MOBILITY-GUIDED ESTIMATION OF COVID-19 TRANSMISSION RATES

Dylan Parker and Oleg Pianykh.

Author affiliations: Department of Statistics, Harvard University, Cambridge, Massachusetts (Dylan Parker); Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts (Oleg Pianykh).

Correspondence address: Correspondence to Oleg Pianykh, Department of Radiology, Massachusetts General Hospital, 25 New Chardon St., Boston, MA 02114 (phone 617-724-2618, email: OPIANYKH@mgh.harvard.edu).

Funding disclosures: none declared.

Conflict of interest: none declared.

Running head: Mobility-Guided Estimation of COVID-19 Transmission.

ABSTRACT

It is of critical importance to estimate changing transmission rates and their dependence on population mobility. A common approach to this problem involves fitting daily transmission rates using a Susceptive Exposed Infected Recovered (SEIR) model (regularizing them to avoid overfitting), and then computing the relationship between the estimated transmission rate and mobility. Unfortunately, there are often several, very different transmission rate trajectories that can fit the reported cases well, meaning that the choice of regularization determines the final solution (and thus the mobility-transmission rate relationship) selected by the SEIR model. Moreover, the classical approaches to regularization—penalizing the derivative of the transmission rate trajectory—do not correspond to realistic properties of pandemic spread. Consequently, models fit using derivative-based regularization are often biased toward underestimating the current transmission rate and future deaths. In this work, we propose mobility-driven regularization of the SEIR transmission rate trajectory. This method rectifies the artificial regularization problem, produces more accurate and unbiased forecasts of future deaths, and estimates a highly interpretable relationship between mobility and the transmission rate. Mobility data for this analysis was collected by Safegraph (San Francisco, CA) from major US cities between March and August 2020.

Keywords: SEIR model, reproduction number, transmission rate, regularization, cell phone mobility, COVID-19

Abbreviations: Susceptible Exposed Infected Recovered (SEIR).

INTRODUCTION

As the COVID-19 pandemic has demonstrated, it is difficult to assess to what degree and under what conditions states/countries may reopen their economies without dramatically increasing the rate of pandemic transmission. Answering this question requires a better understanding of the relationship between population mobility (economic activity) and the transmission rate. The transmission rate, r(t), is defined as the expected number of people a single infected individual transmits the virus to on a single day. Population mobility, m(t), is

defined as the number of unique visits to points of interest in a county, as measured by cell-phone traffic. [1]

The most natural way to estimate the r(t)-to-m(t) relationship should consist of two principal steps: 1) estimating r(t) from an SEIR model fit to death data [2] [3]; 2) comparing the estimated r(t) to the observed m(t). This appears to be the approach taken by several notable pandemic models [4] [5]. Unfortunately, the significant noise in the data often results in several model solutions that fit the same data well, but imply starkly different r(t)-to-m(t) relationships (Figure 1). Moreover, these solutions are highly dependent on the (arbitrary) choice of r(t) regularization, which is necessary to avoid model overfitting.

This makes it impossible to discover an accurate, stable mobility-transmission rate relationship by first estimating the transmission rate and then attempting to align it with mobility. Instead, it is necessary to determine whether there exists a transmission rate trajectory that both fits the data and aligns with the observed mobility trend. In this work, we propose a model that estimates r(t) as a function of m(t), which effectively regularizes r(t) using mobility data instead of artificial, derivative-based constraints. Consequently, the model produces a more accurate, unbiased, and stable solution, even in the presence of the significant variance observed in death data.

METHODS

Data

Aggregated and anonymized mobility data was obtained from Safegraph through an academic partnership program. We elected to fit the model to county-level death data, specifically for the 25 US counties with the highest death counts. Each county's model was fit to death data beginning on the date of the first death in the county and continuing through the end of May 2020; on average, this period was 91 days. While the model can easily be trained on case data (official counts of new infections), we believe death data is more reliable, as the number of reported cases depends strongly on the speed and availability of testing. Moreover, because the state of the pandemic and degree of social distancing varies widely within each state, we determined that state-level data is unsuitable for estimating the m(t)-to-r(t) relationship through an SEIR model, which assumes a uniform transmission rate. For example, in California, economic reopenings occurred soonest in rural regions of the state that were relatively unscathed

by the COVID-19 virus. As a result, these initial reopenings appeared to have little impact on the transmission rate, though cases in California eventually surged in June 2020, a few weeks after reopenings began in major cities. A state-level model would have interpreted rural reopenings as evidence of a weak m(t)-to-r(t) relationship and thus would have vastly underestimated the impact of metropolitan reopenings. In fact, we find that it is often impossible for a state-level model to elucidate any strong, stable m(t)-to-r(t) relationship.

SEIR model specification

Below, we provide the specifications for the SEIR model used in the subsequent sections. In addition to the typical SEIR variables, this model includes the following features: 1) asymptomatic infections [6] and infectious incubation periods [7] both of which are assumed to be 56% percent [8] as infectious as symptomatic infectious; 2) a hospitalization period, which is modeled as an exponential distribution with a mean of 8 days [9]; and 3) a critical care period, which is also modeled as an exponential distribution with a mean of 8 days [9]. Since the model described in the next section is fit to death data, the most important trajectory within this model is Exposed (E), Infected (I, I_U), Hospitalized (H), Critical Care (C), Death (D) (see Table 1 for all model notations). Note that the use of death data does not impact the SEIR model's fundamental assumption that infections drive infections. Deaths are simply used as a delayed signal for past infections. The model assumes deaths occur after infections according to an exponential distribution with a mean of approximately 12 days. Furthermore, incubation periods are assumed to follow an exponential distribution with a mean of 6 days. Thus, on average, mobility on day t is related to deaths on day t+18. The only difference between this approach and a more typical infection-based fitting approach is that model error is calculated with respect to forecasted and observed deaths, rather than forecasted and observed infections. We made the decision to use deaths based on the observation that COVID infections depended heavily on testing rates and capacities, which makes infections counts far less reliable. Moreover, there was no clear way to correct for this source of bias. Therefore, we chose to fit to death data because we believed that this was the most consistent and widely available metric for measuring the extent of the epidemic. The University of Washington's IHME has written in greater depth about the relative merit of death data. [10]

$$S'(t) = -\frac{r(t)S(t)}{N}(I(t) + \phi\big(I_U(t) + E(t)\big)$$

$$E'(t) = S(t+1) - S(t) - \alpha E(t)$$

$$I'(t) = \alpha(1-\pi_U)E(t) - \gamma_I I(t)$$

$$I_U{}'(t) = \alpha\pi_U E(t) - \gamma_U I_U(t)$$

$$H'(t) = \pi_H\gamma_I I(t) - \gamma_H H(t)$$

$$C'(t) = \pi_C\gamma_H H(t) - \gamma_C C(t)$$

$$R'(t) = (1-f)\gamma_C C(t)$$

$$D'(t) = f\gamma_c C(t)$$

Mobility-based transmission rate estimation

One would expect that the relationship between r(t) and m(t) could change over time as social distancing policies are implemented and relaxed. For instance, even if the same number of people visit the grocery store as did before the pandemic, they are less likely to become infected due to masks, shorter dwell times, lower store capacity, etc. Our model accounts for this by modeling an r(t)-to-m(t) relationship that is permitted to evolve over time. Let r_0 be the initial transmission rate and $\beta(t)$ be the r(t)-to-m(t) factor, determining how r(t) scales with mobility over time. Our model assumes that the percent change in r(t) is equal to the product of $\beta(t)$ and the percent change in m(t) :

$$\frac{r(t) - r(t-1)}{r(t-1)} = \beta(t)\frac{m(t) - m(t-1)}{m(t-1)}, t > 0$$

Changes in $\beta(t)$ are quite gradual, as social distancing behavior is unlikely to change from day to day (i.e. not the number of people coming in contact with each other, but *how* people behave when they do come in contact). Therefore, it is sufficient to estimate $\beta(t)$ on a weekly basis. This dramatically reduces the computation time to fit the model and serves as a natural regularizer.

Loss function, regularization, and robustness to outliers

In addition to fitting $\beta(t)$ on a weekly basis, we regularize it by penalizing its first derivative. This encourages the model to find as consistent a m(t)-to-r(t) relationship as possible. This is more realistic than the corresponding assumption used by most r(t) estimation models since, while we expect to observe dramatic changes in r(t), we do not expect to observe dramatic

changes in the relationship between m(t) and r(t). Together, these penalties produce the following minimization problem. Let *D(t)* be the number of observed deaths on day *t* and SEIR(*t*, $r(0), \beta$) be the number of model-forecasted deaths on day *t*, where $r(0)$ is the initial r(t). Then, we find the optimal model parameters by minimizing the following model error E:

$$E = \sum_t (SEIR(t, r(0), \beta) - D(t))^2 + \lambda \sum_t \big(\beta(t) - \beta(t-1)\big)^2$$

where λ is the regularization factor. Without any regularization, the model still produces plausible, stable m(t)-to-r(t) relationships for 68% of the counties we processed. However, there are a few counties for which the model terminates at a local minimum corresponding to a highly unstable m(t)-to-r(t) relationship or even diverges. Our numerical experiments found that any $\lambda \geq 0.4$ consistently prevents the model from diverging. Therefore, to avoid inserting additional bias into the model through stronger regularization, we elected to set λ to its minimum effective value of 0.4. The addition of λ does not provide a theoretical guarantee that the resulting solution is a global minimum. However, it does insert significant convexity into the objective function, which was sufficient to ensure the optimizer avoided local minima for the 2,250 models fit during our experiments. We also achieved faster and more stable results by setting a prior on the initial reproduction number r_0. Most estimates [11] have placed the standard reproduction number between 2.5 and 3, so we set a normal prior with a mean of 2.8 and a standard deviation of 1. Enforcing this prior is equivalent to adding a third term to the objective function:

$$E = \sum_t (SEIR(t, r(0), \beta) - D(t))^2 + \lambda \sum_t \big(\beta(t) - \beta(t-1)\big)^2 + (\gamma_T r(0) - 2.8)^2$$

where γ_T is the average length of the infectious period (which is necessary to convert between the transmission rate and the reproduction number). In the following experiments, the models are fit by minimizing this error function with Scipy's L-BFGS numerical minimization algorithm. When forecasting future deaths, our models calculated future transmission rates using the most recent estimate for $\beta(t)$ and the most recent 7-day moving average of m(t).

RESULTS

When fit to US COVID-19 death data from March through May and used to forecast deaths over the next 5 weeks, the mobility model overestimated deaths by 5.0% among the 25 counties

with the highest death totals. This compared favorably with our highest-accuracy non-mobility model (where r(t) was permitted to vary as needed to fit the death data with regularization of the first derivative), which underestimated deaths by 32.5% over the same period. Figure 2 compares the non-mobility model and our mobility model relative errors for each county, where relative error is defined as the difference between forecasted and observed deaths, divided by the total number of deaths in the county. As one can see, while the non-mobility model systematically underestimated future deaths, the mobility model appeared to be nearly unbiased; its average error was close to zero.

Figure 3 compares forecasted trajectories from the mobility and non-mobility models for six US counties with large numbers of cases and deaths.

The fitted results of our mobility model also appear to be quite stable, meaning that the model only infrequently revised its estimate of the relationship between m(t) and r(t). Among the 25 counties with the highest death totals, the mean change in $\beta(t)$ for models fit daily between June 1 and June 22 was 1.83%. Figure 4 demonstrates the stability of the LA county fitted results over that period, including forecasts for the following month.

During and immediately following lockdowns in the 25 counties with the most deaths, estimates of the mobility-transmission factor clustered closely around 0.89, meaning that an 1% decrease in mobility was associated with a 0.89% decrease in the transmission rate (Figure 5). As expected, social distancing procedures appeared to decrease the sensitivity of the transmission rate to mobility to a median value of 0.67 in week 3 post-shutdown. However, the mobility-transmission factor appeared to increase in subsequent weeks as social distancing was relaxed; it was estimated as 0.72 in week 5, 0.76 in week 7, and 0.78 in week 9 post-shutdown. The dispersion in mobility-transmission factors also increased during this period, suggesting that some counties were more effective than others in reopening portions of their economies without dramatically increasing r(t). Several counties, primarily in Massachusetts, appeared to decrease r(t) even as m(t) increased. This could be due to improved contact tracing, isolation, and adherence to social distancing and mask recommendations.

DISCUSSION

We have proposed a model that uses mobility data to forecast future deaths and estimate the relationship between mobility and the transmission rate. The model finds a stable association

between m(t) and r(t) that is conserved across several major counties, and it outperforms non-mobility models when forecasting future deaths. Moreover, the fitted trends in $\beta(t)$ suggest that public health interventions (e.g. contact tracing, social distancing, mask wearing) initially reduced the sensitivity of the transmission rate to changes in mobility in some counties, but $\beta(t)$ rebounded as social distancing was relaxed. The median, most recent estimate for the mobility transmission factor of $\beta(t) = 0.78$ suggests that transmission rates remain sensitive to increases in mobility. While the model's forecasts are limited to the approximately 25 counties with sufficient daily death data, it provides an improved framework for understanding the relationship between mobility and the transmission rate, as well as estimates of this relationship for many of the US counties with the largest outbreaks.

BIBLIOGRAPHY

[1] Safegraph, "Safegraph COVID-19 Data Consortium," Safegraph, 1 March 2020. [Online]. Available: https://www.safegraph.com/covid-19-data-consortium. [Accessed 1 March 2020].

[2] Department of Statistics, Carnegie Mellon University, "Advanced Methods for Data Analysis: Smoothing Splines," CMU Statistics, Pittsburgh, 2020.

[3] J. H. Jones, "Models of Infectious Disease," in *Stanford Spring Workshop in Formal Demography*, Stanford, 2008.

[4] UW Institute for Health Metrics and Evaluation, "COVID-19 Projections Release Notes," University of Washington, Seattle, 2020.

[5] D. Rubin, "Small-Area Projections of COVID-19 Transmission in the United States," Children's Hospital of Philadelphia, Philadelphia, 2020.

[6] D. Böhning, I. Rocchetti, A. Maruotti and H. Holling, "Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods," *International Journal of Infectious Diseases*, vol. 97, pp. 197-201, 2020.

[7] Centers for Disease Control and Prevention, "Clinical Questions Regarding COVID-19," 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html.

[8] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang and J. Shaman, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," *Science*, vol. 368, no. 6490, pp. eabb3221-493, 2020.

[9] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe and W. Green, "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand," Imperial College COVID-19 Response Team, 2020.

[10] IHME COVID-19 health service utilization forecasting team, Christopher JL Murray, "Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries.," 2020.

[11] Z. Zhuang, S. Zhao, Q. Lin, P. Cao, Y. Lou, L. Yang, S. Yang, D. He and L. Xiao,

"Preliminary estimates of the reproduction number of the coronavirus disease (COVID-19) outbreak in Republic of Korea and Italy by 5 March 2020," *International Journal of Infectious Diseases*, vol. 95, pp. 308-310, 2020.

[12] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich and J. Lessler, "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application," *Annals of internal medicine*, vol. 172, no. 9, pp. 577-582, 2020.

[13] COVID Tracking Project, "COVID Tracking Project Daily State Data," Atlantic Monthly Group, New York.

[14] C. T. Project, "COVID Tracking Project Historical State Data," The Atlantic Monthly Group, 2020. [Online]. Available: https://covidtracking.com.

[15] P. Quah, A. Li and J. Phua, "Mortality rates of patients with COVID-19 in the intensive care unit: a systematic review of the emerging literature," *Critical Care*, vol. 24, no. 1, p. 285, 2020.

Table 1: Explaining model notations

Variable	Description
$S(t)$	Susceptible individuals in the population. Initialized to 90% of population.
$E(t)$	Individuals exposed to the virus. Initialized to 0.
$I(t)$	Individuals with ongoing, detected infections. Initialized to 10.
$I_U(t)$	Individuals with ongoing, undetected infections. Initialized to 0.
$H(t)$	Patients currently hospitalized. Initialized to 0.
$C(t)$	Patients currently in critical care. Initialized to 0.
$R(t)$	Recovered individuals. Initialized to 0.
$D(t)$	Virus-related deaths. Initialized to 0.
$r(t)$	Transmission rate. Initialization set by model.
N	Population size.
ϕ	Relative infectivity of undetected and pre-symptomatic infections. Set to 0.54.
α	Rate of exposed period. Set to 0.2.
π_U	Proportion of cases undetected. Estimated to be 0.88 [8].
γ_U	Rate of undetected infection period. Set to 0.12 [12].
γ_I	Rate of detected infection period. Set to 0.11 [12].
π_H	Proportion of detected infections hospitalized. Set to 0.3 [13].
γ_H	Rate of hospitalization period. Set to 0.125 [9].
π_C	Proportion of hospitalized patients admitted to critical care. Set to 0.3 [14].
γ_C	Rate of critical care period. Set to 0.125 [9].
f	Death rate among critical care patients. Set to 0.55 [15].

Figure 1: Alternate transmission rate trajectories in Montgomery County, Maryland: Derivative-Smoothed Trajectory (A) and Mobility-Based Trajectory (B). Transmission rate is plotted on the y-axis, and the number of days since the first case in the county is plotted on the x-axis. Both trajectories closely fit the observed death data, but show very different trends. The first consists of a slow, sustained decrease in the transmission rate, while the second consists of a sharp initial decrease (corresponding to the lockdown) followed by a slow rise (corresponding to reopening). A model that penalizes either the first or second r(t) derivative would prefer the first result to the second. The mobility-based model, described in subsequent sections, produces the second trajectory. While the second solution reflects the likely scenario that r(t) increased in the summer of 2020 as the US economy reopened, the first r(t) suggests that the economic reopening did not influence the transmission rate.

Figure 2: Relative error comparison for mobility and non-mobility models.

Figure 3: Forecasted deaths for six counties using the mobility model (solid line) and non-mobility model (dashed line). Models were trained on March through May 2020 COVID-19 data (fully shaded curve) and tested on the following 5 weeks (lightly shaded curve).

Figure 4: Forecasted death trajectories for models fit daily between June 1 and June 22 for Los Angeles County.

Figure 5: Mobility-transmission factor estimates by week after initial shutdowns.