Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones

Spencer Woody, Mauricio Tec,
Maytal Dahan, Kelly Gaither, Spencer J. Fox,
Lauren Ancel Meyers, and James Scott*

The University of Texas at Austin COVID-19 Modeling Consortium†

April 16, 2020

1 Introduction

On March 26, 2020, the Institute for Health Metrics and Evaluation (IHME) at the University of Washington released a website that forecasts coronavirus disease (COVID-19) healthcare demand and mortality for all states in the United States. After being cited during a White House briefing on COVID-19 modeling efforts, their forecasting model, described in a preprint on medRxiv [IHME et al., 2020], has received an enormous amount of attention from both the general population and scientific community. IHME has since updated the model several times resulting in considerable revisions to the COVID-19 forecasts.

The IHME approach departs from classic epidemiological modeling. Rather than using systems of equations to project the person-to-person transmission of the virus, the model postulates that COVID-19 deaths will rise exponentially and then decline in a pattern that roughly resembles a bell curve (i.e., normal distribution). The model assumes that the shape of the curve will be curtailed by social distancing measures. Key inputs driving this component of the IHME model include the reported dates of state-wide shelter-in-place orders and shapes of

*Corresponding author: james.scott@mccombs.utexas.edu

†This work was partially funded by a gift from Tito's Handmade Vodka to the UT COVID-19 Modeling Consortium.

COVID-19 epidemiological curves observed in Chinese and European cities following the implementation of similar measures.

In light of the popular appeal of the IHME model and considerable scrutiny from the scientific community, we have developed an alternative curve-fitting method for forecasting COVID-19 mortality throughout the US. Our model is similar in spirit to the IHME model, but different in two important details.

1. For each US state, we use local data from mobile-phone GPS traces made available by SafeGraph[1] to quantify the changing impact of social-distancing measures on "flattening the curve."

2. We reformulated the approach in a generalized linear model framework to correct a statistical flaw that leads to the underestimation of uncertainty in the IHME forecasts.

The incorporation of real-time geolocation data and several key modifications yields projections that differ noticeably from the IHME model, especially regarding uncertainty when projecting COVID-19 deaths several weeks into the future.

2 Model overview

At a high level, our model shares some key properties of the IHME model.

Similarity 1: a statistical curve-fitting approach. Ours is not an epidemiological model, in the sense that we do not try to model disease transmission, nor do we use or attempt to estimate underlying epidemiological parameters like the basic reproductive rate or attack rate. Rather, our model is purely statistical: we are fitting a curve and a probabilistic error model to observed death rates in a state, and we are extrapolating from that curve. The advantage of this approach is that it does not require estimates of critical epidemiological parameters, some of which remain elusive. The disadvantage is that it cannot project longer-term epidemiological dynamics beyond the initial wave of mitigated transmission. For this reason, we do not use the model to make projections beyond a moderate (2-3 week) horizon.

[1] SafeGraph is a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places. To enhance privacy, SafeGraph excludes census block group information if fewer than five devices visited an establishment in a month from a given census block group.

Similarity 2: time-evolving Gaussian curves. The family of curves we use for expected deaths over time is identical to that of the IHME model. Specifically, we assume that expected daily death rates can be *locally* approximated by a three-parameter curve that is proportional to a Gaussian kernel. This approximation is local in the sense that the curve's three parameters are allowed to evolve in time as a function of state-level covariates. Just as in the IHME model, this results in fitted death-rate curves that, when plotted over time, can differ substantially from the shape of a single Gaussian. While epidemic curves do not resemble Gaussian curves, *time-evolving* Gaussian curves do provide a good fit to observed COVID-19 state-level death rates.

Similarity 3: regression on social-distancing covariates to inform projections. As in the IHME analysis, our regression model connects each state's death-rate curve to covariates that describe social distancing within each state. Changes in each state's social-distancing covariates can "flatten the curve" by changing the peak death rate, the timing of that peak, and the deceleration in death rate near the peak. The strength of this approach is that it can leverage readily available data on social distancing without requiring a mechanistic transmission model.

However, our model differs from the IHME model in at least three key ways, which collectively result in different forecasting behavior.

Difference 1: real-time daily social-distancing data. We use data on Americans' actual social-distancing behavior, derived from GPS traces from tens of millions of mobile phones across the country. This data source quantifies two main types of distancing behavior: 1) changes in visitation patterns to public places like restaurants, bars, schools, parks, pharmacies, grocery stores, etc.; and 2) time spent at home versus at work. The IHME model, by contrast, uses a much coarser measure of social distancing: the timing of state-level policy implementations like school closures and stay-at-home orders. But the mobile-phone data reveals substantial differences among states in the timing and extent of peoples' actual distancing behavior, even for states with nominally similar policies. In Texas, for example, many large cities issued their own stay-at-home orders before the state did, affecting the movement patterns of many millions of people days before a statewide policy was in place—a fact that is clearly visible in the data. Our measures capture this substantial state-level and temporal variation that is obscured by regressing only on policies.

Difference 2: U.S. data only. The IHME model assumes that data on death rates and social distancing policies in other countries (specifically China, at least

in the original formulation of the model) can inform U.S. state-level forecasts. For a variety of reasons, we find this assumption problematic. Our forecasts therefore rely solely on U.S. data, with state-level parameters shrunk toward a common mean in a hierarchical Bayesian model.

Difference 3: valid uncertainty quantification. We address a problem with the IHME model by relying on a fundamentally different statistical assumption about model errors. Briefly: the IHME model fits *cumulative* death rates using a least-squares-like procedure on the log scale and applying standard large-sample statistical theory to get confidence intervals. For this procedure to result in valid uncertainty quantification, one must assume that successive model errors are independent. But in the IHME fitting procedure, this assumption is violated: today's cumulative death rate is yesterday's plus an increment, so the two must be correlated. Our model repairs this problem by fitting *daily* (noncumulative) death rates using a mixed-effects negative-binomial generalized linear model.

3 Model details: a negative-binomial mixed-effects GLM for daily COVID-19 deaths

3.1 Model structure

We let i index the geographic area; in our analysis this is U.S. states, but it could be at any level of spatial resolution (e.g. country, city, etc). To make our results comparable to the IHME model, let t denote the number of days elapsed since a threshold death rate of 3 per 10 million residents was reached in a given area. Thus t doesn't represent calendar days, but rather a notion of "epidemic days."

Let y_{it} denote observed number of deaths in area i at time t. Let $\tilde{y}_{it}$ denote per-capita death rate. The IHME model assumes that the expected daily death rate λ_{it} can be locally approximated by a curve proportional to a Gaussian kernel:

$$E(y_{it}) = \lambda_i(t) = \kappa_{it} \cdot \exp\left\{-\frac{\omega_{it}}{2} \cdot (t - \mu_{it})^2\right\} \tag{1}$$

This leads to the following interpretation of the parameters:

- κ is the maximum daily expected death rate
- μ is the day on which the expected death rate achieves its maximum
- ω is a steepness parameter: higher ω means the death rate rises more rapidly as t approaches μ, and also falls more rapidly on the far side of

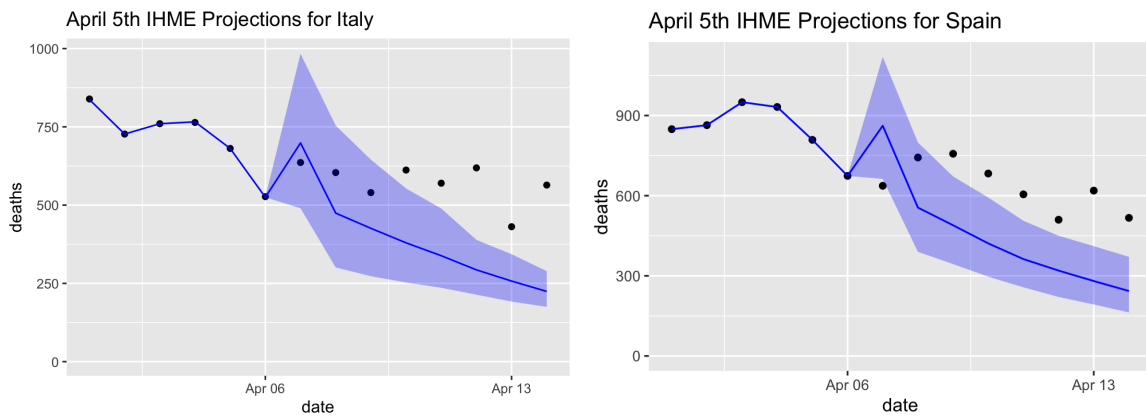

Figure 1: The IHME model's April 5 projections ten days ahead for Italy and Spain (blue lines). The actual data (black dots) fall noticeably outside the model's 95% error bars (shaded blue region), illustrating the model's underestimate of forecasting uncertainty, even in countries whose peak daily death rates seem to be well characterized by the data.

μ. Specifically, the slope at the inflection point of the death-rate curve is $\kappa\sqrt{\omega}$.

Equation (1) expresses the most general form of the IHME model, where $\theta_{it} = (\kappa_{it}, \mu_{it}, \omega_{it})$ changes from area to area and day to day. This is highly overparametrized; it is necessary to enforce some type of shrinkage on these parameters in order to make the model identifiable. We address model overparametrization in a similar way to the original IHME analysis (i.e. via a hierarchical model). But we use very different data based on mobile-phone GPS traces that quantify actual distancing behavior, rather than the timing of state-level social-distancing policies (e.g. school closures, stay-at-home orders, etc.).

The IHME model-fitting process. We note that the IHME model parameterizes its Gaussian curves in a slightly different way, but the underlying family is identical to Equation (1), in the sense that there is a bijection between our parameterization and theirs. Briefly, the IHME model assumes that the *cumulative* death rate is proportional to Gaussian CDF, and they fit the three model parameters by optimizing a penalized least-squares objective on the log-cumulative-deaths scale, interpreting the result as a Bayesian maximum a posteriori (MAP) estimate under an assumed prior. More specifically, let N_i be the population in state i and define $r_{it} = N_i^{-1} \sum_{s<t} y_{it}$ be the per-capita cumulative death rate in

state i. The IHME model assumes that

$$\log r_{it} = \log \left\{ \int_0^t \lambda_i(t)\, ds \right\} + \epsilon_i \,. \tag{2}$$

The model is fit by a penalized nonlinear least-squares procedure, encouraging the parameters of each state's $\lambda_i(t)$ curve to be shrunk towards a common *conditional* mean, given social-distancing covariates. They interpret the result as a hierarchical Bayesian MAP estimate under an assumed prior, and use the inverse Hessian matrix at the MAP estimate as plug-in estimate of the posterior covariance matrix for the model parameters.

In the Appendix, we describe why this procedure does not result in valid confidence intervals. Briefly, it ignores the problems of heteroscedasticity and intra-day correlation in model errors associated with fitting to cumulative data. Together these have major consequences for uncertainty quantification. Figure 1 briefly illustrates the problems that can arise: it shows the IHME projections for Italy and Spain on April 5 looking ten days ahead, together with 95% prediction intervals. The actual data fall noticeably outside the model's claimed range of uncertainty. This is especially worrisome, given that peak daily death rates in Italy and Spain seem to be well-characterized by the data itself, versus the situation in many U.S. states that have yet to reach their peak, where we must rely on a model to extrapolate the location of the peak.

An April 8 technical report by Marchant et. al points to similar problems with the IHME's U.S. projections. They found that, in evaluating the model's U.S. projections made on April 1, only 27% of the actual data points on the subsequent day actually fell within the 95% confidence bands.

Luckily, this is an easily correctable problem, by placing the model in Equation (1) into the framework of generalized linear modeling. This allows for better uncertainty quantification, as we describe in the next section.

3.2 Our fitting approach

Observe that if we move to the log scale, Equation (1) becomes:

$$\log \lambda_{it} = \log \kappa_i - 0.5\omega_i(t - \mu_i)^2 \equiv \beta_{i0} + \beta_{i1} t + \beta_{i2} t^2$$

for some β vector. Keeping in mind that λ_{it} is the expected value for a count outcome (daily deaths y_{it}), we recognize this as the expression for the linear predictor in a generalized linear model (GLM) with a log link function, as might arise in a Poisson or negative-binomial regression model for y. On the right-hand side, we have a locally quadratic regression on t, the number of elapsed days since deaths crossed the threshold value of 3 per 10 million. Moreover, there

is a simple relationship between the regression coefficients β and the original parameters of the curve:

$$\begin{bmatrix} \omega \\ \mu \\ \kappa \end{bmatrix} = \begin{bmatrix} -2\beta_2 \\ \frac{\beta_1}{2\beta_2} \\ \exp\left\{\beta_0 - \frac{\beta_1^2}{2\beta_2}\right\} \end{bmatrix}$$

Thus to fit the model, we estimate a hierarchical negative binomial regression model with mean λ and overdispersion parameter r, as follows:

$$y_{it} \sim \text{NegBin}(\lambda_{it}, r) \tag{3}$$

$$\log \lambda_{it} = \log N_i + \beta_{it,0} + \beta_{it,1} t + \beta_{it,2} t^2 \tag{4}$$

$$(\beta_{it,0}, \beta_{it,1}, \beta_{it,2})^T = \mu + \Gamma x_{it} + \eta_i \tag{5}$$

$$\eta_i \sim N(0, \Sigma)\,, \tag{6}$$

where $\text{E}(y_{it}) = \lambda_{it}$ and $\text{Var}(y_{it}) = \lambda_{it}(1 + \lambda_{it}/r)$. Here we have included an offset for the logarithm of N_i, the population in area i, so that the linear predictor can be interpreted as the log per-capita death rate. Here x_{it} is a vector of social-distancing predictors that are allowed to "flatten the curve" by changing its shape, via the second-stage regression. The negative binomial model naturally handles the heteroscedasticity and overdispersion that we observe in the data.

One note: we have phrased this model as a raw polynomial. But we actually fit the model using orthogonal polynomials to stabilize estimation; one may convert between the raw and orthogonal parameterizations easily.

We fit the model using Markov Chain Monte Carlo, sampling from the posterior distribution of all model parameters. We use weakly informative priors on the fixed effects β, the second-stage regression coefficients Γ, and the covariance matrix Σ of the random effects. We also explored the possibility of using more informative priors based on daily death rates from European countries. But this had almost no effect on the fit of the model. Model fitting was carried out using the `rstanarm` package in R [Goodrich et al., 2020]. Forward-looking model projections are based on draws from the posterior predictive distribution. Because of the way that our covariates are constructed (see below), most of the x values for these forward-looking projections correspond to social-distancing behavior that has already been observed. However, at a longer horizon some extrapolation of x is necessary to generate projections; here we assume that social-distancing behavior in a state remains unchanged from the average behavior over the seven most recent days of available data.

3.3 Social-distancing predictors

To define the social-distancing predictors x_{it}, we take a weighted averages of past social-distancing metrics made available to us by SafeGraph. The metrics considered as predictors include daily time series per location for:

- the median duration of time that people in a given area spend at home, as well as the number of people in an area exhibiting "full-time work" behavior, at their normal place of work;[2]
- and total visitation counts for various points of interest aggregated by category, including grocery stores, hospitals, parks, restaurants, bars, colleges, etc.[3]

These data are derived from GPS traces of tens of millions of mobile phones nationwide. "Home" and "work" locations for a device were inferred by SafeGraph based on daytime and overnight locations over an extended period of time. The data were provided to us in an aggregated format; no device-level data or individual GPS traces were accessible by the research team. Figure 2 shows a selection of these social-distancing measures over time in both New York and Texas.

We denote these D distancing metrics $s_{it,1}, \ldots, s_{it,D}$, observed each day in each state. To construct useful covariates out of this information, we proceed as follows. For each distancing metric s_j, define a corresponding lagged version $\tilde{s}_j$ as follows:

$$\tilde{s}_{it,j} = \sum_{l=1}^{L} w_l s_{i,t-l,j}$$

where $w = (w_1, \ldots, w_L)$ is a fixed vector of backward looking weights. The lagging is necessary to account for the time from infection to death. Social distancing actions taken on day t will not result in an immediate change in death rates; rather, the effect will show up many days in the future. We therefore tune this weight vector based on prior knowledge of the distribution of lags between infection with COVID-19 and death. Specifically, w is a Gaussian kernel centered 23.5 days in the past, with a standard deviation of 6 days. This is based on published estimates of time from contraction to illness onset [Lauer et al., 2020] and on time from illness onset to death [Zhou et al., 2020].

To construct predictors x for our regression model, we then take the first four principal components of the lagged social distancing metrics. These four components explain about 95% of the variation in the state-level data.

[2] https://docs.safegraph.com/docs/social-distancing-metrics

[3] https://docs.safegraph.com/docs/places-schema

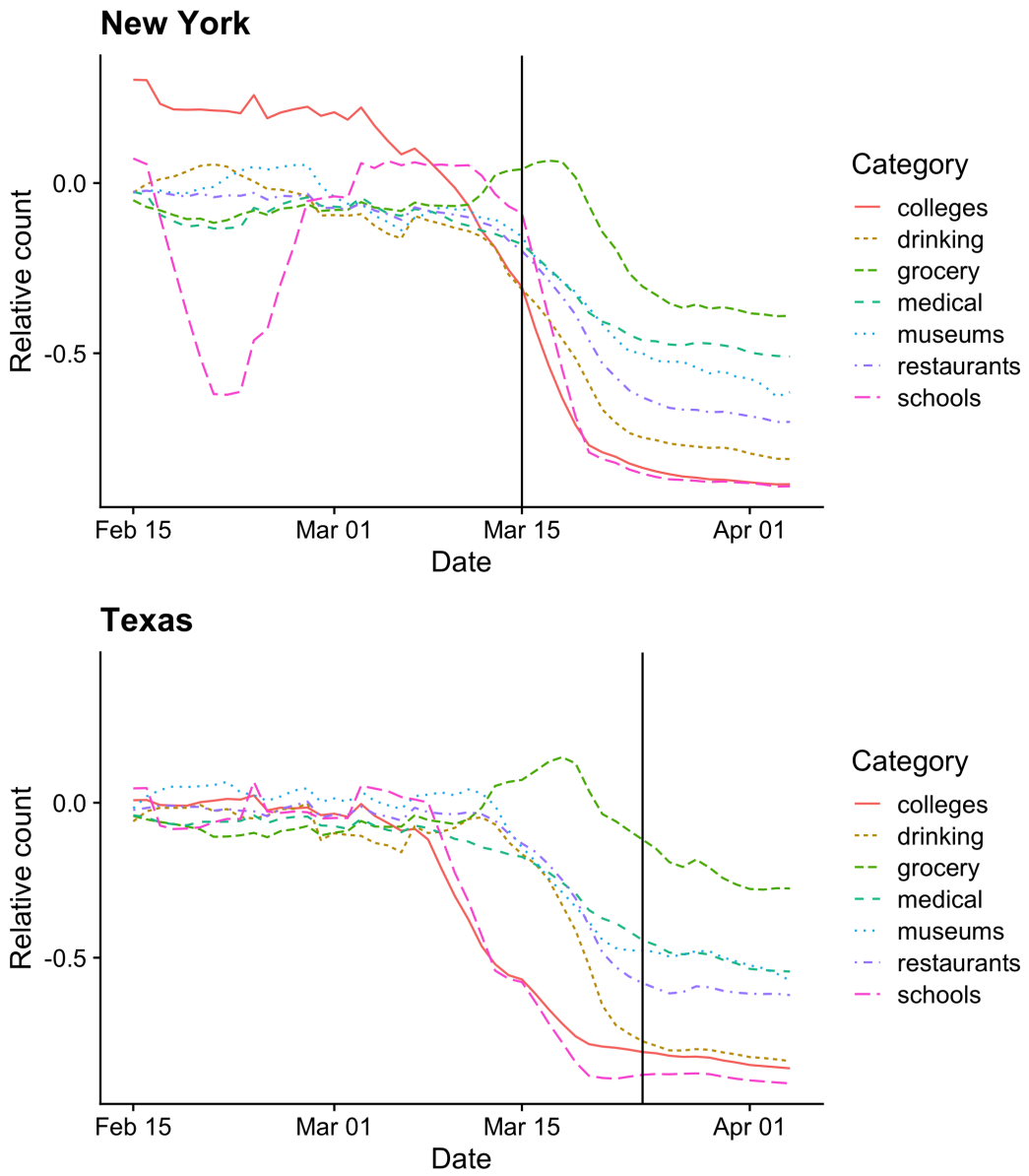

Figure 2: SafeGraph social-distancing data in New York versus Texas. The social-distancing predictors we use in our model quantify visitation patterns to various points of interest, using GPS traces derived from mobile phones. The vertical axis is standardized so that 0 represents a pre-pandemic baseline, and -0.5 indicates a 50% decrease in visitations relative to that baseline. The vertical lines in each panel represent the day in which the state death rate reached 3 per 10 million residents. One can see that social distancing in Texas began substantially before this threshold day was reached, but much nearer this day in New York.

4 Results

We update our model results daily and post them at https://covid-19.tacc.utexas.edu/projections/.

A Uncertainty quantification in the original IHME model

In the IHME model, the choice of using penalized-least-squares fit on the log-cumulative-deaths scale has major consequences for statistical inference. In particular, the authors use the inverse-Hessian matrix at the MAP estimate in order to produce uncertainty estimates. This uncertainty quantification procedure, however, implicitly assumes that successive observations are independent. Indeed, without this assumption, it is not generally true that the inverse-Hessian at the MAP provides a valid large-sample estimate for the covariance matrix of an estimator, Bayesian or otherwise. This important technical condition simply cannot be true on the scale used for fitting the IHME model, for the simple reason that the data used for fitting are cumulative: if today's prediction for cumulative death rate is too high, then tomorrow's prediction is more likely to be too high as well.

This is easily verified by a simple calculation. The covariance of two successive cumulative death rates r_{it} is:

$$\begin{aligned}\mathrm{cov}(r_{i,t}, r_{i,t+1}) &= \mathrm{cov}(r_{i,t}, r_{i,t} + y_{i,t+1}) \\ &= \mathrm{var}(r_{it}) + \mathrm{cov}(r_{i,t}, y_{i,t+1}) \\ &= N^{-2}\mathrm{var}\left(\sum_{s<t} y_{it}\right) + \mathrm{cov}(r_{i,t}, y_{i,t+1})\end{aligned}$$

This covariance cannot be zero unless $\mathrm{cov}(r_{i,t}, y_{i,t+1})$ is negative, and of sufficient magnitude to exactly cancel out the first (positive) term—which is highly implausible. Moving to the logarithmic scale does not repair the basic fallacy of assuming independent errors. This likely accounts for much of the understatement in uncertainty seen in Figure 1.

References

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2020. URL https://mc-stan.org/rstanarm. R package version 2.19.3.

IHME, COVID-19 health service utilization forecasting team, and Christopher JL Murray. Forecasting COVID-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months, 2020. URL https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1. doi: https://doi.org/10.1101/2020.03.27.20043752.

Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*, 03 2020. ISSN 0003-4819. doi: 10.7326/M20-0504. URL https://doi.org/10.7326/M20-0504.

Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 2020. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30566-3.