Analyzing the Impacts of Public Policy on COVID-19 Transmission: A Case Study of the Role of Model and Dataset Selection Using Data from Indiana

George Mohler∗  Martin B. Short†  Frederic Schoenberg‡  Daniel Sledge§

Abstract

Dynamic estimation of the reproduction number of COVID-19 is important for assessing the impact of public health measures on virus transmission. State and local decisions about whether to relax or strengthen mitigation measures are being made in part based on whether the reproduction number, $R_t$, falls below the self-sustaining value of 1. Employing branching point process models and COVID-19 data from Indiana as a case study, we show that estimates of the current value of $R_t$, and whether it is above or below 1, depend critically on choices about data selection and model specification and estimation. In particular, we find a range of $R_t$ values from 0.47 to 1.20 as we vary the the type of estimator and input dataset. We present methods for model comparison and evaluation and then discuss the policy implications of our findings.

∗Department of Computer and Information Science, Indiana University - Purdue University Indianapolis
†Department of Mathematics, Georgia Institute of Technology
‡Department of Statistics, University of California Los Angeles
§Department of Political Science, University of Texas Arlington
1 Introduction

During the first months of 2020, nations responded to the COVID-19 pandemic by adopting a variety of public health interventions, including contact tracing, disease surveillance, and mandated social distancing [33, 34, 41]. Within the United States, the ongoing transmission of COVID-19 represents a serious public health threat and an ongoing strain on local, state, and federal resources. In the US, direct public health authority is largely vested in states and localities, with local decision-makers playing a critical role in shaping public health responses and in deploying resources during times of crisis. The federal government, meanwhile, seeks to play a coordinating role through the Centers for Disease Control and Prevention (CDC), funds research through agencies such as the National Institutes of Health (NIH), and helps shape the regulatory environment through the Food and Drug Administration (FDA) and other agencies [38, 16, 8, 39].

While this system has the potential to be highly responsive and adaptive, it is prone to problems including divergent outcomes across political jurisdictions and difficulty coordinating responses to emergent events. The lack of widespread testing in the early stages of transmission in the US forced policymakers to make decisions without high-quality data and foreclosed the possibility of effective disease surveillance, which might under different circumstances have proved a powerful public policy tool [20]. In the absence of testing and of pharmaceutical interventions such as a vaccine or anti-viral therapies, social distancing measures (including shelter-in-place orders and mandated closure of non-essential businesses) emerged as the primary tool at the disposal of state and local decision-makers [17].

Despite the very real public health benefits of such interventions, they have potentially large economic and social costs, which are distributed unevenly across society. As the pandemic has continued, public and political pressure to relax public health interventions has increased. Policymakers, as a result, confront a complex set of problems and high levels of uncertainty [4, 12, 19].

Given these circumstances, the swift assessment of how differing public health strategies impact the transmission of COVID-19 is critical to fostering flexible, focused, and data-driven policy-making [15]. One means of measuring the impact of public health interventions is through the effective reproduction number of a virus, $R_t$, e.g. the average number of individuals an infected person directly infects. When $R_t > 1$ and the majority of the
population is susceptible, as during the initial stages of the pandemic, the number of new daily infections exhibits exponential growth. However, when $R_t < 1$, the virus is no longer self-sustaining and will die out before most people in the population are exposed.

COVID-19’s initial reproduction number, $R_0$ (when the entire population is susceptible and policy interventions are not in place) has been estimated across several studies to be around 3.28 (1.4, 6.5) [26]. A study conducted with data from China through mid February estimated that, as a result of public health interventions, the effective time-varying reproduction number, $R(t)$, was reduced from 2 to 1 [44]. In Singapore, the impact of social distancing on $R$ was estimated to be between 78.2% and 99.3% [25]. Research on interventions in Europe observed that a combination of school closings, bans on mass gatherings, and other social distancing measures reduced $R(t)$ below 1 [14]. In the United States, state and local decisions about implementing, modifying, and relaxing social distancing measures have been informed in part based on whether the reproduction number, $R_t$, falls below the self-sustaining value of 1.

1.1 A need for model comparison and evaluation

While many of the forecasting models guiding policy-makers on COVID-19 capture uncertainty in parameter estimates, a large number of these analyses are presented via stand-alone models: model comparison, evaluation, and goodness-of-fit tests are often not presented. In addition, there are a variety of data sources available to researchers, ranging from data aggregation websites [11, 1] to local government data portals [2].

In this article, we show how estimates of the impact of policy interventions can vary depending on modeling and estimation choices, as well as the dataset that is used as an input. Understanding the role of model and dataset selection, we argue, is critical to high-quality policymaking during the COVID-19 pandemic and may help policymakers to more effectively prepare for and respond during the early stages of future disease outbreaks.

A variety of frameworks have been employed for modeling COVID-19 [5], including agent based models, compartmental models, and branching point process models. Given our expertise and their broad use in estimating the reproduction number of a virus [28, 13, 36, 9, 42], we focus here on the point process type of model. Within this framework, we compare three choices for modeling the impact of interventions on the transmission of COVID-19: 1) a
step function modeling an immediate impact on $R_t$ at key policy change dates that is employed in the highly cited paper [14], 2) a constant $R_t$ up until a key policy change date followed by exponential decay [21, 23], and 3) a non-parametric histogram estimator that adapts to changes in the reproduction number over time. These choices are not meant to be exhaustive, but rather illustrative of the variation in estimates that can arise based on differing model and data choices.

We apply these models to both daily case and mortality COVID-19 data in Indiana from three different sources: the widely-used COVID-19 data portal hosted at Johns Hopkins University [11] (abbreviated as “jhu” throughout), the Covid Tracking project [1] (abbreviated as “covidtracking” throughout), and the local Indiana data portal hosted at [2] (abbreviated as “in.gov” throughout). Differences in COVID-19 data collection and reporting standards represent a critical challenge for both researchers and policymakers. One issue we investigate is the reporting lag of new cases and deaths posted to the Indiana state department of health dashboard, often several days after the testing date (with reporting sometimes paused on the weekend). While data on the Indiana state health department website are retrospectively updated and corrected, data on aggregation websites like covidtracking and jhu are based on daily updates to cumulative counts and are not retrospectively corrected. Consequently, artificial peaks and valleys are present in the covidtracking and jhu count data. As a result, Indiana provides an excellent test case to investigate different choices in data processing and their impact on critical parameters in models for the spread of Covid-19.

In Section 2 we present the branching point process modeling framework and discuss how these models can be estimated from data. Then, in Section 3 we present our results when the models are applied to Indiana COVID-19 data. We show that estimates of the value of $R_t$ in Indiana, and whether it is above or below 1, depend on the model and dataset used for estimation. We also present several methods that can either be used to compare competing models or used to assess the goodness-of-fit of a particular model. We find a range of $R_t$ values from 0.47 to 1.20 as we vary the type of estimator and input dataset. In Section 4 we discuss the policy implications of our findings.
2 Methods

We consider a branching point process [29, 18] framework to estimate a time-varying reproduction number $R(t)$ [9, 42, 31, 37]. The conditional intensity (rate) of infections is modeled as

$$\lambda(t) = \mu + \sum_{t > t_i} R(t_i) w(t - t_i), \quad (1)$$

where $R(t)$ and $w(t)$ are the dynamic reproduction number and inter-infection time (also known as serial interval [42]) distribution respectively. We also include an exogeneous rate $\mu$ modeling imported infections.

The conditional intensity models the rate of new infections and is connected to the reproduction parameter $R(t)$ through the serial interval distribution $w(t)$. In particular, the expected number of new secondary infections on day $t$ caused by an infection on day $t_i$ is given by $R(t_i) w(t - t_i)$. The point process governed by Equation 1 can be viewed as an approximation to the common SIR (susceptible-infected-removed) model of infectious diseases during the initial phase of an epidemic when the total infections is small compared to the overall population size and $w(t)$ is specified to be exponential [35]. 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 [27]. One can also allow the reproduction parameter to vary parametrically as a function of the overall infection rate, due to mitigation efforts and herd immunity, as in [37].

We consider three competing models for $R(t)$:

1. A **step function** that changes at the Indiana stay-at-home order date $t_{sh}$, effective on March 24, 2020:

   $$R(t) = \begin{cases} R_0 & t \leq t_{sh} \\ R_1 & t > t_{sh} \end{cases} \quad (2)$$

   This is analogous to the step function used in [14] to assess the impact of public health interventions on COVID-19 in Europe.

2. An **exponential decay** [21, 23] after the stay-at-home date of the form:

   $$R(t) = \begin{cases} R_0 & t \leq t_{sh} \\ R_0 \exp(-ct) & t > t_{sh} \end{cases} \quad (3)$$

5
3. A **histogram estimator** that adapts to dynamic changes in $R_t$ over time

$$R(t) = \sum_{k=1}^{B} r_k \{t \in I_k\}. \tag{4}$$

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 $k$. In the remainder of the paper we use a bin-width of 1 week. We also merge the bins of the first 3 weeks due to the low number of events during that time period.

The branching process can be estimated via an expectation-maximization (EM) algorithm for maximum likelihood inference [40, 30, 24]. Given initial guesses for the model parameters and $r_k$, the EM algorithm iteratively updates the parameters and branching probabilities by alternating between the **E-step update**:

$$p_{ij} = R(t_j) w(t_i - t_j) / \lambda(t_i) \tag{5}$$

$$p_{ii} = \mu / \lambda(t_i) \tag{6}$$

and **M-step update**:

$$w(t) \sim \text{MLE}(\{t_i - t_j; p_{ij}\}) \tag{7}$$

$$\mu = \sum_i p_{ii} / T \tag{8}$$

$$r_k = \sum_{t_i > t_j} p_{ij} \{t_j \in I_k\} / N_k \tag{9}$$

where $T$ is the total length of the observation period, $N_k$ is the total number of new infections in interval $k$, and $w(t)$ is estimated via weighted MLE (maximum likelihood estimation) using the inter-event times as observations and branching probabilities as weights. The branching probability $p_{ij}$ corresponds to the probability that secondary infection $i$ was caused by infection $j$ in the dataset. While $w(t)$ can be estimated using a Weibull, Gamma, or log-normal distribution, we use a non-parametric histogram estimator with bin-width of 1 day to prevent model mis-specification.

Competing models can be compared using the Akaike Information Criterion [3], $AIC = 2p - 2\log(L)$. The AIC balances goodness of fit measured by
the log-likelihood, $\log(L)$, and over-parametrization by penalizing the number of parameters, $p$ (lower AIC is better). Alternatively, the goodness of fit of a branching process model can be assessed using residual analysis of rescaled event times [32],

$$
\tau_i = \int_0^{t_i} \lambda(t)dt
$$

(10)

The rescaled times are distributed according to a unit rate Poisson process if the model is correctly specified.

2.1 Using reported death data to estimate the reproduction number

We estimate the reproduction number of COVID-19 in Indiana from both new reported case data and mortality data. While it is standard to use reported infections to estimate the reproduction number, in the case of Indiana the daily rate of testing has steadily increased since the start of the pandemic [2]. In such a scenario, the reproduction number may be over-estimated as the rate of increase of reported infections is partly explained by an increase in testing. On the other hand, reported death counts also suffer from under-counting [43].

We note that under certain modeling assumptions, the reproduction number can be estimated from either reported infections or deaths, though the latter estimate will be lagged due to the time between a confirmed case and a subsequent fatality. For example, consider a Susceptible-Infected-Recovered-Death (SIRD) model governed by 

$$
\frac{dS}{dt} = -\beta SI/N, \quad \frac{dI}{dt} = \beta SI/N - \gamma I, \\
\frac{dR}{dt} = (1-c)\gamma I, \quad \text{and} \quad \frac{dD}{dt} = c\gamma I.
$$

Here it is assumed that some fraction, $c$, of those who are infected will subsequently correspond to a fatality. In Figure 1, we simulate such a model with $N = 10^6$, $\gamma = .2$, $\beta = .2$ and $c = .01$. We note that during the initial phase of the simulation when $S \approx N$, the exponential growth rates of new daily infections and new daily deaths are the same. When we apply the EM algorithm with the histogram estimator in Equations 5-9 to the simulated SIRD data, we obtain similar estimates for $R_t$, close to the true value of $R_0 = 2$, when using new infections ($\beta SI/N$) or new deaths ($c\gamma I$).

Similarly, in Figure 1 we also display a simulation of a Hawkes process of the form of Equation 1 with $\mu = 1$, $R_0 = 2$, and inter-infection time distribution $w(t)$ given by a Weibull(6,2). We assume a fraction $c = .1$ of infections
lead to a fatality where the infection-death inter-time distribution is given by a Weibull(5,3). Again, we observe that the growth rate of new infections and new deaths is the same. When we apply the EM algorithm with the histogram estimator in Equations 5-9 to the simulated Hawkes process data, we obtain similar estimates for $R_t$, close to the true value of $R_0 = 2$, when using new infections or new deaths.

However, there are certainly scenarios where the estimated reproduction number will be different when infections or deaths are analyzed. For example, transmission may not be a homogeneous process across the population, and instead sub-populations with higher (or lower) case-fatality rates could have higher (or lower) contact rates. There also could be temporal trends, for example the case-fatality rate could go down over time as the quality of medical interventions improves. Including these types of effects in estimates of the dynamic reproduction number are outside of the scope of this paper.

3 Results

We apply the estimation procedure outlined above to Indiana COVID-19 case and mortality daily counts (new cases rather than cumulative) from March 5, 2020 to April 26, 2020. While school and business closings occurred on March 16, 2020, there is limited case data available before this date and we therefore assume $R_0$ is constant across all models up until the stay-at-home order on March 24, 2020.

We present estimated $R_t$ curves in Figure 2, estimated intensities $\lambda_t$ (Figure 3) and inter-infection time distributions $\omega(t)$ (Figure 4), and in Table 1 we present the corresponding AIC values for the different models (step, exponential, histogram), different data sources (Johns Hopkins [11], Covid Tracking Project [1], State of Indiana [2]), and different data types (case counts or mortality).

The first observation of note is that the AIC values are lower for all models using the local in.gov data rather than data from aggregation websites. In the case of Indiana, new cases and deaths on a given date are routinely reported to the state department of health several days after the fact (with reporting often paused on the weekend). While the data on Indiana’s state health website are retrospectively updated and corrected, the aggregation website data are based on daily updates to cumulative counts. Importantly, these sites do not go back and correct historical cumulative count data. Consequently,
Figure 1: New infections and new deaths in a SIRD model simulation with $N = 10^6$, $\gamma = .2$, $\beta = .2$ and $c = .01$ (top left) along with estimated reproduction number $R_t$ (bottom left). New infections and new deaths in a Hawkes process simulation with $\mu = 1$, $R_0 = 2$, inter-infection time distribution Weibull(6,2), infection-death time distribution Weibull(5,3), and case fatality rate .1 (top right) and estimated reproduction number $R_t$ using EM algorithm with histogram estimator (bottom right).

artificial peaks and valleys are present in the covidtracking and jhu estimates (see Figure 3). Our findings are consistent with recent recommendations to use local data whenever possible [22].

Variation in estimates of $R_t$ that arise from using either case or death counts is higher earlier in the Indiana epidemic. Estimates of $R_t$ are initially as high as 5 when using case data but are between 2 and 4 using death counts. The high value of $R_t$ early on for cases may be due to the initial lack of testing followed by a rapid growth in testing over a several week period.

We find a larger variation in $R_t$ across models than across data sources. For example, when applied to jhu case data the models provide final $R_t$ values (at our study’s end date of April 26th) of 1.20 (histogram), 0.83 (step) and 0.66 (exponential). Across the datasets, the estimated value of $R_t$ at the final
Figure 2: Fitted $R_t$ curves for histogram estimator (blue), step function (red) and exponential decay (yellow) applied to different Indiana COVID-19 data types and data sources.

Time tends to be higher for the histogram as it adapts to changes in reproduction after Indiana’s March 24th stay-at-home order. While $R_t$ fell below 1 according to the histogram several weeks after the order, it later rose back above 1 (possibly due to lack of adherence to social distancing or the emergence of new clusters in counties outside of the Indianapolis Metropolitan area). The histogram estimator also consistently has the lowest AIC values of the competing models because of its ability to adapt to local changes in time.

While the AIC is useful for comparing competing models, goodness-of-fit of a point process model can be evaluated using residual analysis. In Figure 3 we plot point process intensities fit to Indiana COVID-19 cases and deaths per
<table>
<thead>
<tr>
<th>model</th>
<th>dataset</th>
<th>datatype</th>
<th>log-likelihood</th>
<th>AIC</th>
<th>$R_{final}$ (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hist</td>
<td>covidtracking</td>
<td>deaths</td>
<td>1965.80</td>
<td><strong>-3921.60</strong></td>
<td>1.17 (.40)</td>
</tr>
<tr>
<td>step</td>
<td>covidtracking</td>
<td>deaths</td>
<td>1824.10</td>
<td>-3644.20</td>
<td>0.68 (.14)</td>
</tr>
<tr>
<td>exp</td>
<td>covidtracking</td>
<td>deaths</td>
<td>1947.40</td>
<td>-3890.90</td>
<td>0.49 (.12)</td>
</tr>
<tr>
<td>hist</td>
<td>in.gov</td>
<td>deaths</td>
<td>1980.10</td>
<td><strong>-3950.20</strong></td>
<td>0.97 (.37)</td>
</tr>
<tr>
<td>step</td>
<td>in.gov</td>
<td>deaths</td>
<td>1951.50</td>
<td>-3899.00</td>
<td>0.95 (.16)</td>
</tr>
<tr>
<td>exp</td>
<td>in.gov</td>
<td>deaths</td>
<td>1967.50</td>
<td>-3931.00</td>
<td>0.59 (.13)</td>
</tr>
<tr>
<td>hist</td>
<td>jhu</td>
<td>deaths</td>
<td>1954.60</td>
<td><strong>-3899.30</strong></td>
<td>1.12 (.39)</td>
</tr>
<tr>
<td>step</td>
<td>jhu</td>
<td>deaths</td>
<td>1877.30</td>
<td>-3750.60</td>
<td>0.85 (.16)</td>
</tr>
<tr>
<td>exp</td>
<td>jhu</td>
<td>deaths</td>
<td>1944.60</td>
<td>-3885.20</td>
<td>0.47 (.11)</td>
</tr>
<tr>
<td>hist</td>
<td>covidtracking</td>
<td>cases</td>
<td>81521.00</td>
<td><strong>-163030.00</strong></td>
<td>1.19 (.41)</td>
</tr>
<tr>
<td>step</td>
<td>covidtracking</td>
<td>cases</td>
<td>80809.00</td>
<td>-161610.00</td>
<td>0.83 (.15)</td>
</tr>
<tr>
<td>exp</td>
<td>covidtracking</td>
<td>cases</td>
<td>80838.00</td>
<td>-161670.00</td>
<td>0.67 (.14)</td>
</tr>
<tr>
<td>hist</td>
<td>in.gov</td>
<td>cases</td>
<td>81534.00</td>
<td><strong>-163060.00</strong></td>
<td>1.19 (.41)</td>
</tr>
<tr>
<td>step</td>
<td>in.gov</td>
<td>cases</td>
<td>80727.00</td>
<td>-161450.00</td>
<td>0.83 (.15)</td>
</tr>
<tr>
<td>exp</td>
<td>in.gov</td>
<td>cases</td>
<td>80853.00</td>
<td>-161700.00</td>
<td>0.67 (.14)</td>
</tr>
<tr>
<td>hist</td>
<td>jhu</td>
<td>cases</td>
<td>81516.00</td>
<td><strong>-163020.00</strong></td>
<td>1.20 (.41)</td>
</tr>
<tr>
<td>step</td>
<td>jhu</td>
<td>cases</td>
<td>80741.00</td>
<td>-161480.00</td>
<td>0.83 (.15)</td>
</tr>
<tr>
<td>exp</td>
<td>jhu</td>
<td>cases</td>
<td>80793.00</td>
<td>-161580.00</td>
<td>0.66 (.14)</td>
</tr>
</tbody>
</table>

Table 1: Comparison of dynamic $R_t$ model fits across estimator type, data sources and data types for Indiana COVID-19 data. Number of bins is $B = 5$ for the histogram estimator. The other two models each have $p = 2$ parameters for the dynamic reproduction number.
day from March 5, 2020 to April 26, 2020. We again use a histogram estimate for the inter-infection time distribution $\omega(t)$ which we plot in Figure 4. In Figure 5, we plot the normalized cumulative distribution of rescaled event times and compare them to confidence bounds of the cumulative distribution for a unit rate Poisson process. We find that the estimated intensity using the histogram and exponential decay for $R_t$ provides a good fit to Indiana new deaths per day, whereas the intensity that uses a step function for $R_t$ under-estimates the empirical death rate. This is in comparison to Figure 3, where all of the intensities appear to give plausible fits to the data based upon visual inspection.
4 Discussion

Over the course of the COVID-19 pandemic, state and local policymakers have employed mathematical models and projections to inform decisions about implementing, relaxing, and reimposing a variety of public health interventions. Policymakers and public officials have at times focused on the predictions of individual models without making comparisons to other models that might yield different results. During the first months of the pandemic, notably, top presidential advisors regularly referenced the Institute for Health Metrics and Evaluation (IHME) model [10], which generated more optimistic projections than several high-profile alternatives [7].

It is critical that policymakers and political leaders consider the role of model and data selection when attempting to respond to an outbreak such as the COVID-19 pandemic. Focusing on one model and data source, decision-makers might easily draw misleading conclusions about the impacts of pub-
lic health interventions on the transmission of COVID-19. Confronted with the social and economic costs that flow from stringent social distancing measures, political leaders may be drawn towards the most optimistic projections. Here, however, we show that conclusions about when COVID-19 transmission might peak depend on the type of model that is used and on the data that is used.

In the present study, we find that a histogram estimator for dynamic $R(t)$ provides the most accurate fit to Indiana COVID-19 data during the early stages of the COVID-19 pandemic, with the exponential model also providing a plausible fit according to residual analysis. We emphasize, however, that these results may not generalize to other datasets and time periods. For other data and for different time periods, model comparison and goodness-of-fit analysis should again be applied. We also note that while this analysis was limited to temporal data, in the future, access to detailed spatial-temporal data on individual cases may enable more precise estimation of spatial-temporal triggering in these types of models.

During the early stages of the COVID-19 pandemic, it was important that forecasting models be rapidly developed in order to inform decision-making. Now that these models have been implemented and knowledge of COVID-19 transmission dynamics has matured, research is needed on the trade-offs between competing model and data choices. Here, we showed that, when varying the estimator of dynamic $R_t$ between three simple choices, along with 3 different data sources, we get dramatically different answers to the
question of whether public health interventions in Indiana during Spring 2020 reduced the reproduction number to below one.

As the COVID-19 pandemic has progressed, decision-makers have faced a myriad of new challenges and cross-pressures. Economic concerns, social costs, and public fatigue with health interventions have interacted with misinformation and partisan positioning to help foster an environment in which key political leaders have publicly expressed comfort with ongoing community spread. As COVID-19’s reproduction number declined during the late spring and early summer of 2020, states and localities began relaxing mandated social distancing measures. Often framed in terms of the need for increased economic activity, these policy changes had a substantial impact on individual behavior and on public perceptions of the threat of ongoing transmission. In many areas, relaxation was followed by periods where $R_t$ again rose above 1, resulting in new cases and deaths.

In several high-profile cases, state Governors implemented a new round of public health interventions. In California and Texas, for instance, Governors allowed bars to reopen and then moved to close them again as they became associated with increases in COVID-19 transmission. Similar patterns were observed during the 1918 Influenza pandemic [6]. In Indiana, meanwhile, Governor Eric Holcomb announced a new set of restrictions in November 2020 as case counts and hospitalizations began to soar. As of November 15, 2020, the histogram model estimate for the reproduction number was $R_t = 1.3$ in Indiana.

Moving forward, modelling will continue to play a critical role in informing policy decisions. The results presented here emphasize both the complex nature of the pandemic and the critical importance of acknowledging that findings about the impacts of public policy interventions will vary as a result of model and data selection. In making decisions about public health measures and in gauging the impact of various interventions, policy-makers should be careful to consider model specification, goodness-of-fit, and the sensitivity of models to the choice of input data.

**Acknowledgements**

This research was supported by NSF grants SCC-1737585, ATD-1737996 and ATD-1737925.
References


