Optimizing & Extending a COVID-19 Simulation in Python

Ross Bernstein, Sree Govindaprasad, Vishwaesh Rajiv, Brina Seidel, Ben Stadnick

May 2020

Abstract

With the COVID-19 pandemic creating unprecedented levels of uncertainty, the ability to efficiently simulate numerous possible outcomes has become crucial to informed decision-making. We improved the efficiency of one such simulation using Cython and multiprocessing, ultimately reducing its runtime by 66%. We apply this simulation to New York City using a novel dataset of movement through the city, and model numerous scenarios using our optimized code.

1 Introduction

As the novel coronavirus COVID-19 has ravaged populations around the world, policymakers everywhere have relied heavily on simulations of the disease's spread in deciding how to respond.

With so much about the virus still unknown, such simulations rely on numerous assumptions and parameters whose values are sometimes little more than guesses. Furthermore, there is a degree of randomness inherent to the spread of the virus. In light of this uncertainty, it is crucially important to be able to run numerous simulations as efficiently as possible.

In this project, we improved the efficiency of one such simulation released by Yeghikyan [2020]. We first compile data on transit and taxi use in New York City in order to apply the model to the current epicenter of the virus. We then work to improve the efficiency of the baseline code by changing how data is loaded into the program, compiling with Cython, and leveraging multiprocessing to run multiple simulations. Our optimized version reduces the runtime of the original by 66%. Finally, using the fastest version of the code, we model numerous potential scenarios for the spread of the virus through New York City.

2 Methodology: The Susceptible-Infected-Recovered Model

Yeghikyan [2020] uses the Susceptible-Infected-Recovered (SIR) method to estimate the portion of of the population that is vulnerable, sick, and immune at any given time. Such models are well-established in the field of epidemiology, with early versions proposed over a century ago [Brauer, 2017, Hamer, 1906]. Such models assume that those who have had the disease cannot be reinfected.

SIR models estimate the spread of the disease using estimates of urban mobility, the transmission rate β, and the recovery rate γ. Urban mobility is modeled using an origin-destination matrix M, where cell $m_{j,k}$ measures the share of the total population N_j originating in region j of the city who travels to region k by the end of the time period, as well as a parameter α_t measuring mobility at time t as a share of the normal baseline mobility. The transmission rate is modeled as a random variable, drawn separately from a gamma distribution for each location.

At time $t = 0$, infections are introduced over the regions based on user-defined criteria for population constraints and initial infection percentage. For example, we use a default infection percentage of 5% in regions with a population of less than 200.

For each subsequent time t, the number of individuals that are susceptible (S), infected (I), and recovered (R) in region j at time t is calculated as follows:

$$\begin{aligned}
S_{j,t+1} &= S_{j,t} - \frac{\beta_{j,t} S_{j,t} Ij,t}{N_j} - \frac{\alpha_t S_{j,t} \sum_k m_{j,k} \beta_{k,t} \frac{I_{k,t}}{N_k}}{N_j + \sum_k m_{j,k}} \\
I_{j,t+1} &= I_{j,t} + \frac{\beta_{j,t} S_{j,t} Ij,t}{N_j} + \frac{\alpha_t S_{j,t} \sum_k m_{j,k} \beta_{k,t} \frac{I_{k,t}}{N_k}}{N_j + \sum_k m_{j,k}} - \gamma I_{j,t} \\
R_{j,t+1} &= R_{j,t} + \gamma I_{j,t}
\end{aligned}$$

This basic framework provides an intuitive way to estimate the spread of the disease while allowing for numerous extensions to model more complex features of the world.

3 Data Collection

We approximated the movement of individuals throughout New York City by combining transit data, taxi data, and population data. The Metro Transit Authority (MTA) does not usually collect origin-destination data – after all, subway riders are only required to swipe their cards on the way *in*. However, such information is available for a representative sample of New Yorkers between May and November 2008 from a MTA survey [MTA, 2008]. We use the weekday ridership estimates from this survey. The Taxi and Limousine Commission (TLC) provides trip data for all rides taken in yellow taxis [TLC, 2009]. We use weekday trips from May 2009, the first year for which data is available, to match the MTA data as closely as possible.

For both types of trip information, we mark the census tract in which each trip started and ended and use population data from the 2010 US Census to determine how many individuals are in each census tract at the start of a given time period. By counting all rides from each pair of census tracts, we can estimate the number of individuals moving between tracts on a typical weekday. Finally, we added one to each cell of the matrix for smoothing purposes.

These numbers likely understate the true number of New Yorkers moving throughout the city. Nevertheless, they allow us to apply Yeghikyan [2020]'s original work on Yerevan to New York City, a highly salient city at the center of the global pandemic.

4 Improvements to Original Code

The original code is written using Numpy and relies heavily on vectorized operations, which means it is already pretty efficient. In order to see where improvements could be made, we used the line profiler to see which parts of the code were taking the most time. We found that loading in the origin-destination matrix from file accounted for approximately 40% of the execution time of the simulation. In order to speed up this part, we experimented with using a different library (Pandas) to load in the file as well as changing the format of the file from csv to feather format. Both of these changes drastically reduced the time it took to load the matrix in.

While changing the way the matrix was loaded into the simulation accounted for the majority of the speed up, we also tried compiling with Cython as well as removing unnecessary functionality, like the tqdm progress bar, from the simulation code. We found that the fastest performance was achieved by using Cython, feather loading, and removing tqdm. Making these changes reduced the runtime of a single simulation by 66% (from 12.7s to 4.26s).

We also added multiprocessing functionalities to this optimized code to enable many simulations to be run in parallel. We chose multiprocessing over threading for two reasons. First, Python's GIL lock only allows one thread to be in execution at any point in time. Second, and more importantly, SIR modeling is compute-intensive and needs no interaction with other servers or online APIs that introduce latency and delays (which would merit the use of concurrent threads). Multiprocessing also adds to ease-of-use. The user can run many simulations in parallel over a range of values for each random variable in the model, allowing the user to determine the interactions between these random variables and how they affect the infection curves over time.

Data Loading	**Cython**	**Includes tqdm**	**Serial**	**Multiprocessing**	**Single**
NumPy	No	Yes	49.5s ± 819ms	31.1s ± 2.48s	12.4s ± 291 ms
Pandas	No	Yes	28.1s ± 220ms	16.9s ± 625ms	4.9s ± 279ms
Feather	No	Yes	25.1s ± 736 ms	12.0s ± 1.14s	4.09s ± 82ms
Feather	Yes	Yes	18.8s ± 626ms	11.5s ± 240ms	4.08s ± 25.2ms
Feather	Yes	No	16.1s ± 331ms	8.52s ± 126ms	3.65s ± 53.9ms
Feather	Yes, with types	No	18.6s ± 522ms	9.35s ± 141ms	5.2s ± 343ms

Table 1: Optimizations and Run Time for Four Simulations and Single Simulation

5 Extending the Model: Simulating New York City's Future

We extend the original model by exploring the impact of 1) different estimates of the transmission rate of the disease, and 2) different estimates of transit use as the disease progresses.

In order to accurately simulate the transmission of COVID-19 in New York City, several parameters needed to be estimated. Based on the findings of Lin et al. [2020] and Peng et al. [2020], the transmission rate (β) was assumed to be within the range of (0.59, 1.68). This value represents the average number of secondary infections caused each day by an infected individual. For initial simulations, the median value 1.14 was used. The recovery rate (γ) was set to 0.2, which assumes a mean infectious period of 5 days. This interval only accounts for the period of time that an individual is able to infect others, and does not represent the total duration of the infection. It also takes into account that most symptomatic individuals eventually quarantine themselves and can no longer spread the virus.

To visualize the impact of transit intensity, simulations were run with different levels held constant over time. The results of these simulations are shown in Appendix Figure 1. The total share of the population that gets infected can be inferred from the upper limit of the 'Recovered' curve, since everyone that recovers has had the virus. We can see that at very low levels of transit intensity, an outbreak does not occur. Conversely, at higher levels, both the magnitude of infections and the speed of transmission increases with transit intensity. This makes sense because the transit intensity at time t multiplies the transmission rate at time t, so that a 50% reduction in transit intensity effectively lowers the transmission rate by 50%.

Next, a transit intensity vector that decreased over time was implemented to reflect changes in mobility patterns due to social distancing and government restrictions. Transit intensity was estimated over time using reports released by the MTA, which stated that ridership had decreased to 40% by March 17th, 8% by April 8th, and 7% by April 17th [Goldbaum, 2020, Rapier, 2020, Frost, 2020]. Daily percentages were extrapolated with a logistic function, as shown in Appendix Figure 2.

It is assumed that changes in mobility patterns across the city can be estimated by changes in the amount of people using public transit. The extent to which the public has followed social distancing guidelines and refrained from travel can be inferred by the lower limit of the transit intensity curve.

As shown in Appendix Figure 3, varying this value did not have a large impact on simulated outcomes.

Since substantial data was available to justify estimates for both the transit intensity and the recovery rate, the most crucial parameter left to be identified within a narrow range was the transmission rate (β). Prior estimates for this value are based on data gathered in other cities, and are not necessarily applicable to New York. This is due to the fact that the transmission rate is highly dependent on localized mobility patterns and population density. Thus, it can vary with respect to both location and time. Alternative approaches take this into account by estimating β as a function of time rather than a constant value. We designed our model to handle these complexities in two ways. First, β was implemented as a Gamma-distributed random variable, drawn separately for each zone, to simulate variations with respect to location. Second, since our transit intensity vector (α_t) acts as a modifier on β, $\beta \cdot \alpha_t$ effectively turns β into a function of time. It suffices to estimate the value of β at $t = 0$ by running simulations to find a starting value that causes total infections to approach the most recently published estimates, which now show that approximately 21% of the population of New York City has been infected. As shown in Appendix Figure 4, setting $\beta = 0.75$ came closest to simulating this outcome.

With the set of initial parameters established, we can now attempt to answer important questions using this model. The effectiveness of the government's reaction time in limiting the spread of the virus can be evaluated by shifting the transit intensity curve. Shifting the curve to the left represents faster intervention, while shifting to the right represents a slower response. Appendix Figure 5 shows that early intervention plays an important role. We can see that if the decline in transit intensity occurs one week earlier, the proportion of the population that is ultimately infected is reduced from 21.9%. to 15%, while delaying travel restrictions by one week only increases the infected proportion to 24.1%.

6 Conclusions

We have created a tool for efficiently simulating the transmission characteristics of COVID-19 in New York City that utilizes multiprocessing, Cython, and vectorized operations. Several important features were added to the basic SIR model, such as incorporating urban mobility patterns with a novel origin-destination flow matrix, and modeling transmission rates as a function of public transit usage over time. The core model remains simplistic enough that we can easily see the effects of varying key parameters, and evaluate policy decisions. Given that publicly available data is based on confirmed cases only, a simulation such as this may reflect the transmission characteristics of the virus more accurately than a model that relies more heavily on reported data.

There are many potential improvements that could be made to our model. For example, it could be converted into an SEIR model, which adds a compartment for individuals that have been exposed to the virus but are not yet infectious. This period, known as the *latent period*, may allow for more accurate estimates of transmission characteristics.

Taken as a whole, our work demonstrates how basic Python optimization techniques can be useful during a global pandemic.

References

Fred Brauer. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, 2 (2):113–127, May 2017. ISSN 2468-0427. doi: 10.1016/j.idm.2017.02.001.

Mary Frost. New York City subway ridership down 92 percent due to coronavirus, Apr 2020. URL `https://bit.ly/2W5OOkZ`.

Christina Goldbaum. M.T.A., citing huge drop in riders, seeks $4 billion virus bailout, Mar 2020. URL `https://nyti.ms/2WlYblW`.

W.H. Hamer. Epidemic disease in England: The evidence of variability and of persistency of type. *The Lancet*, 167(4305):569 – 574, 1906. ISSN 0140-6736. Originally published as Volume 1, Issue 4305.

Qianying Lin, Shi Zhao, Daozhou Gao, Yijun Lou, Shu Yang, Salihu Musa, Maggie Wang, Weiming Wang, Lin Yang, and Daihai He. A conceptual model for the outbreak of Coronavirus disease 2019 (COVID-19) in Wuhan, China with individual reaction and governmental action. *International Journal of Infectious Diseases*, 93, 03 2020. doi: 10.1016/j.ijid.2020.02.058.

Metropolitan Transit Authority (MTA). 2008 New York customer travel survey: Final report. 2008.

NY Taxi and Limousine Commission (TLC). New York City taxi trip data. 2009.

Liangrong Peng, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. Epidemic analysis of COVID-19 in China by dynamical modeling. 2020. doi: 10.1101/2020.02.16.20023465.

Graham Rapier. NYC subway ridership cratered 93% as all but essential workers stay home - and the transit system says it needs an additional $3.9 billion bailout to keep running, Apr 2020. URL `https://www.businessinsider.com/new-york-subway-rides-plunge-mta-requests-second-bailout-93-2020-4`.

Gevorg Yeghikyan. Modelling the coronavirus epidemic in a city with Python: Are cities prepared for epidemics? *Towards Data Science*, February 2020.

Appendix

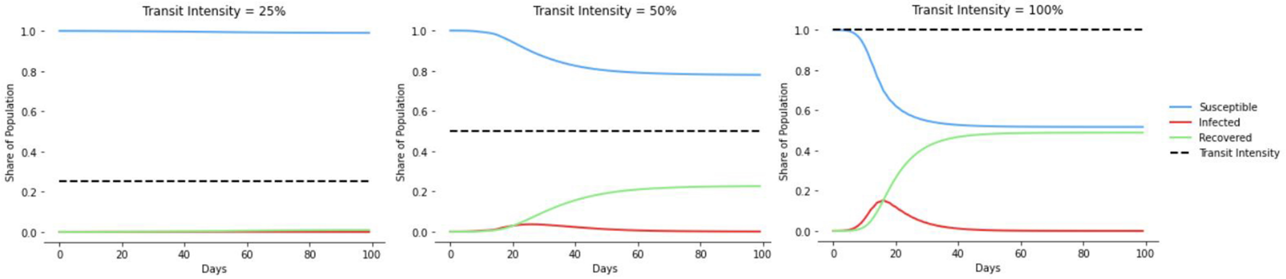

Figure 1: Simulations holding transit intensity constant at different levels

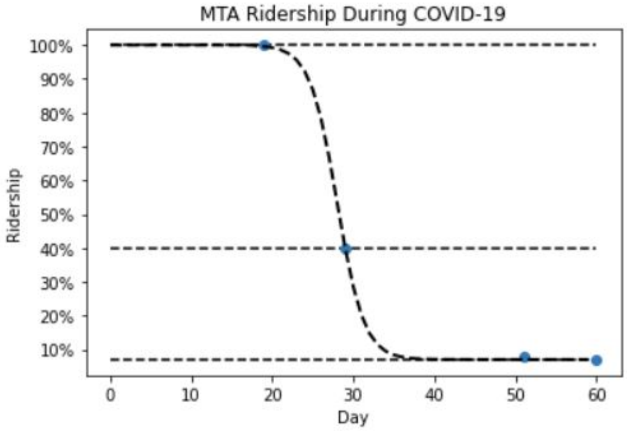

Figure 2: MTA ridership during COVID-19, as a percentage of pre-pandemic values. This curve will be used to adjust transmission rates, and help to simulate the effects of social distancing and government regulations.

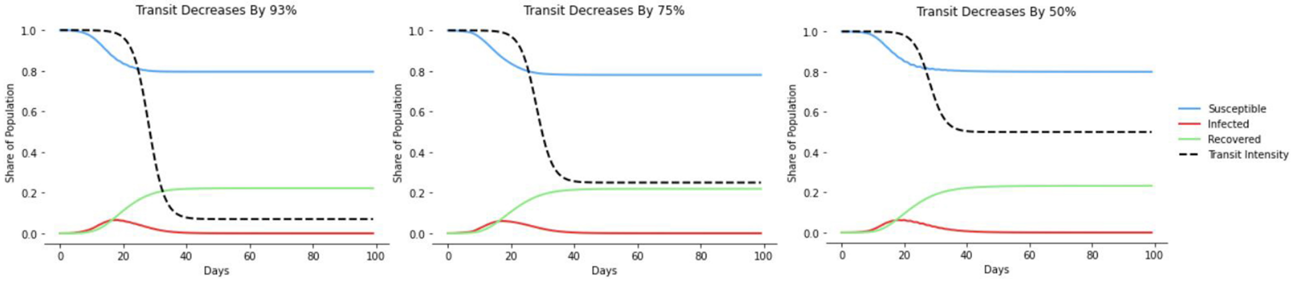

Figure 3: Simulations varying the magnitude of the drop in transit intensity. From left to right, the plots reflect total infected population shares of 22.2%, 21.9%, and 23.4%

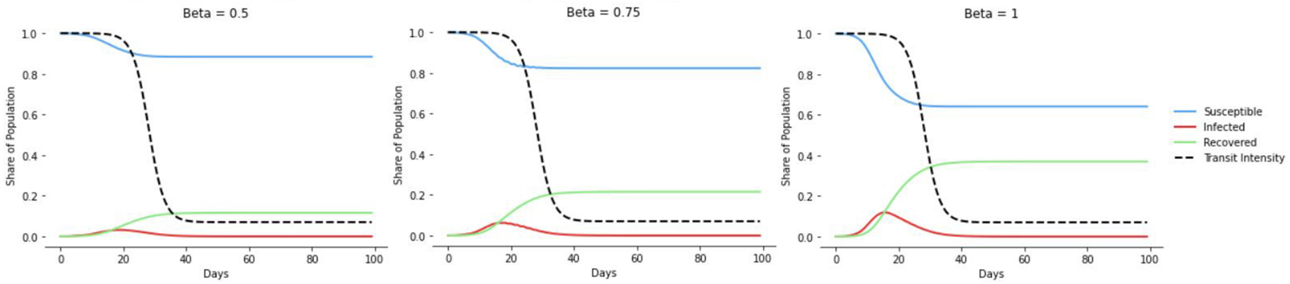

Figure 4: Fitting β to available data. From left to right, the plots reflect total infected population shares of 10.9%, 21.9%, and 40.2%.

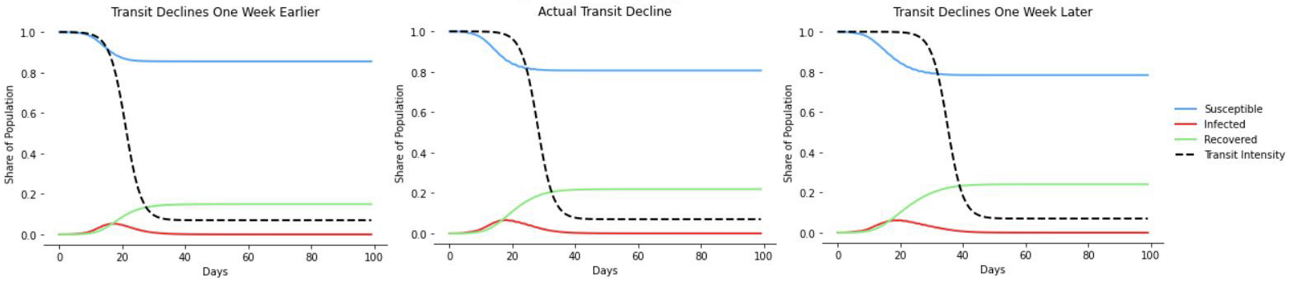

Figure 5: Simulations with varying start-dates for the intial decline in transit intensity. From left to right, the plots reflect total infected population shares of 15%, 21.9%, and 24.1%.