in \{2, \ldots, k_I\} \\
\dot{H}_1(t) &= \varepsilon I(t) - \left(\frac{k_H}{\mu_H}\right) H_1(t) \\
\dot{H}_n(t) &= \left(\frac{k_H}{\mu_H}\right) H_{n-1}(t) - \left(\frac{k_H}{\mu_H}\right) H_n(t) && \text{for } n \in \{2, \ldots, k_H\} \\
\dot{R}(t) &= \left(\frac{k_I}{\mu_I}\right) I_{k_I}(t) + \left(\frac{k_H}{\mu_H}\right) H_{k_H}(t).
\end{aligned}$$

For $X \in \{E, I, H\}$, $\mu_X \in \mathbb{R}_{++}$ gives the expected amount of time an individual will spend in compartment $X$. Parameter $\beta \in \mathbb{R}_{++}$ denotes the transmission rate of the virus and function $C : \mathbb{R}_+ \to [0, 1]$ denotes the time-varying contact rate of the population. Though they both relate to the ease of spread, we can think of $\beta$ as a function of the immutable characteristics of the virus (*Given that I come into contact with an infectious person, with what probability will they infect me?*), whereas $C$ is a function of social behavior (*With what probability will I come into contact with another person?*).

The parameter $\varepsilon \in \mathbb{R}_{++}$ denotes the infection-hospitalization rate. Because recorded case numbers early on in the pandemic could not possibly capture the true infection rate in the city, it is important to consider the hospital census, which is a much more reliable figure: whereas an infected individual might never be tested, a hospitalized individual almost certainly will. Given the above specification, we have innately assumed that an individual has a constant chance of being hospitalized throughout their infection period. We do this because there is not good data on the true distribution of the infection-to-hospitalization period, and because this specification affords us a convenient closed-form solution to the time-varying reproduction number.

# The Reproduction Number

The reproduction number indicates the expected number of individuals an index patient will go on to infect. Mathematically, it also characterizes the unstable equilibrium of our system of ordinary differential equations. To compute the reproduction number inherent to our model specification, we follow an article by van den Driessche. Let $k \equiv k_E + k_I$ denote the number of infectious disease compartments in our model. We compute the force of infection matrix $F$ and the infection transition matrix $V$ to be the following $k \times k$ matrices.

$$F = \frac{1}{N} \begin{pmatrix} 0 & \cdots & 0 & \beta C(t) S(t) & \cdots & \beta C(t) S(t) \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} \gamma_E & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\gamma_E & \gamma_E & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\gamma_E & \gamma_E & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \gamma_E & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -\gamma_E & \gamma_I + \varepsilon & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\gamma_I & \gamma_I + \varepsilon & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \gamma_I + \varepsilon & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & -\gamma_I & \gamma_I + \varepsilon & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -\gamma_I & \gamma_I + \varepsilon \end{pmatrix}$$

Above, in order to write our matrices more compactly, we have defined $\gamma_X \equiv k_X \mu_X^{-1}$ for each $X \in \{E, I, H\}$, which can be thought of as the transition rate between sub-compartments. Let $M$ denote the set of all $k \times k$ matrices over the complex field. We define the function $\rho : M \to \mathbb{R}_+$ to map a matrix to its spectral radius, which is a well-defined map by the fundamental theorem of algebra. It follows that the time-varying reproduction number $\mathcal{R} : \mathbb{R}_+ \to \mathbb{R}_+$ is given by

$$\begin{aligned} \mathcal{R}(t) &\equiv \rho(FV^{-1}) \\ &= \frac{1}{N} \left[ \frac{1}{\gamma_I} \sum_{m=1}^{k_I} \left( \frac{\gamma_I}{\gamma_I + \varepsilon} \right)^m \right] \beta C(t) S(t) \\ &= \frac{1}{N} \left[ \frac{1}{\varepsilon} \left( 1 - \left( \frac{\gamma_I}{\gamma_I + \varepsilon} \right)^{k_I} \right) \right] \beta C(t) S(t) \\ &= \left( \left( 1 - \left( \frac{k_I}{k_I + \mu_I \varepsilon} \right)^{k_I} \right) \frac{\beta}{\varepsilon} \right) \frac{C(t) S(t)}{N}, \end{aligned}$$

and we further define our baseline reproduction number to be

$$\mathcal{R}_0 \equiv \sup_{t \in \mathbb{R}} \mathcal{R}(t) = \left( 1 - \left( \frac{k_I}{k_I + \mu_I \varepsilon} \right)^{k_I} \right) \frac{\beta}{\varepsilon}.$$

# Hyperparameter Estimation

As it stands, we have empirical estimates of the pairs $(k_X, \mu_X)$ for each $X \in \{E, I\}$ from literature, and we can infer $(k_H, \mu_H)$ from available hospital data. Thus, to execute our model, we need to choose a pair $(\beta, \varepsilon)$ along with a vector of initial conditions

$$(E_1(0), \ldots, E_{k_E}(0), I_1(0), \ldots, I_{k_I}(0), H_1(0), \ldots, H_{k_H}(0), R(0)).$$

Given that we have $k_E + k_I + k_H + 1$ initial conditions, choosing the correct specification can be computationally expensive. To reduce the complexity of our parameter search, we choose to begin our simulation early in the pandemic, such that many of our disease compartments are close to zero. In particular, we let $t = 0$ correspond to 2020 February 1, and we assume that all disease compartments at $t = 0$ are empty with the exception of $E_1(0)$.

We have now reduced our parameter search to identifying the transmission rate $\beta$, the infection-hospitalization rate $\varepsilon$, and the initial number of exposed individuals $E_1(0)$. Recall that the reproduction number $\mathcal{R}_0$ is determined once we choose $\beta$ and $\varepsilon$. As such, our parameter search is equivalent to finding the triple $(E_1(0), \mathcal{R}_0, \varepsilon)$. We proceed via a Tree-structure Parzen estimator model, which incorporates machine learning into an otherwise intensive grid search procedure. To train the algorithm, we look at squared deviations in $H$ from the observed daily hospital census in the city from February through April. We find that deviations are minimized when

$$(E_1(0), \mathcal{R}_0, \varepsilon) \approx (30{,}669, 3.3591, 0.0028).$$

That is, we find that each sick New Yorker would infect an average of 3.36 other New Yorkers, that 365 per 100,000 New Yorkers were incubating the virus as early as February 1, and that 280 per 100,000 infected New Yorkers would be hospitalized at some point in the course of their illness. Given that New York is a densely populated city that relies heavily on public transit, it makes sense that the estimated reproduction number is higher than that of the global estimate of 3.00.

We chose this time period for the training data because we believe that our transit mobility data was a good proxy for the contact rate $C$ at the outset of the outbreak in the city. Now, however, as individuals are increasingly gathering at bars, at restaurants, and in homes, the relationship between mobility and the contact rate is less clear. As such, we then implement a second-stage optimization procedure to back out the trend in the contact rate given the above estimates and the trend in the hospital census.

# Primary Results

We now present the primary output of the model. That is, we chart the paths of the functions $C, S, E, I, H, R$ implied by our hyperparameter estimation, and we compare them to the seroprevalence data and the hospital census data to which they were trained.

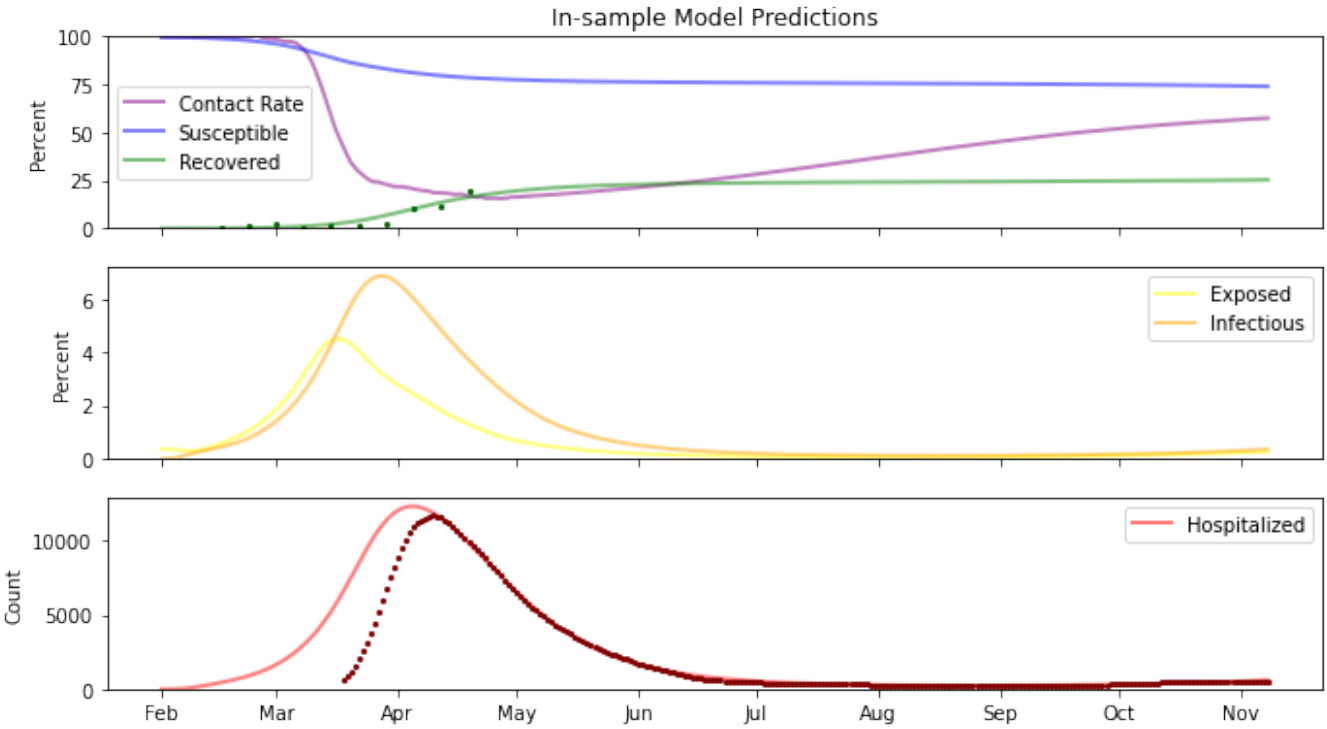


Figure 1: Solid lines denote model output; dotted lines denote seroprevalence data and hospital census data. The contact rate is a function of mobility through April; after April, the contact rate is backed out via the hospital census data.

Do note that $H$ reaches its first-wave peak more quickly than the observed hospital census. When we originally ran the simulation weighing equally the data from the hospital census, the model could not reconcile the gap between the sharp decline in mobility and the peak in the hospital census. As such, we give more weight to minimizing errors following the peak in the hospital census when executing the optimization procedure. We argue that this is acceptable because many individuals were likely misdiagnosed with the flu in early March, especially given our limited testing infrastructure at that time.

In particular, we find that more than 7% of New Yorkers were actively infected during the first-wave peak, and that more than 25% of New Yorkers had been infected with the virus by November. These numbers square well with models that employ cumulative death data in lieu of hospital census data (e.g., the work of Youyang Gu). Unlike many of those models, however, this framework enables us to forecast infection rates and hospitalization rates entirely via forecasting the contact rate. For next steps, we will forecast the trend in the contact rate in order to identify the expected second-wave peak given the absence of any further policy interventions.

# Appendix of Secondary Results

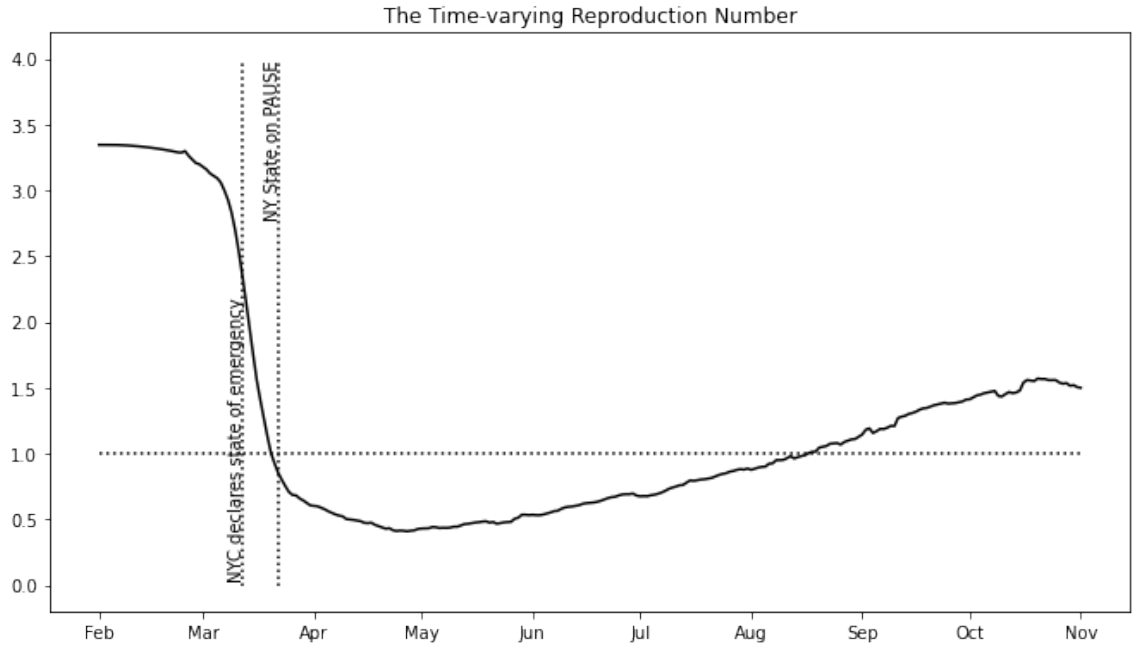


Figure 2: As shown, our model enables us to directly compute the time-varying reproduction number. Our estimates suggest that $\mathcal{R}$ has exceeded unity since the summer, likely due to reopening guidelines and isolation fatigue.

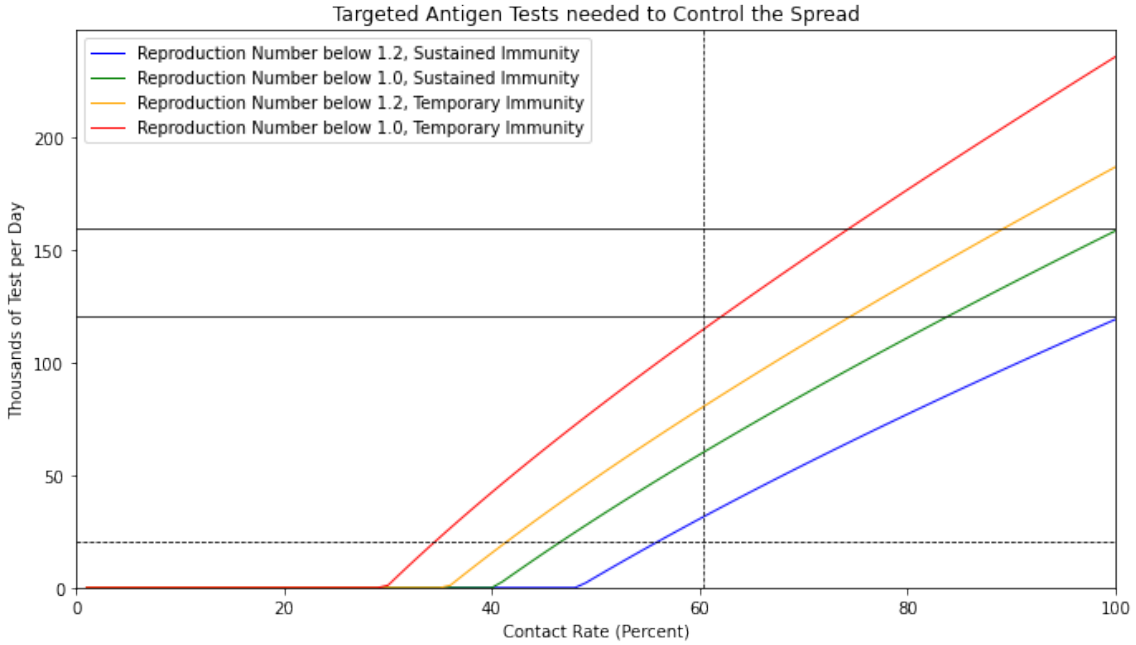


Figure 3: We also estimate the number of antigen tests needed to keep the spread of disease below particular values of $\mathcal{R}$. In particular, we assume that when an individual receives a positive test result, they immediately enter quarantine. The dashed lines intersect at the conditions in which we presently find ourselves.