medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates

Wen-Hao Chiang, Xueying Liu, George Mohler*

Department of Computer & Information Science
Indiana University–Purdue University Indianapolis
Indianapolis, IN 46224
{chiangwe, xl17, gmohler}@iupui.edu

June 6, 2020

Abstract

Hawkes processes are used in machine learning for event clustering and causal inference, while they also can be viewed as stochastic versions of popular compartmental models used in epidemiology. Here we show how to develop accurate models of COVID-19 transmission using Hawkes processes with spatial-temporal covariates. We model the conditional intensity of new COVID-19 cases and deaths in the U.S. at the county level, estimating the dynamic reproduction number of the virus within an EM algorithm through a regression on Google mobility indices and demographic covariates in the maximization step. We validate the approach on short-term forecasting tasks, showing that the Hawkes process outperforms several benchmark models currently used to track the pandemic, including an ensemble approach and a SEIR-variant. We also investigate which covariates and mobility indices are most important for building forecasts of COVID-19 in the U.S.

1 Introduction

Mathematical modeling and forecasting is playing a pivotal role in the ongoing SARS-CoV-2 (COVID-19) pandemic. In mid-March 2020, a report out of Imperial College London [8] forecasted severe consequences in the US and UK without significant public health interventions. In both nations, governments responded by closing schools, non-essential businesses and releasing general stay-at-home (shelter-in-place) orders. In the U.S., state and local policymakers are using mathematical models and projections to inform decisions about when and how to relax public health measures that have been put in place. By and large, compartmental models that explicitly incorporate transmission characteristics of infectious diseases have been favored over machine learning approaches. High profile Susceptible-Exposed-Infected-Removed (SEIR) models include those out of Columbia University [22], MIT [13], and UCLA [32] (in the case of the UCLA model, a SEIR-variant with an unreported compartment is fit using least-squares to reported infection and recovery data). A major exception is the well-known IHME model [3], that employs Gaussian curve fitting to COVID-19 case and death count time series in locations further along (e.g. China, Europe), to estimate curves in locations where the outbreak is more recent (e.g. the United States). The IHME model has been called into question by epidemiologists because it lacks explicit transmission dynamics in the model [11].

Our goal in this paper is to show that Hawkes processes, widely used in the machine learning community to model contagion patterns in event data, are well suited for modeling and forecasting COVID-19 case and mortality data. They have several advantages that we will highlight, including being highly flexible in accommodating auxiliary spatio-temporal features such as county-level demographics and temporal mobility patterns, yet mathematically they are connected to compartmental models [24] and allow for explicit incorporation of transmission dynamics (which we briefly review in the following section). Furthermore, extensive research has been conducted in the past several years to couple machine learning techniques with the point process framework. Non-parametric Hawkes processes can be constructed where the triggering kernel is learned [31] and more recently fully neural network based point processes have been developed

*Corresponding author

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

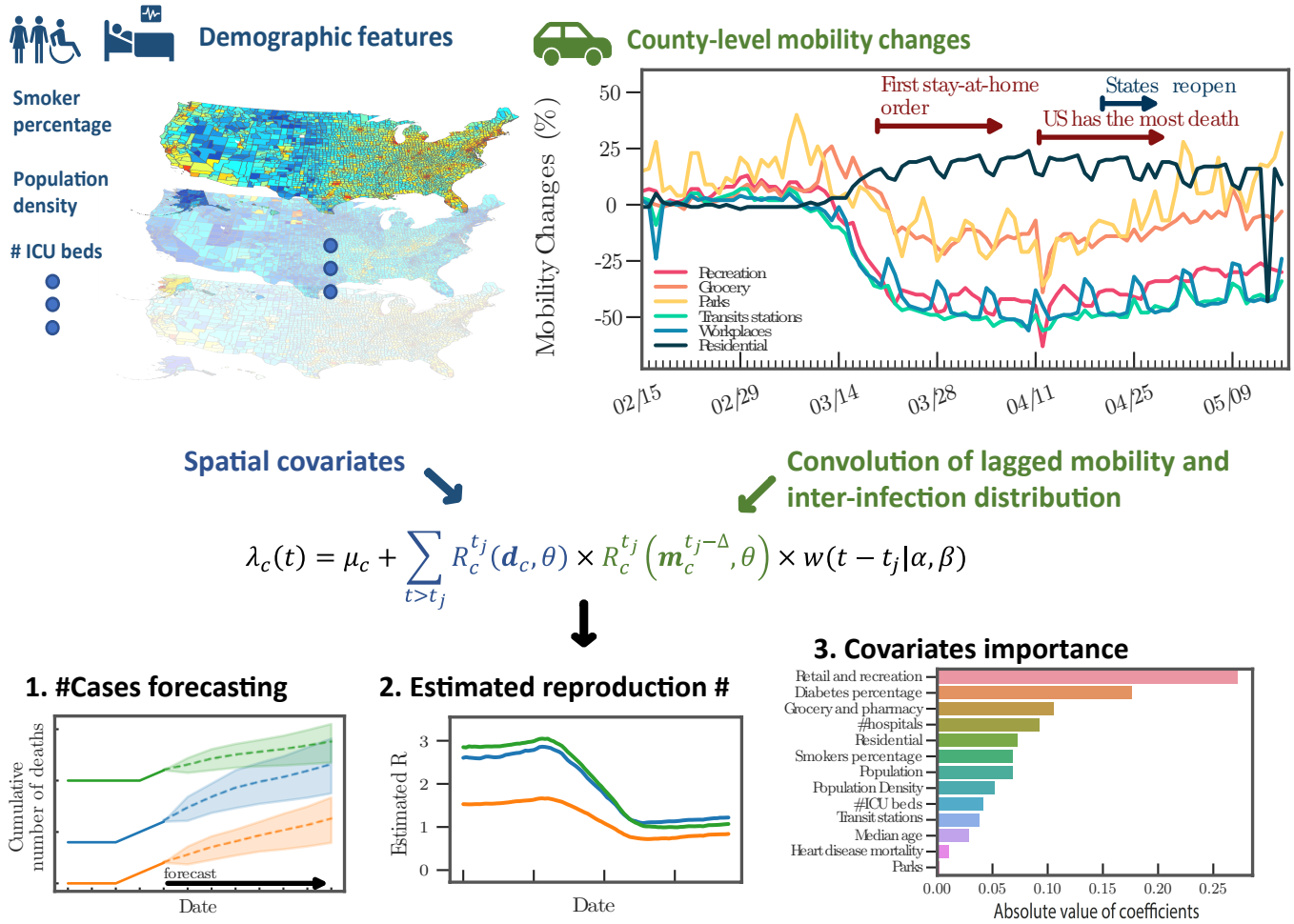


Figure 1: Framework of Hawkes process model for COVID-19 transmission. Demographic features at the county level impact the reproduction number of the Hawkes process. Lagged changes in mobility impact future secondary infections through a convolution with the inter-infection distribution $w(t)$. Output of the model includes 1) forecasts of future cases and mortality through simulation of the Hawkes process intensity, 2) an estimate of the dynamic reproduction number of the virus, and 3) regression results that allow for interpretation of the covariates that influence transmission differences across counties.

[16, 21]. Sparse linear combinations of Hawkes processes were a winning solution in the 2017 NIJ Crime Forecasting Challenge [19]. In certain cases a mixture of Hawkes processes may be needed to model more complex event contagion with high dimensional marks through dirichlet processes [30, 7]. Hawkes processes can also be used for causal inference on networks [29] and recent efforts have also focused on training point processes through reinforcement learning [27, 14]. We believe all of these methods have potential applications to modeling infectious diseases that are highly complex due to heterogeneity in the population, environment, and disparate public policies across regional and local jurisdictions. Despite these advantages, to our knowledge the only U.S. state where a Hawkes process is being used to inform COVID-19 policy is in New Jersey (a collaboration with Facebook AI Research) [1].

The outline of the paper is as follows. In Section 2 we introduce our Hawkes process model whose productivity (reproduction number) is dynamic and depends on spatio-temporal covariates. Unlike recently introduced models that incorporate covariates into the background rate of a Hawkes process [18, 23], our Hawkes process model may be viewed as a convolution of lagged mobility with an inter-infection time distribution to estimate the intensity of secondary infections in the future. This is important as phased reopening in the U.S. leads to mobility changes, the effects of which are not realized in case and mortality data until days or weeks later. Hence the model can be used to forecast changes in transmission and new cases in real-time as mobility changes (see Figure 1). We estimate the intensity along with the dynamic reproduction number of the virus within an EM algorithm through a regression on Google mobility indices and

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

demographic covariates in the maximization step. In Section 3, we validate the approach on short-term forecasting tasks, showing that the Hawkes process outperforms several benchmark models including the Columbia University SEIR model [22] and an ensemble model from Berkeley that uses combined linear and exponential predictors with spatial covariates [2]. We also investigate which covariates and mobility indices are most important for building forecasts of COVID-19 in the U.S. In Section 4 we discuss directions for future research and how the machine learning community may be able to help improve Hawkes process models of COVID-19 as the pandemic continues to unfold.

2 Hawkes process model of COVID-19 transmission

In this section we introduce a Hawkes process with spatio-temporal covariates for modeling COVID-19 case and death data. We then discuss the connection of the model to compartment models used in epidemiology and develop an expectation-maximization algorithm for inference.

2.1 Incorporating covariates into the Hawkes process

We propose a novel Hawkes process model that simultaneously estimates the intensity of events and tracks the dynamic reproduction number of the virus. Given the timestamps (or dates), $\mathcal{T} = \{t_1, t_2, \cdots t_n\}$, of daily reported positive test cases or deaths, we model the rate of new cases (or deaths) in each country c as,

$$\lambda_c(t) = \mu_c + \sum_{t > t_j,\, t_j \in \mathcal{T}} R_c^{t_j}(\boldsymbol{d}_c, \theta_d) \times R_c^{t_j}(\boldsymbol{m}_c^{t_j - \Delta}, \theta_m) w(t - t_j), \tag{1}$$

where μ_c is the background rate modeling imported infections, $w(t)$ is the inter-infection time distribution, $\boldsymbol{m}_c^t = [m_1^t, m_2^t, \cdots]^\intercal$ are mobility indices on day t, and $\boldsymbol{d}_c = [d_1, d_2, \cdots]^\intercal$ are static demographic features. The time-varying reproduction number R_c^t can be interpreted as the average number of secondary infections caused by a primary infection. Because we are modeling reported infections rather than time of exposure, we introduce the parameter Δ to capture a potential lag between a mobility change and the time t_j of a reported primary infection.

We model the dynamic reproduction number R_c^t through a Poisson regression,

$$E[R_c^{t_j} | \boldsymbol{x}_c^{t_j - \Delta}, \theta] = \exp(\theta^\intercal\, \boldsymbol{x}_c^{t_j - \Delta}), \text{where } \boldsymbol{x}_c^{t_j - \Delta} = \begin{bmatrix} \boldsymbol{d}_c \\ \boldsymbol{m}_c^{t_j - \Delta} \end{bmatrix} \tag{2}$$

where we have combined the spatial and temporal covariates to simplify notation in the rest of the paper. Our approach is related to those in recent preprints that incorporate mobility into compartment models [17], however those approaches typically involve large-scale Monte Carlo simulations when performing inference. As we will show, the Hawkes process likelihood can be maximized without simulation via an efficient expectation-maximization algorithm.

2.2 Mathematical connection between Hawkes processes and compartmental models

Here we briefly review several variations of the Hawkes process in Equation 1 that can be connected to SEIR-type compartment models. The first variant is the SIR-Hawkes process, also referred to as HawkesN. This model captures the long-term evolution of a pandemic by incorporating a pre-factor that accounts for the dynamic decrease in the number of susceptible individuals [24]:

$$\lambda^{SIR}(t) = (1 - I_c(t)/N)(\mu + \sum_{t_i < t} R_0 w(t - t_i)). \tag{3}$$

Here $I_c(t)$ is the cumulative number of infections that have occurred up to time t and N is the total population size. The point process governed by Equation 3 is a continuous time analog of a discrete stochastic SIR model when $w(t)$ is specified to be exponential [24]. When $w(t)$ is chosen to be gamma distributed, the Hawkes process also can approximate staged compartment models, like SEIR, if the average waiting time in each compartment is equal [15]. More complex parametric (or non-parametric) inter-infection time distributions $w(t)$ may be employed within the Hawkes process framework in situations where disease dynamics cannot be captured by a SIR or SEIR model. In the early exponential growth stage of an epidemic, before finite population effects play a role (which is the case with current U.S. data), the Hawkes process in Equation 1 without the prefactor can be used to model new infections arising from SIR and SEIR models, as $I_c(t)/N$ will be small (see Section S1 in the supplemental material for more details).

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2.3 EM algorithm for parameter inference

We use an expectation–maximization (EM) algorithm to estimate the model in Equation 1. First, we introduce latent random variables, $p_c(i,j)$, that represent the event that secondary infection i is caused by primary infection j in county c. We let $p_c(i,i)$ represent the event that case i is imported. The complete data log-likelihood is then given by,

$$\begin{aligned}\mathcal{L} = \sum_{c=1}^{|\mathcal{C}|} \Bigg\{ \sum_{i=1}^{n} p_c(i,i)\log(\mu_c) - \int_0^T \mu_c \, dt + \\ \sum_{j=1}^{n} \Bigg\{ \sum_{i=j+1}^{n} p_c(i,j) \log\big[R_c^{t_j}(\boldsymbol{x}_c^{t_j-\Delta},\theta) w(t_i - t_j|\alpha,\beta)\big] - \int_{t_j}^{T} R_c^{t_j}(\boldsymbol{x}_c^{t_j-\Delta},\theta) w(t-t_j|\alpha,\beta)\, dt \Bigg\}\Bigg\}.\end{aligned} \tag{4}$$

Here we use a Weibull distribution with shape α and scale β to model inter-infection times, which is used in other studies of epidemics [20, 4, 10] and we find accurately captures transmission in the present data.

As the branching structure of the process is unobservable, we optimize the complete data log-likelihood in Equation 4 by iteratively alternating between an expectation step where the branching probabilities p_c are estimated and a maximization step where model parameters are updated by maximizing Equation 4. The EM-algorithm is equivalent to a projected gradient ascent on the likelihood of the Hawkes process [12].

2.3.1 Expectation step

During the expectation step, we estimate the latent variables $p_c(i,j)$ for each county. Given the parameters θ, α, β, and μ_c estimated from the last iteration, the probabilities that case i was caused by case j or was imported are given by:

$$p_c(i,j) = \frac{R_c^{t_j}(\boldsymbol{x}_c^{t_j-\Delta},\theta) w(t_i - t_j|\alpha,\beta)}{\lambda_c(t_i)}, \tag{5}$$

$$p_c(i,i) = \frac{\mu_c}{\lambda_c(t_i)}. \tag{6}$$

2.3.2 Maximization step

We then maximize the complete data log-likelihood with respect to the model parameters, conditioned on the estimated branching structure $p_c(i,j)$. During estimation we do not include event pairs (i,j) when j is within Ψ days of the last day of the dataset, as the offspring events i have not yet been realized and the inclusion of these incomplete data biases parameter estimates.

We approximate the integrals in Equation 4 as is done in [25] by noting that $\int_{t_j}^{T} w(t-t_j|\alpha,\beta) \approx 1$ (given we are removing the last Ψ days from the estimation).

Maximization of Equation 4 then decouples into three independent optimization problems. The first is a Poisson regression of observations $P_c(j) = \sum_{i=j+1}^{n} p_c(i,j)$ against the covariates $\boldsymbol{x}_c^{t_j}$:

$$\hat{\theta} \coloneqq \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}_\theta = \underset{\theta}{\operatorname{argmax}} \sum_{c=1}^{|\mathcal{C}|} \Bigg\{ \sum_{j=1}^{n} P_c(j)\theta^\intercal \boldsymbol{x}_c^{t_j-\Delta} - \exp(\theta^\intercal \boldsymbol{x}_c^{t_j-\Delta}) \Bigg\}. \tag{7}$$

The second optimization problem is weighted maximum likelihood estimation for the Weibull shape and scale parameters:

$$\hat{\alpha}, \hat{\beta} \coloneqq \underset{\alpha,\beta}{\operatorname{argmax}}\, \mathcal{L}_{\alpha,\beta} = \underset{\alpha,\beta}{\operatorname{argmax}} \sum_{c=1}^{|\mathcal{C}|} \Bigg\{ \sum_{j=1}^{n} \Bigg\{ \sum_{i=j+1}^{n} p_c(i,j) \log\big[w(t_i - t_j|\alpha,\beta)\big] \Bigg\}\Bigg\}. \tag{8}$$

where $p_c(i,j)$ is the weight of each inter-infection time observation $t_i - t_j$.

Third, the background rate μ_c is determined analytically:

$$\hat{\mu_c} \coloneqq \underset{\mu_c}{\operatorname{argmax}}\, \mathcal{L}_{\mu_c} = \underset{\mu_c}{\operatorname{argmax}} \sum_{i=1}^{n} p_c(i,i)\log(\mu_c) - \int_0^T \mu_c \, dt, \;\; \hat{\mu_c} = \sum_{i=1}^{n} \frac{p_c(i,i)}{T}. \tag{9}$$

Pseudo code for the EM algorithm is provided in Algorithm S1 in the supplementary material. We note that the EM algorithm of the Hawkes process is also connected to the dynamic reproduction number estimator of Wallinga and Teunis [28], as the latter can be viewed as a 1-iteration EM algorithm where a histogram estimator is used for R_c^t with initial guess $R_c^t \equiv 1$ (see Section S2 in the supplemental material for a derivation).

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2.4 Hawkes process forecasting

We forecast future events using the branching process representation of the Hawkes process. In particular, for each event in the history of the process we simulate a Poisson random variable with mean $R_c^{t_j}(\boldsymbol{x}_c^{t_j}, \theta)$ representing the number of secondary infections caused by event j. For each of these infections we simulate the time of infection by drawing inter-event times from the estimated Weibull distribution. Events falling in the future (past the forecasting date) are then used to update the forecasted intensity through Equation 1. We simulate multiple realizations of this process to estimate a mean intensity forecast along with confidence intervals.

3 Experiments and Results

In this section we first provide details on the datasets and baseline models used in our experiments. We then discuss the experimental results of several COVID-19 retrospective forecasting tasks at the U.S. county level. The source code and dataset are both available online*.

3.1 Datasets

3.1.1 Covid-19 daily cases and deaths reported by The New York Times

The New York Times (NYT) [26] † releases a daily report of the cumulative numbers of COVID-19 cases in the United States at the county level and over time. While NYT data closely tracks data aggregated by a project at Johns Hopkins University [5], NYT county level reporting started earlier and is therefore used in this study. In total, there are $2,920$ counties with cases and/or deaths in the dataset. The time series data are compiled from state and local government health departments. In order to have sufficient data for statistical inference, we select the counties with confirmed cases greater than and equal to 10 (denoted by $\mathcal{D}_{\text{conf}}$) and the counties with at least 1 death (denoted by $\mathcal{D}_{\text{death}}$) by 05/20/2020. In total, there are $2,113$ and $1,236$ counties in these two datasets.

Parameter sharing may improve models in counties with less data through variance reduction, but can potentially bias estimates in more populated counties with more cases. We therefore assess model performance over different subsets of counties grouped by case volume. We first rank counties by the number of confirmed cases and deaths by the cut-off date, 05/20/2020, and we then evaluate forecasting accuracy on the top-10% of counties (denoted by $Q_{10\%}^{\text{top}}$), the top-25% counties ($Q_{25\%}^{\text{top}}$), and counties between the top-25% and top-50% quantiles (denoted by $Q_{50\%}^{25\%}$).

3.1.2 Google mobility index reports

We use Google daily mobility index reports at the county level [9] to estimate a dynamic reproduction number that tracks changes in movement patterns due to stay at home orders (and their staged removal). In total, there are 6 mobility types, including grocery & pharmacy, Parks, transit stations, retail & recreation, residential and workplaces. Mobility indices for each category and county are calculated with respect to a baseline value for that day of the week ‡. We drop "workplace" mobility from our analysis due to high collinearity with "residential" mobility. Some mobility data are missing when data is sparse for a given date. To deal with missing values, we adopt multivariate feature imputation §, which estimates each missing mobility entry as a function of other mobility types on the same day in the same county. We show heatmaps of mobility patterns across counties and time in the Figure S4 in the supplemental material, where a major change can be observed coinciding with stay at home orders (the first state-wide stay-at-home order was issued at 03/21/2020).

3.1.3 County-level demographic covariates

We incorporate spatial demographic features that may be predictive of symptomatic cases of COVID-19 (which are more likely to result in testing and mortality). The dataset is available in a curated form [2] and is derived from CDC and census datasets. The data is at the county level and includes population, median age, number of hospitals and ICU beds, percentage of smokers and diabetes, and heart disease mortality.

*https://github.com/chiangwe/HawkPR

†https://github.com/nytimes/covid-19-data

‡The baseline is the median value, for the corresponding day of the week calculated during the 5-week period, 01/03/2020 to 02/06/2020.

§https://scikit-learn.org/stable/modules/impute.html#multivariate-feature-imputation

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1: Model performance on MAE

Mdl	Confirmed cases $\mathcal{D}_{\text{conf}}$									Death cases $\mathcal{D}_{\text{death}}$								
	3-days			5-days			7-days			3-days			5-days			7-days		
	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$
PROJ	187.47	87.43	5.79	324.50	153.04	9.59	497.29	230.73	14.70	14.88	7.66	1.09	31.05	14.69	1.64	45.23	20.97	2.09
CLEP	116.46	57.94	6.96	201.27	98.16	12.27	352.17	170.89	16.63	15.35	7.71	1.26	22.64	11.63	1.94	27.05	23.42	2.52
Hawkes	103.61	51.52	6.12	183.42	89.57	9.66	227.23	110.27	13.13	13.35	6.77	1.08	20.43	10.40	1.68	23.69	12.37	2.14
HawkPR$_{\text{m}}$	95.95	**47.91**	5.89	134.60	68.97	8.54	178.93	93.56	11.63	11.96	6.14	1.04	16.25	8.52	1.52	23.07	11.28	1.87
HawkPR$^{+}_{\text{m}}$	**95.53**	47.94	**5.76**	**133.13**	**68.47**	**8.39**	**174.35**	**89.89**	**11.35**	**11.59**	**6.00**	**1.02**	**15.82**	**8.28**	**1.48**	**21.57**	**11.09**	**1.85**

The best performance is marked in **bold**.

Table 2: Model performance on NDCG

Mdl	Confirmed cases $\mathcal{D}_{\text{conf}}$									Death cases $\mathcal{D}_{\text{death}}$								
	3-days			5-days			7-days			3-days			5-days			7-days		
	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$
PROJ	0.933	0.932	0.776	0.940	0.939	0.818	0.958	0.959	0.826	0.916	0.915	0.716	0.925	0.923	0.763	0.927	0.926	**0.819**
CLEP	0.928	0.929	0.796	0.947	0.949	0.824	0.866	0.855	**0.848**	0.939	0.938	0.717	0.972	0.963	0.766	0.940	0.949	0.803
Hawkes	0.964	0.964	0.794	0.974	0.974	**0.844**	0.938	0.939	0.845	0.955	0.951	**0.721**	**0.979**	**0.977**	**0.768**	**0.971**	**0.970**	0.790
HawkPR$_{\text{m}}$	**0.971**	**0.974**	**0.796**	0.974	0.973	0.839	**0.982**	**0.980**	0.834	0.952	0.951	0.702	0.975	0.975	0.760	0.966	0.965	0.789
HawkPR$^{+}_{\text{m}}$	0.966	0.966	**0.796**	**0.976**	**0.976**	0.836	0.971	0.970	0.835	**0.962**	**0.960**	0.708	0.969	0.968	0.753	0.965	0.965	0.786

The best performance is marked in **bold**.

3.2 Baseline models and experimental setup for retrospective forecasting comparison

We compare the Hawkes process model in Equation 1 with several benchmarks including a SEIR model used in a pandemic tracking dashboard out of Columbia University [22] (denoted by **PROJ**) and an ensemble model with linear and exponential predictors from University of California, Berkeley [2] (denoted by **CLEP**). A simplified Hawkes process, denoted by **Hawkes**, where the reproduction number is held constant is used for comparison to demonstrate the effectiveness of tracking the reproduction number dynamically. We also compare our full Hawkes process model, denoted by **HawkPR**$^{+}_{\text{m}}$, to a Hawkes process, **HawkPR**$_{\text{m}}$, with only mobility features to determine the marginal improvement of adding demographics.

We backtest the five competing models on the $\mathcal{D}_{\text{conf}}$ and $\mathcal{D}_{\text{death}}$ datasets using the "walk-forward" validation approach. In particular, for 7-day forecasts we first train the models based on cases and deaths before the first cut-off date, 04/15/2020, and then forecast through 04/21/2020. We then slide the forecasting window, training on data before 4/22/2020 and forecasting from 04/22/2020 to 04/28/2020. We repeat this process until the final date of 05/19/2020 (a similar approach is used for 3 and 5 day forecasts). The multivariate imputation models are also trained in the same walk forward fashion to avoid possible data leakage. We evaluate the models according to mean absolute error, MAE, averaged across counties and forecasting windows of the same length, along with percentage error, PE. We also compare the ranking quality of the competing models using Normalized Discounted Cumulative Gain (NDCG).

3.3 Experimental results

In Table 1 and 2, we present the experimental results for all models applied to both confirmed cases ($\mathcal{D}_{\text{conf}}$) and deaths ($\mathcal{D}_{\text{death}}$) and in Figure 2 and 3, we show example forecasts, along with confidence intervals, for the Top-6 counties in $Q^{\text{top}}_{10\%}$ and $Q^{25\%}_{50\%}$. ¶.

In Table 1 we show the MAE of each method for 3, 5, and 7 day window forecasts and in Table 2 we show NDCG scores for the same experiments. In terms of MAE, both our proposed models, **HawkPR**$_{\text{m}}$ and **HawkPR**$^{+}_{\text{m}}$, outperform the benchmarks, **PROJ** and **CLEP**, by a large margin in all three forecasting periods and across quantile subsets of the data. The proposed **HawkPR**$_{\text{m}}$ and **HawkPR**$^{+}_{\text{m}}$ models are also the best performing models in terms of PE (see Table S1 in the supplementary material).

We also find that adding mobility indices improves the baseline Hawkes process, **Hawkes**, where **HawkPR**$_{\text{m}}$ has improvements of up to 27% (5 day forecast) and 23% (7 day forcast) over **Hawkes** when applied to $Q^{\text{top}}_{10\%}$. By adding demographic features, we can marginally boost the MAE of **HawkPR**$^{+}_{\text{m}}$ over **HawkPR**$_{\text{m}}$. Generally, the proposed models have a better NDCG performance when applied to confirmed cases for most of the quantile subsets. The baseline Hawkes process, **Hawkes**, performs the best in terms of NDCG on the $\mathcal{D}_{\text{death}}$ dataset.

¶More examples of $\mathcal{D}_{\text{conf}}$ forecasts are presented in the Figure S2 and Figure S3 in the supplementary file.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3: Model coefficients ($\mathcal{D}_{\text{death}}$)

Covariate	coef	SE	pValue
Retail/recreation	0.2720	0.0107	$2.92 \times 10^{-141}*$
Grocery/pharmacy	0.1058	0.0058	$6.82 \times 10^{-073}*$
Residential	−0.0723	0.0071	$2.37 \times 10^{-023}*$
Transit stations	−0.0383	0.0086	$8.28 \times 10^{-005}*$
Parks	0.0021	0.0030	4.90×10^{0}
Diabetes percentage	−0.1766	0.0104	$3.41 \times 10^{-064}*$
# hospitals	−0.0927	0.0096	$5.14 \times 10^{-021}*$
Smokers percentage	−0.0682	0.0070	$1.90 \times 10^{-021}*$
Population estimate	0.0681	0.0076	$3.60 \times 10^{-018}*$
Population Density	0.0521	0.0020	$8.72 \times 10^{-148}*$
# ICU beds	0.0416	0.0058	$5.26 \times 10^{-012}*$
Median age	0.0290	0.0064	$6.08 \times 10^{-005}*$
Heart disease mort.	0.0107	0.0082	1.92×10^{0}
Intercept	0.1786	0.0095	-

The first 5 covariates are mobility indices, followed by static demographic covariates.

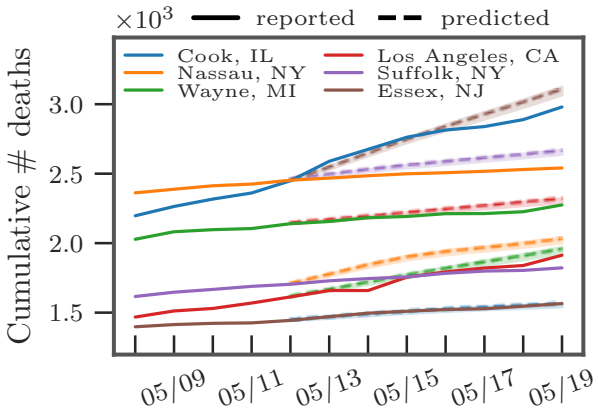


Figure 2: Top-6 counties in $Q^{\text{top}}_{10\%}$ of $\mathcal{D}_{\text{death}}$

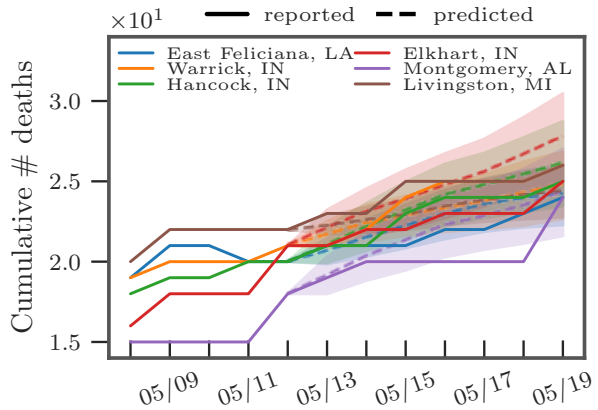


Figure 3: Top-6 counties in $Q^{25\%}_{50\%}$ of $\mathcal{D}_{\text{death}}$

In Table 3, we show the dynamic reproduction number coefficients of $\mathbf{HawkPR}^{+}_{\text{m}}$ estimated from the Poisson regression when applied to $\mathcal{D}_{\text{death}}$.[‖] The absolute value of the coefficients indicates the magnitude of the correlation between the reproduction number and the features. With the exception of heart disease mortality and parks, the coefficients of all variables are statistically significant at the 10^{-4} level or below. The dynamic reproduction number is positively correlated with retail and recreation, meaning that as mobility shifted away from commercial areas towards residences the reproduction number decreased. The reproduction number is negatively correlated with percent of the population with diabetes. Several possible explanations for this observation include high-risk individuals are being more cautious or that they tend to live in areas with less cases, potentially with less population.

In Figure 4, we find that the estimated dynamic reproduction number closely tracks lagged mobility, where the optimal lag parameter is determined as $\Delta = 14$ days. The top-2 counties have estimated reproduction number initially above 3, however after stay-at-home orders mobility decreased and the reproduction number fell to around 1 (which explains why curves are relatively "flat" in many areas in the U.S.). As we observe from the tail of the reproduction number curve, a "reoponing" after 04/20/2020 coincides with a slight uptick in the reproduction number.

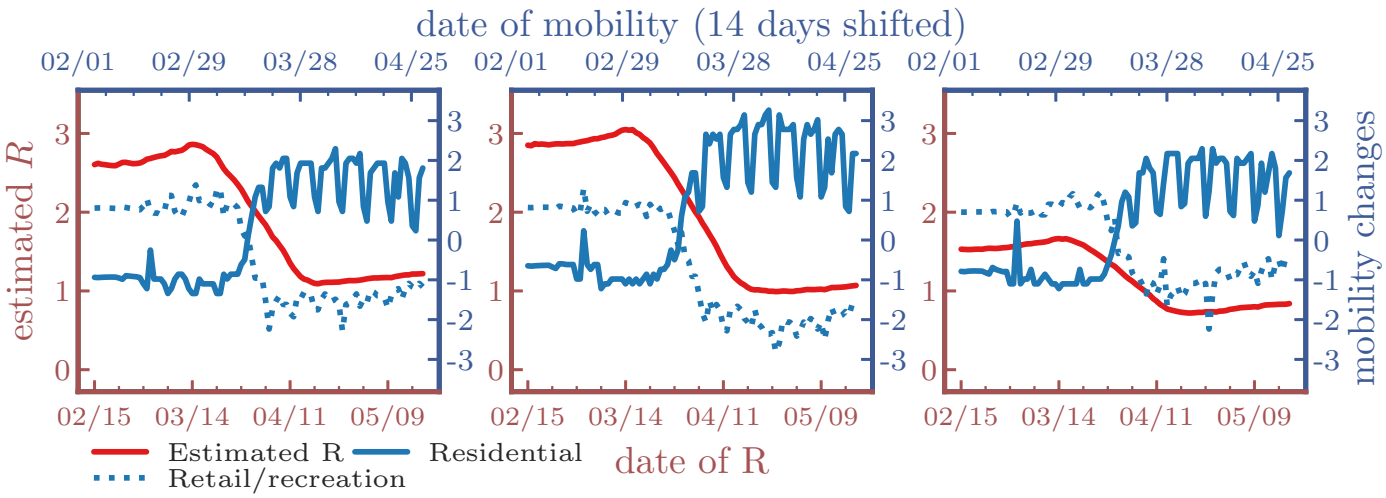
4 Conclusion

We showed how Hawkes processes can be combined with spatio-temporal covariates to accurately model COVID-19 transmission and forecast future cases and deaths. The model is competitive with several benchmark models used to forecast the pandemic, achieving improved MAE and NDCG scores on a majority of the experiments we conducted. Our hope is that this work will encourage more research into Hawkes process models of disease spreading that incorporate more advanced features and machine learning principles.

One potential direction for future research is extending the work here to neural network based point process models

[‖] The model coefficients for confirmed cases is included in Table S3 of the supplementary file.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



(a) Cook, IL (1st in $Q^{\text{top}}_{10\%}$) (b) Nassau, NY (2nd in $Q^{\text{top}}_{10\%}$) (c) East Feliciana, LA (1st in $Q^{25\%}_{50\%}$)

Figure 4: Estimated R of death cases $\mathcal{D}_{\text{death}}$ and lagged mobility changes ($\Delta = 14$ days).

[16, 21]. These models may be able to capture more complicated relationships between mobility patterns, demographics, and transmission. The challenges of such an approach include the potential for over-fitting with added parameters and determining how best to realistically model transmission in a neural point process (analagous to the SIR-Hawkes process), which will be important if neural point processes are to be used in long-term forecasting.

Many of the preprints and models currently being released on academic archives and websites present a single model without model evaluation, goodness of fit analysis, or comparison to baselines. Here we believe the "common task framework" [6] could be beneficial in model selection and validation. The machine learning community can contribute to pandemic modeling efforts by performing careful benchmarking of methodologies, creating standardized datasets and tasks, and comparing competing models that come from different fields such as epidemiology, statistics, and machine learning.

Broader Impact Statement

We propose a novel Hawkes process model that incorporates spatio-temporal mobility and demographic features to improve forecasts of COVID-19 transmission. These models can help guide health policy decisions during the pandemic to prevent spread in forecasted hotspots and also may serve as a bridge between the machine learning community, who have developed high-dimensional and non-linear versions of these models, and the epidemiological modeling community, who utilize compartment models to which Hawkes processes are mathematically connected. We caution that forecasting COVID-19, especially in the long-term, is challenging due to uncertainties in the data, human behavior, and future policy decisions.

5 Acknowledgements

This research was supported by NSF grants SCC-1737585 and ATD-1737996.

References

[1] https://ai.facebook.com/blog/using-ai-to-help-health-experts-address-the-covid-19-pandemic.

[2] Nick Altieri, Rebecca L Barter, James Duncan, Raaz Dwivedi, Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Chandan Singh, Yan Shuo Tan, et al. Curating a covid-19 data repository and forecasting county-level death counts in the united states. *arXiv preprint arXiv:2005.07882*, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[3] IHME COVID, Christopher JL Murray, et al. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv*, 2020.

[4] Benjamin J Cowling, Max SY Lau, Lai-Ming Ho, Shuk-Kwan Chuang, Thomas Tsang, Shao-Haei Liu, Pak-Yin Leung, Su-Vui Lo, and Eric HY Lau. The effective reproduction number of pandemic influenza: prospective estimation. *Epidemiology (Cambridge, Mass.)*, 21(6):842, 2010.

[5] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 2020.

[6] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[7] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J Smola, and Le Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 219–228, 2015.

[8] Ferguson et al. Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. 2020. DOI: https://doi.org/10.25561/77482.

[9] Google. Covid-19 community mobility report, 2020.

[10] Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Fiona Sun, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 2020.

[11] Nicholas P Jewell, Joseph A Lewnard, and Britta L Jewell. Caution warranted: using the institute for health metrics and evaluation model for predicting the course of the covid-19 pandemic. *Annals of Internal Medicine*, 2020.

[12] Erik Lewis and George Mohler. A nonparametric em algorithm for multiscale hawkes processes. 2011.

[13] Bouardi Hamza Li, Michael and Omar Lami. Coronavirus in the u.s.: Latest map and case count, Mar 2020.

[14] Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song. Learning temporal point processes via reinforcement learning. In *Advances in neural information processing systems*, pages 10781–10791, 2018.

[15] Alun L Lloyd. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1470):985–993, 2001.

[16] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.

[17] Andrew C Miller, Nicholas J Foti, Joseph A Lewnard, Nicholas P Jewell, Carlos Guestrin, and Emily B Fox. Mobility trends provide a leading indicator of changes in sars-cov-2 transmission. *medRxiv*, 2020.

[18] George Mohler, Jeremy Carter, and Rajeev Raje. Improving social harm indices with a modulated hawkes process. *International Journal of Forecasting*, 34(3):431–439, 2018.

[19] George Mohler and Michael D Porter. Rotational grid, pai-maximizing crime forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(5):227–236, 2018.

[20] Thomas Obadia, Romana Haneef, and Pierre-Yves Boëlle. The R_0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC medical informatics and decision making*, 12(1), 2012.

[21] Takahiro Omi, Kazuyuki Aihara, et al. Fully neural network based model for general temporal point processes. In *Advances in Neural Information Processing Systems*, pages 2120–2129, 2019.

[22] Sen Pei and Jeffrey Shaman. Initial simulation of sars-cov2 spread and intervention effects in the continental us. *medRxiv*, 2020.

[23] Alex Reinhart and Joel Greenhouse. Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1305–1329, 2018.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[24] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. Sir-hawkes: linking epidemic models and hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference*, pages 419–428, 2018.

[25] Frederic Paik Schoenberg. Facilitated estimation of etas. *Bulletin of the Seismological Society of America*, 103(1):601–605, 2013.

[26] The New York Times. Coronavirus in the u.s.: Latest map and case count, Mar 2020.

[27] Utkarsh Upadhyay, Abir De, and Manuel Gomez Rodriguez. Deep reinforcement learning of marked temporal point processes. In *Advances in Neural Information Processing Systems*, pages 3168–3178, 2018.

[28] Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516, 2004.

[29] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726, 2016.

[30] Hongteng Xu and Hongyuan Zha. A dirichlet mixture model of hawkes processes for event sequence clustering. In *Advances in Neural Information Processing Systems*, pages 1354–1363, 2017.

[31] Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309, 2013.

[32] Difan Zou, Lingxiao Wang, Pan Xu, Jinghui Chen, Weitong Zhang, and Quanquan Gu. Epidemic model guided machine learning for covid-19 forecasts in the united states. *medRxiv*, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates (Supplementary material)

Wen-Hao Chiang, Xueying Liu, George Mohler**

Department of Computer & Information Science
Indiana University–Purdue University Indianapolis
Indianapolis, IN 46224
{chiangwe, xl17, gmohler}@iupui.edu

June 6, 2020

S1 Linear Hawkes processes can approximate SEIR during early stages

While a pre-factor in the Hawkes process involving the cumulative number of infections, $I_c(t)$, is necessary to model long-term disease dynamics [24],

$$\lambda(t) = (1 - I_c(t)/N)(\mu + \sum_{t_i<t} R_0 w(t - t_i)), \tag{1}$$

in the early stages of transmission a linear Hawkes process can be used (as the prefactor will be close to 1),

$$\lambda(t) \approx \mu + \sum_{t_i<t} R_0 w(t - t_i). \tag{2}$$

To illustrate this, we simulate a SEIR differential equation $dS/dt = -\beta SI/N$, $dE/dt = \beta SI/N - \mu E$, $dI/dt = \mu E - \gamma I$, $dR/dt = \gamma I$, $\beta = \gamma R_0$, where the parameters are chosen similar to those of COVID-19: $\gamma = .1$, $R_0 = 2$, $\mu = 1$, and $N = 5 \cdot 10^8$. We then fit the linear Hawkes process model in Equation 2 to new infections, μE, generated by the SEIR model. We use a non-parametric histogram estimator for $w(t)$ and find a close fit between the Hawkes process and the SEIR model in Figure S1.

S2 Connection of EM algorithm for Hawkes Process and dynamic R estimator of Wallinga and Teunis

Here we make the connection between the EM algorithm for the Hawkes process and the popular dynamic reproduction number estimator of Wallinga and Teunis [S1, 28, S2]. The dynamic R estimator of Wallinga and Teunis is constructed as follows. The probability that individual i at time t_i was infected by individual j at time t_j is defined to be,

$$p_{ij} = \frac{w(t_i - t_j)}{\sum_{t_i > t_k} w(t_i - t_k)}, \tag{3}$$

where the distribution of inter-infection times $w(t_i - t_j)$ is typically modeled as Weibull, Gamma, or log-normal [S2]. The expected total number of individuals that j infects is then given by:

$$R_j = \sum_{i>j} p_{ij} \tag{4}$$

*Corresponding author
**Corresponding author

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Wallinga and Teunis then obtain an estimate of the dynamic reproduction number $R(t)$ by averaging R_j over all observed cases j where the time of infection t_j occurred on day t:

$$R(t) = \frac{1}{N_t} \sum_{t \leq t_j < t+1} R_j, \tag{5}$$

(here N_t is the number of observed infections on day t).

On the other hand, for the Hawkes process the intensity (rate) of infections is modeled as

$$\lambda(t) = \mu + \sum_{t > t_i} R(t_i) w(t - t_i), \tag{6}$$

where $w(t)$ and $R(t)$ are the inter-infection time distribution and dynamic reproduction number respectively. Rather than modeling $R(t)$ as dependent on mobility, we can instead model $R(t)$ as a piecewise constant function:

$$R(t) = \sum_{k=1}^{B} r_k 1\{t \in I_k\}. \tag{7}$$

Here the I_k are intervals discretizing time, B is the number of such intervals, and r_k is the estimated reproduction rate in interval k.

Given initial guesses for the model parameters and r_k, the EM algorithm for the Hawkes process iteratively updates the parameters and branching probabilities by alternating between the **E-step update**:

$$p_{ij} = R(t_j) w(t_i - t_j) / \lambda(t_i) \tag{8}$$

$$p_{ii} = \mu / \lambda(t_i) \tag{9}$$

and **M-step update:**

$$w(t) \sim MLE(\{t_i - t_j; p_{ij}\}) \tag{10}$$

$$\mu = \sum_i p_{ii} / T \tag{11}$$

$$r_k = \sum_{t_i > t_j} p_{ij} 1\{t_j \in I_k\} / N_k \tag{12}$$

where T is the total length of the observation period, N_k is the total number of events in interval k, and the $w(t)$ is estimated via weighted MLE (for either a Gamma, Weibull or log-normal) using the inter-event times as observations and branching probabilities as weights.

Finally, we can compare Equation 8 to Equation 3. The dynamic $R(t)$ estimator in Equation 3 is what you obtain with 1 step of the EM algorithm in Equation 8 with initial guess $R(t) \equiv 1$, $\mu = 0$ and 1 day chosen as the bin width for the histogram estimator.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

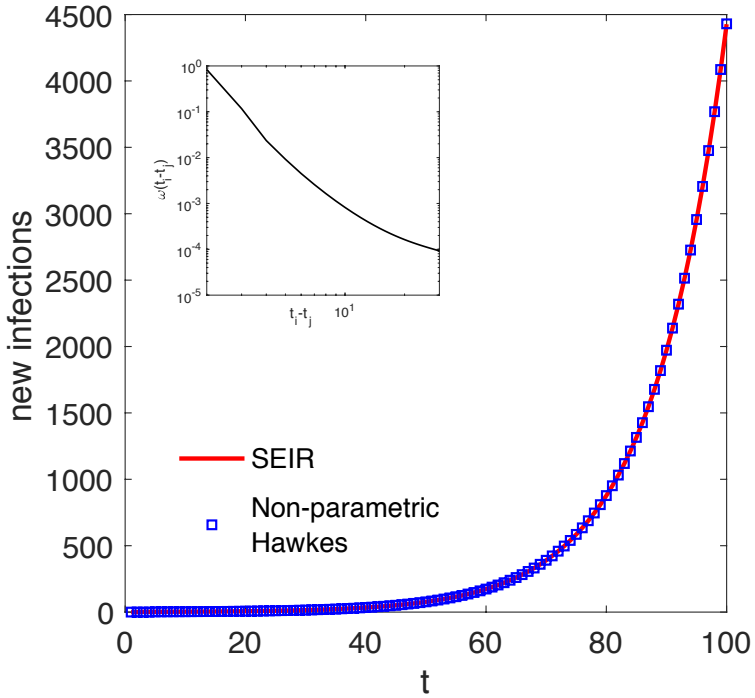


Figure S1: Main figure: (Red) SEIR differential equation $dS/dt = -\beta SI/N$, $dE/dt = \beta SI/N - \mu E$, $dI/dt = \mu E - \gamma I$, $dR/dt = \gamma I$, where $\beta = \gamma R_0$, $\gamma = .1$, $R_0 = 2$, $\mu = 1$, and $N = 5 \cdot 10^8$. (Blue squares) linear Hawkes process $\lambda_t = \mu + \sum_{t>t_i} R_0 w(t - t_i)$ fit to the SEIR curve of new infections. Inset: Non-parametric histogram estimate for $w(t)$.

S3 Hawkes process model of COVID-19 transmission

S3.1 EM algorithm for parameter inference

In the Algorithm S1, we present the pseudo code for the model $\mathbf{HawkPR}_{\mathrm{m}}^{+}$. The details of the derivation is elaborated in the Section 2.3 in the manuscript. After initialization, we iterate between the expectation and maximization steps to estimate the parameters until the algorithm converges (parameter changes between consecutive iterations are within a specified tolerance).

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Algorithm S1 EM algorithm optimization

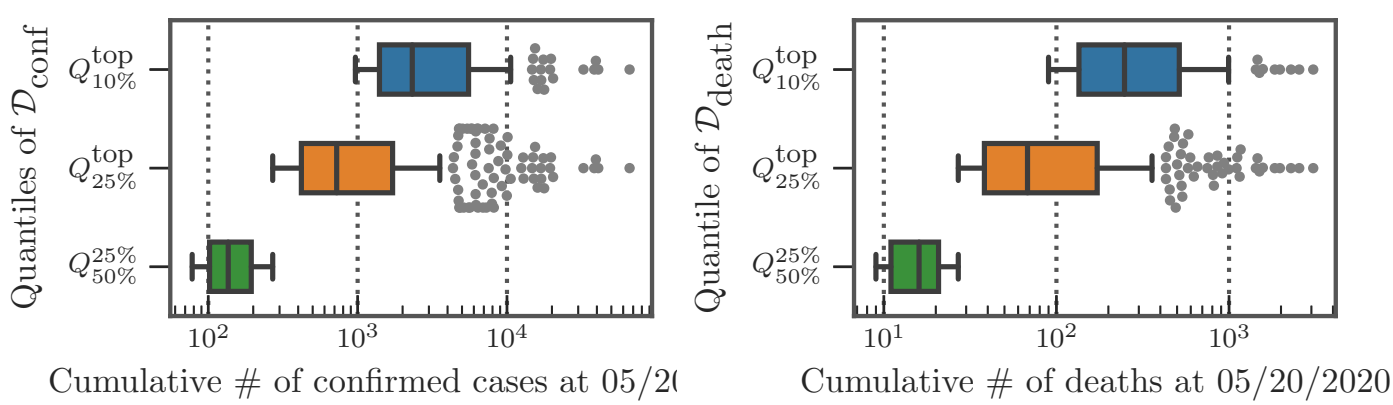
procedure $\textbf{HawkPR}^{+}_{\text{M}}(\mathcal{T}, \boldsymbol{x}, \Delta)$
  $T \leftarrow \max \mathcal{T},\ \alpha \leftarrow 2,\ \beta \leftarrow 2.$ ▷ Initialization
  $\mu_c \leftarrow 0.5,\ R^t_c(t) \leftarrow 1,\ \forall c \in \mathcal{C}$ and $0 < t < T$.
  while $\|\Delta\theta\|,|\Delta\alpha|,|\Delta\beta|,\|\Delta\mu\| >$tol **do**
    Expectation step:
    for $\forall i \geq j$ and $0 < i, j < T$ and $\forall c \in \mathcal{C}$ **do**
      if $i > j$ **then**
        $p_c(i,j) \leftarrow \frac{R_c^{t_j}(\boldsymbol{x}_c^{t_j-\Delta},\theta) w(t_i - t_j|\alpha,\beta)}{\lambda_c(t_i)}.$
      else if $i = j$ **then**
        $p_c(i,i) \leftarrow \frac{\mu_c}{\lambda_c(t_i)}.$
      end if
    end for

    Maximization step:
    $\theta \leftarrow \text{argmax}_\theta \sum_{c=1}^{|\mathcal{C}|} \left\{ \sum_{j=1}^{n} P_c(j)\theta^\intercal \boldsymbol{x}_c^{t_j-\Delta} - \exp(\theta^\intercal \boldsymbol{x}_c^{t_j-\Delta}) \right\}.$
    $\alpha, \beta \leftarrow \text{argmax}_{\alpha,\beta} \sum_{c=1}^{|\mathcal{C}|} \left\{ \sum_{j=1}^{n} \left\{ \sum_{i=j+1}^{n} p_c(i,j) \log\left[w(t_i - t_j|\alpha,\beta)\right] \right\} \right\}$
    for $\forall c \in \mathcal{C}$ **do**
      $\mu_c \leftarrow \sum_{i=1}^{n} \frac{p_c(i,i)}{T}.$
    end for
  end while
end procedure

S4 Experiments

S4.1 Datasets

S4.1.1 Covid-19 report in the USA by The New York Times



(a) Cumulative number of confirmed cases ($\mathcal{D}_{\text{conf}}$)

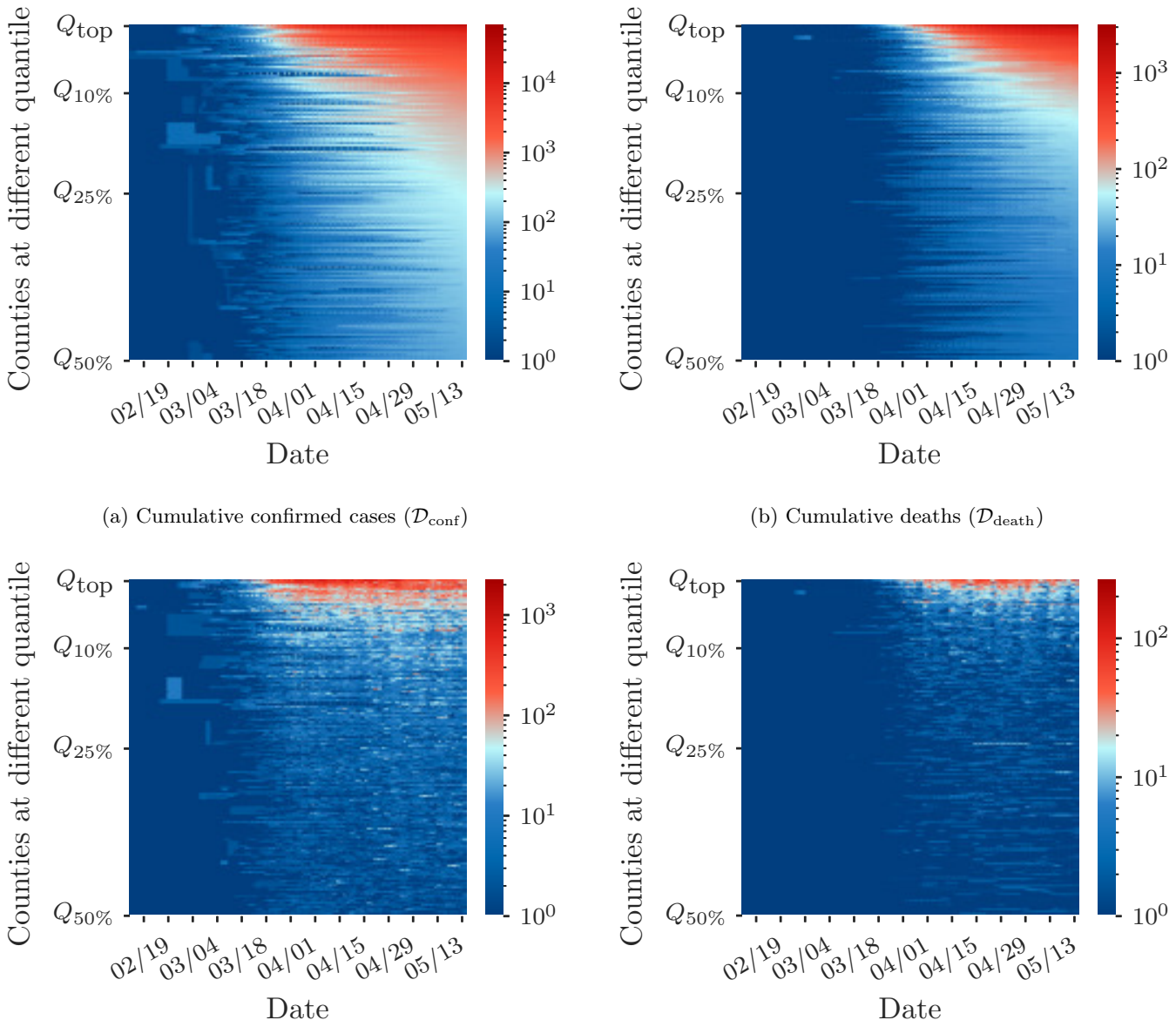
(b) Cumulative number of deaths ($\mathcal{D}_{\text{death}}$)

Figure S2: Distribution at different quantile

As described in Section 3.1.1 in the manuscript, we first rank counties by the number of confirmed cases and deaths by the cut-off date, 05/20/2020 and evaluate models on the top-10% of counties (denoted by $Q^{\text{top}}_{10\%}$), the top-25% counties ($Q^{\text{top}}_{25\%}$), and counties between the top-25% and top-50% quantiles (denoted by $Q^{25\%}_{50\%}$). In Figure S2, we show boxplot of the distribution of the cumulative number of confirmed cases/deaths in $Q^{\text{top}}_{10\%}$, $Q^{\text{top}}_{25\%}$, and $Q^{25\%}_{50\%}$ by the cut-off date.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Counties with cumulative number of cases more than 1.5 times the inter-quartile range are shown individually (gray circles). The groups $Q^{\text{top}}_{10\%}$ and $Q^{\text{top}}_{25\%}$ represent the counties hit hardest by COVID-19.



(a) Cumulative confirmed cases ($\mathcal{D}_{\text{conf}}$)

(b) Cumulative deaths ($\mathcal{D}_{\text{death}}$)

(c) Daily confirmed cases ($\mathcal{D}_{\text{conf}}$)

(d) Daily death cases ($\mathcal{D}_{\text{death}}$)

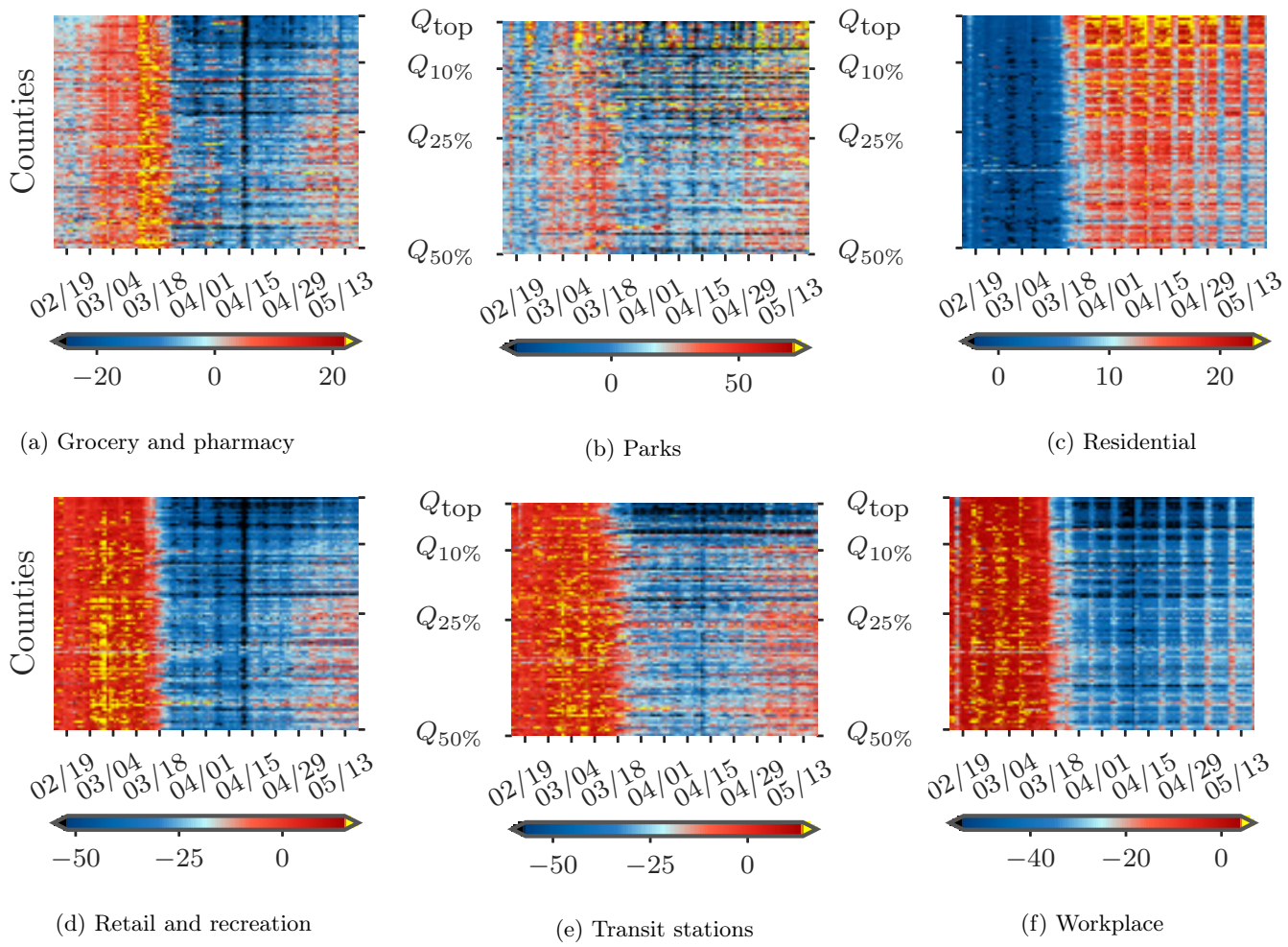
Figure S3: Heat map of the cumulative (top) and daily (bottom) number of cases (left) and deaths (right) over time and across counties.

In Figure S3, we show how the cumulative and daily number of confirmed cases and deaths progress through time. As of 05/20/20, in the U.S. about 25% of counties have more than 1,000 confirmed cases and 10% of counties have a death toll higher than 1,000. We observe a significant increase of confirmed cases in both Figure S3a and S3c starting around 03/25/2020. We can also see a jump in Figure S3b and S3d in the reported deaths around 04/07/2020, when the the U.S. reported more than 2,000 deaths in a single day for the first time.

S4.1.2 Google mobility index reports

In Figure S4, we present heat maps of mobility indices over time for counties in $\mathcal{D}_{\text{conf}}$. There are 6 types of mobility indices provided by Google, including grocery and pharmacy, park, residential, retail and recreation, transit stations,

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



(a) Grocery and pharmacy (b) Parks (c) Residential

(d) Retail and recreation (e) Transit stations (f) Workplace

Figure S4: Heat map of mobility indices across counties in $\mathcal{D}_{\text{conf}}$ and over time.

and workplace. Starting around 03/19/2020, the mobility patterns change drastically in most counties corresponding to school and non-essential business closings, along with stay at home orders. In Figure S4, we observe a mobility shift at this time away from recreation, transit and workplace towards residential areas. We can also observe a spike in grocery and pharmacy mobility preceding the stay at home orders, potentially caused by stockpiling of food and other essential items. As states across the U.S. begin staged reopening after 04/16/2020, grocery, parks, retail and transit mobility indices have moved towards baseline levels (though are still lower than before). Workplace and residential patterns have not yet returned to baseline as of 05/20/20.

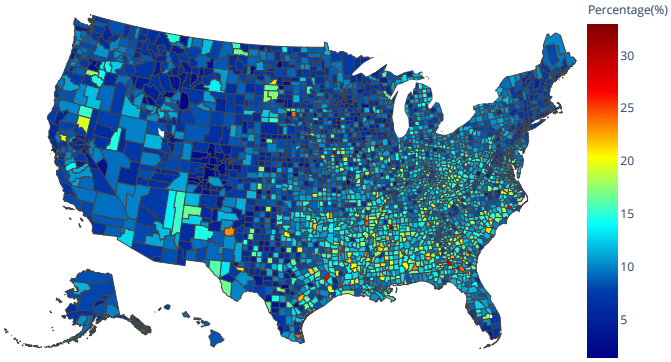
S4.1.3 County-level demographic covariates

In Figure S5, we present four examples of spatial demographic features at the county-level used to model variations in the reproduction number. In Figure S5a we can observe that the east part of USA is more densely populated compared to the west (population and population density are positively correlated with the reproduction number). Smoker and diabetes percentage, on the other hand, are negatively correlated in our regression with the reproduction number of COVID-19.

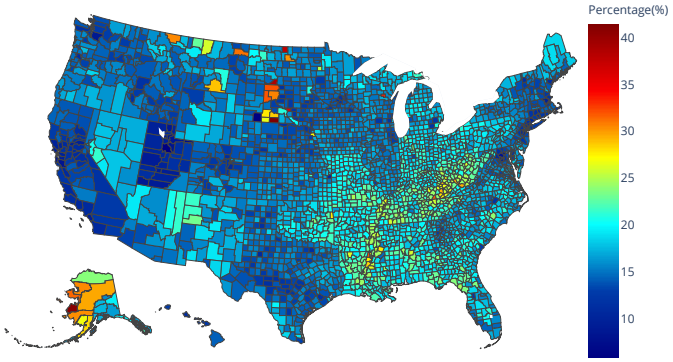
S4.2 Experimental results

In Table S1, we present the percentage error for all models applied to both confirmed cases ($\mathcal{D}_{\text{conf}}$) and deaths ($\mathcal{D}_{\text{death}}$) and in Figure S6 and S7, we show example forecasts, along with confidence intervals, for the Top-6 counties in $Q^{\text{top}}_{10\%}$ and $Q^{25\%}_{50\%}$ for confirmed cases. The proposed **HawkPR**$_{\text{m}}$ and **HawkPR**$^{+}_{\text{m}}$ models outperform the benchmarks, **PROJ** and **CLEP**, across all three forecasting periods and across quantile subsets of the data. Compared to the baseline **Hawkes**

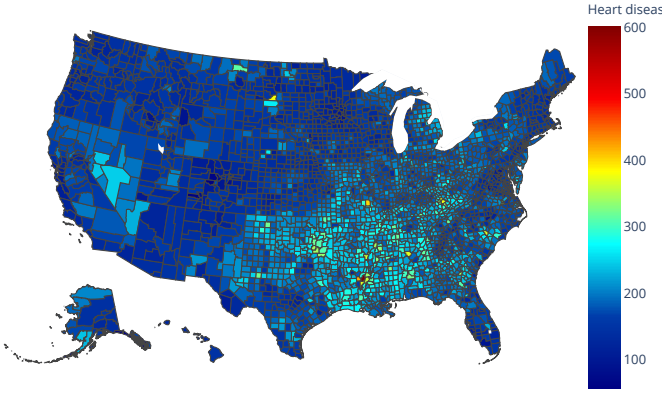
medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



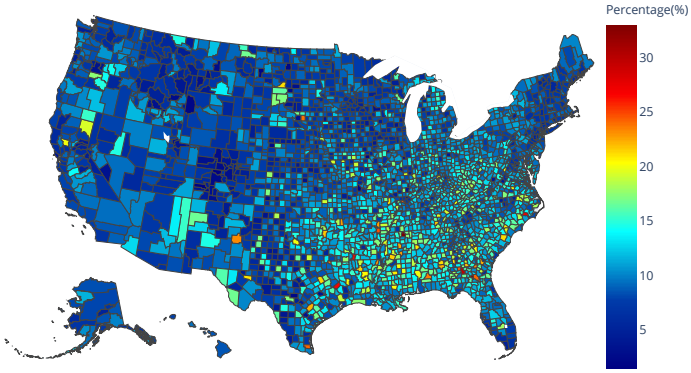
(a) Population density estimated at 2019



(b) Smoker percentage (%)



(c) Heart disease mortality



(d) Diabetes percentage (%)

Figure S5: Examples of spatial demographic and health features at the county-level.

where the reproduction number is held constant, our model can also lower the percentage error through the introduction of the dynamic reproduction number estimation.

In Table S2, we show the dynamic reproduction number coefficients of $\mathbf{HawkPR}_{\mathrm{m}}^{+}$ estimated from the Poisson regression when applied to $\mathcal{D}_{\mathrm{conf}}$. With the exception of "Diabetes percentage," the coefficients of all variables have significance below the level 10^{-4}. Again, the dynamic reproduction number is most positively correlated with "retail and recreation." Whereas the coefficient of heart disease mortality was not significant in the regression using death data, the coefficient is statistically significant when using cases. We note that heart disease mortality and diabetes percentage are positively correlated with each other, potentially explaining why they switched importance across regressions on cases and deaths. The sign of the coefficients is consistent, however, across regressions indicating that the reproduction number is lower for cases in counties with higher levels of smoking, heart disease and diabetes.

In Figure S8, we present examples of the estimated reproduction number R along with lagged mobility (optimal lag chosen as $\Delta = 14$ days). Before stay-at-home orders, the top-2 counties have estimated reproduction number above 3. However, as more Americans adopted social distancing and sheltered at home, mobility decreased and the reproduction number fell to around 1. As we observe from the estimated R after staged reopening begins (around 04/20/2020), the increased mobility is associated with a slight increase in the reproduction number.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table S1: Performances on PE (%)

Mdl	Confirmed cases $\mathcal{D}_{\text{conf}}$									Death cases $\mathcal{D}_{\text{death}}$								
	3-days			5-days			7-days			3-days			5-days			7-days		
	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$	$Q^{\text{top}}_{10\%}$	$Q^{\text{top}}_{25\%}$	$Q^{25\%}_{50\%}$
PROJ	82.15	23.02	23.86	30.57	30.23	10.20	57.65	52.52	11.37	30.95	30.39	37.01	34.67	24.61	**15.39**	44.96	35.09	23.87
CLEP	22.65	21.36	30.91	25.58	25.06	19.48	28.86	28.18	17.25	46.30	47.05	51.73	39.40	36.73	33.69	14.00	14.08	14.37
Hawkes	20.50	19.25	**17.69**	18.44	**16.40**	**6.64**	19.12	17.87	11.88	37.56	31.62	25.03	33.35	30.32	19.36	14.76	12.55	9.79
HawkPR$_{\text{m}}$	**9.65**	**9.22**	19.98	**13.76**	17.78	16.08	**9.19**	**10.07**	10.86	25.34	**25.95**	32.85	13.99	**12.93**	24.97	**11.76**	**11.58**	11.81
HawkPR$^{+}_{\text{m}}$	10.01	10.39	19.21	21.01	18.06	6.72	12.80	13.21	**8.48**	**25.23**	30.04	**22.51**	**13.61**	14.58	16.66	17.55	14.52	**5.45**

The best performance is marked in **bold**.

Table S2: Model coefficients ($\mathcal{D}_{\text{conf}}$)

Covariate	coef	SE	pValue
Retail/recreation	0.2938	0.0022	0^{*}
Residential	−0.1061	0.0015	0^{*}
Transit stations	0.0633	0.0018	$1.6495 \times 10^{-277*}$
Grocery/pharmacy	−0.0451	0.0012	$8.3248 \times 10^{-313*}$
Parks	−0.0169	0.0006	$9.8958 \times 10^{-168*}$
Smokers percentage	−0.0807	0.0015	0^{*}
Heart disease mort.	−0.0549	0.0017	$2.4739 \times 10^{-221*}$
# hospitals	−0.0455	0.0016	$5.0219 \times 10^{-177*}$
Population Density	0.0329	0.0003	0^{*}
Population estimate	0.0295	0.0013	$1.3470 \times 10^{-117*}$
# ICU beds	0.0218	0.0010	$2.4117 \times 10^{-108*}$
Median age	−0.0149	0.0013	$4.5930 \times 10^{-030*}$
Diabetes percentage	−0.0064	0.0021	$1.8368 \times 10^{-002*}$

The first 5 covariates are mobility indices, followed by static demographic covariates.

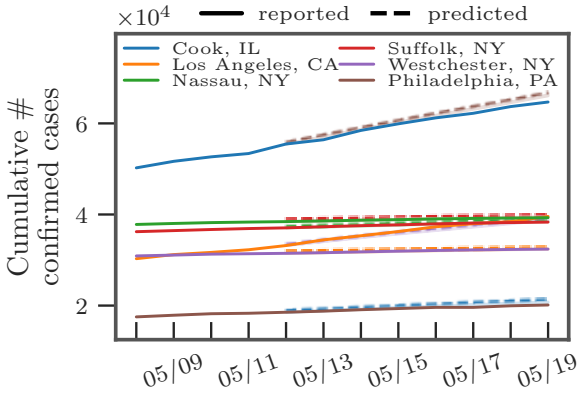


Figure S6: Top-6 counties in $Q^{\text{top}}_{10\%}$ of $\mathcal{D}_{\text{conf}}$

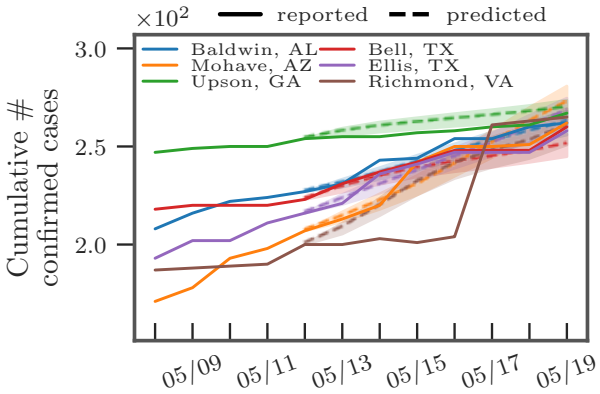


Figure S7: Top-6 counties in $Q^{25\%}_{50\%}$ of $\mathcal{D}_{\text{conf}}$

References

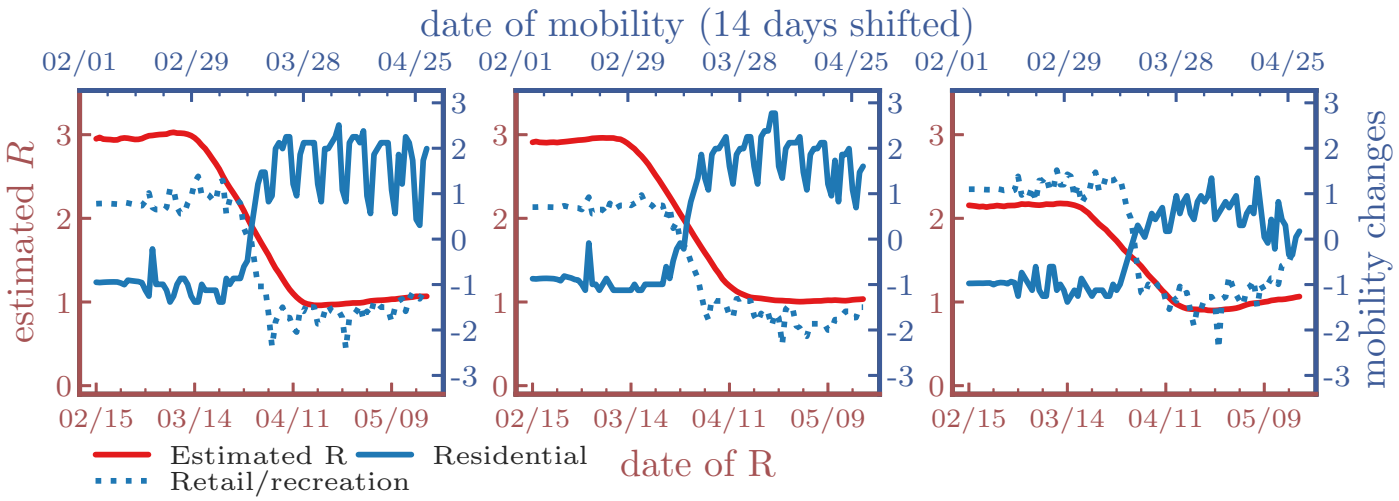
[S1] Simon Cauchemez, Pierre-Yves Boëlle, Christl A Donnelly, Neil M Ferguson, Guy Thomas, Gabriel M Leung, Anthony J Hedley, Roy M Anderson, and Alain-Jacques Valleron. Real-time estimates in early detection of sars. *Emerging infectious diseases*, 12(1):110, 2006.

[S2] Thomas Obadia, Romana Haneef, and Pierre-Yves Boëlle. The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC medical informatics and decision making*, 12(1):147, 2012.

[S3] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. Sir-hawkes: linking epidemic models and hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference*, pages 419–428, 2018.

[S4] Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516, 2004.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.06.20124149; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



(a) Cook, IL (1st in $Q^{top}_{10\%}$) (b) Los Angeles, CA (2nd in $Q^{top}_{10\%}$) (c) Baldwin, AL (1st in $Q^{25\%}_{50\%}$)

Figure S8: Estimated R of confirmed cases $\mathcal{D}_{\text{conf}}$ and lagged mobility changes ($\Delta = 14$ days).