Stay Home Save Lives: A Machine Learning Approach to Causal Inference to Evaluate Impact of Social Distancing in the US

Syed Muhammad Ishraque Osman
Long Island University
Nazmus Sakib (nazmus.sakib@ttu.edu)
Texas Tech University https://orcid.org/0000-0003-4930-2862

Research Article

Keywords: Covid-19, Corona Virus, 2019-nCoV, Social Distancing, Pandemic, Policy Evaluation, Machine Learning

Posted Date: September 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-28199/v2

Abstract

Introduction: This study presents a machine learning based evaluation of the social distancing measures implemented in the US states.

Objectives: Although there are a few studies that provide estimations of the impact of COVID-19 pandemic, there is a need for an actual policy evaluation of the already implemented social distancing measures. This paper presents an evaluation of the social distancing measures implemented by the US states.

Methods: This research uses a machine learning based Generalized Synthetic Control Method. In doing so, it considers the US states that adopted early social distancing approaches as the treatment group and the states that adopted social distancing much later as the control group and it has controlled for state and time fixed effects, to cancel out the possible selection bias and endogeneity. **Results**: The results show that the first round of social distancing in the US is associated with lower COVID-19 infection growth rate (by -167%) when compared to the no policy intervention counterfactual.

Conclusions: The findings from this policy evaluation establishes a robust scientific basis of the efficacy of social distancing measures on slowing down the contagion of a pandemic.

Introduction

Although few post-COVID-19 studies have attempted to provide an estimate of possible US gains (e.g. Greenstone and Nigam, 2020) and simulation model of COVID-19's spread and mortality impacts in the US (e.g. Ferguson et al. 2020), there is a need for a causal policy evaluation of the already implemented social distancing to measure what have we achieved so far compared to a no intervention counterfactual. Using Generalized Synthetic Control Method (GSCM) developed by Xu (2017), considering the US states that adopted early social distancing approaches as the treatment group and the states that adopted social distancing much later as the control group and controlling for state and time fixed effects (to cancel out the selection bias and endogeneity), this paper finds that social distancing is associated with lower COVID-19 infection growth rate (by 167%). GSCM calculates weights for the untreated (control) units in order to create a synthetic twin of the treatment unit in the pre-treatment period, it uses an interactive fixed effect model (discussed later) in this re-weighting phase (Xu, 2017). As in case of predictive machine learning, GSCM then makes out-of-sample (post-treatment period) prediction using the calculated weights (based on the interactive fixed effect model) in order to create a counterfactual for the treatment unit. In this sense, this method is in the same spirit of machine learning predictive modeling.

Since COVID-19 is the only major pandemic in our recent memory, the pre-COVID-19 literature is reasonably inadequate regarding policy evaluations of social distancing. Reluga (2010) finds that social distancing is especially helpful when the implementation is comparatively less expensive and can defer the outbreak until a vaccine becomes broadly accessible. Similarly, although in a slightly different

context, Glass et al. (2006) provides a simulation of targeted social distancing to mitigate flu pandemic in a small town in the United States. For a flu epidemic as infectious as 1957–58 Asian flu (which had close to 50% infection rate), the simulation shows that shutting down schools and keeping young population at home decreased the infection growth by >90%.

Few post-COVID-19 studies have tried to study the pandemic from different perspectives. Using the Ferguson et al. (2020) simulation model of COVID-19's proliferation and mortality in the US, Greenstone and Nigam (2020) estimates that three to four months of modest social distancing starting in the last week of March 2020 would save 1.7 million lives by the beginning of October. Using the US Government's value of a statistical life, Greenstone and Nigam (2020) find that the monetized values of preventing mortality by social distancing are about $8 trillion at aggregate level or $60,000 per household.

Although these are impressive estimates of the impact of social distancing, however, to the best of our knowledge, there has not been any study conducted yet to evaluate the actual impact of social distancing on the growth rate of infection comparing to the no policy counterfactuals. This study is necessary because, not only dozens of major US cities and state capitals were packed with anti-lockdown protests but also studies (e.g. Hatzius et al. 2020) have projected a bleak future for the US job market due to the historically unprecedented US unemployment claims after the implementation of social distancing measures. Analyzing the serious negative economic consequences of social distancing from China, South Korea, and Singapore, question is raised as to whether social distancing is worth the economic cost (Hilsenrath and Armour 2020; Thunstrom et al. 2020). This paper intervenes in this juncture to provide a machine learning informed policy evaluation of whether the already implemented social distancing measures had any impact on restraining the infection growth rate of the COVID-19 virus.

This paper proceeds as follows: section 2 presents the data sources and empirical strategy of the paper, which justifies the method and assignment of treatment and control group; section 3 presents the results and visualization of social distancing measures; and finally, section 4 presents the conclusion.

Data And Method

2.1 Data

The dataset and the interactive dashboard are created and maintained by a group of researchers at Johns Hopkins University[1] (Dong et al, 2020). This dataset reports infection and death cases of Covid-19 disease in real time. It reports cases at the city level in the US, which is later aggregated at state levels. Our data starts reporting from 2020-01-21 and ends on 2020-04-13. There are a total 4536 observations. Our outcome variable is growth rate of daily cases or growth rate of infection and the right-hand side indicator variable is social distancing. We collect cumulative number of confirmed cases and death cases at the Federal Information Processing System (FIPS) code level from Johns Hopkins University (JHU) Center for Systems Science and Engineering[i]. JHU posts the data in a wide time series format. We follow preliminary lines of the R (open source language and environment) code posted by Tim Churches

(Mar 05, 2010) in a blog to extract the data from JHU and put it into a panel format[ii]. The rest of the data cleaning and manipulation tasks are done in R and the data is made ready for feeding into the model of our choice- Generalized Synthetic Control Method. Our outcome variable is growth rate of confirmed cases, which we estimate from the number of daily confirmed cases. We use the data up to April 13. We do not go beyond 13 April as the independent effect of the social distancing was becoming harder to identify as people started learning more about the benefits of social distancing and control states were also catching up in implementing social distancing. Later, states ordered 'lockdown' or shelter in place orders, which made it extremely hard to identify treatment from the control group. Thus, considering the immediate time frame following the social distancing measures makes sense.

2.2 Treatment and Control Status

Glanz et al. (2020) published an article in New York Times titled "Where America Didn't Stay Home Even as the Virus Spread." Using location data from the data intelligence firm Cuebiq, the authors provide a map of the United States showing, following the implementation of social distancing, "… when the average distance travelled first fell below 2 miles[iii]". This is our functional definition of social distancing that we are using for this paper. Glanz et al. (2020) divide the places into 5 categories in terms of the date on which the mean distance travelled first dropped under two miles. These dates are March 16, March 19, March 24, and March 26. We considered the states without social distancing until March 26 in the control group and other states, which implemented social distancing earlier than March 26, in the treatment group. There are some states that had some counties that did not show social distancing until March 26, but majority of the counties did. We put even these states in the control pool in order to get a clean and conservative treatment group. The control group for this study includes the following states: Idaho, Wyoming, Utah, Arizona, New Mexico, Oklahoma, Arkansas, Alabama, Louisiana, Mississippi, South Carolina, Tennessee, Virginia, West Virginia, Kentucky, Kansas, North Dakota, South Dakota, Nebraska, Missouri, Iowa, Illinois, Wisconsin, Indiana, North Carolina, Florida, Maryland, Pennsylvania, Vermont, and Texas. While the treatment group consists of California, Colorado, Connecticut, Delaware, District of Columbia, Georgia, Guam, Hawaii, Maine, Massachusetts, Michigan, Minnesota, Montana, Nevada, New Hampshire, New Jersey, New York, Northern, Mariana Islands, Ohio, Oregon, Puerto Rico, Rhode Island, Virgin Islands, and Washington. The event/treatment date is Mar 26, 2020. Figure 1 shows the average outcome of the treatment and the created counterfactual of the treatment in pre and post treatment periods.

States vary in terms of the first day of infection tested. Assuming the novel corona virus has a life cycle independent of the states' first reported infection date, different states may show slower or faster growth rate of infection depending on the first day of infection. Also, this first day may significantly influence how seriously people take social distancing as the number of cases are low at the beginning and later the cases grow exponentially. Thus, our treatment status can be endogenous to the first day of infection,

which is a proxy for the life cycle of the virus. We control for the first day of infection in our model in order to break the dependency between growth rate of infection and treatment status.

2.3 Summary Statistics

In this section we provided some descriptive statistics and some visualization. First, we looked at the average growth rate of infection for treated and control groups, before and after the intervention. In the pre-treatment period, for treatment group, average growth rate of infection was 33%, while for the control group it was 19%. In the post-treatment period, for treatment group, average growth rate of infection was 88%, while for the control group it was 75%. Figure 2 presents average growth rate of infection by period and groups for our sample time frame.

Here, it should be noted that treatment states show higher average growth rate before as well as after the treatment time of March 26. Thus, their infection growth rate is level up from the control states always in our period of data. That makes our scientific analysis more relevant as well as interesting because we want to see if, even after social distancing, the treatment states continued to show reasonably high growth rate of infection relative to the control states. This is because, in the absence of a proper causal inference technique, the descriptive statistics and/or graphics could deceive the reader by implying that social distancing has had no effect in curbing the infection rates. We want to see if a sophisticated model still supports or nullify what we are seeing with bare eyes. We also show the variance of growth rate between the treated and the control states before and after the social distancing in figure 3.

In the pre-treatment period, for treatment group, the variance of infection was 6.57 SD, while for the control group it was 2.62 SD. In the post-treatment period, for treatment group, the variance of infection was 9.35 SD, while for the control group it was 7.41 SD. It is evident that the treated states vary much more from each other than that of the control states in terms of the growth rate of infection. We believe the smaller variation amid the control group makes them a better donor pool for creating the synthetic counterfactual for our treatment group, as far as the method we are using in this paper is concerned.

2.3 Method

The classic model generally used to understand the impact of an event on an outcome of interest is Difference-in-difference (DID). This is the most popularly used model to answer causal questions. A major limitation of this model is it heavily depends on the assumption that the treatment and control units' mean outcome follow parallel time trends in the pre-treatment period. The same parallel trend is assumed in the post-treatment period in the absence of the treatment. Also, in order to identify the treatment effect, we need exogeneity of the treatment event. In other words, the treatment status cannot be determined by any factor(s) that also impacts the outcome variable of interest. Sometimes, a workable assumption is conditional independence (also known as 'selection-on-variables'), which states that if we can identify the variables the treatment is endogenous to, we can control for those variables in the model. In this way, we can break the dependency between the treatment status and outcome of interest created by those variables.

Another method that has gained momentum is the Synthetic Control Method (SCM) proposed by Abadie, Diamond and Hainmueller (2010). SCM relaxes the parallel trend assumption in DID and essentially computes a “synthetic twin” to the treatment unit by reweighting the control units using the pre-treatment data on outcome and other covariates. In our opinion, SCM uses a machine learning approach to create the counterfactual for the treatment unit in the post-treatment period. It calculates the weights for each control unit using the pre-treatment period and then plug in those weights in the post-treatment control unit data to create the counterfactual for the treated unit. One caveat is SCM is applicable for a single treatment unit. In this paper, we use a more sophisticated approach of Generalized Synthetic Control method (GSCM) proposed by Xu (2017) that combines SCM with another approach to model time-varying unit specific factors, known as Interactive Fixed Effect model. These time varying unit specific factors are not observed in the data, but yet taken care of. GSCM uses interactive fixed effect model on the control unit data to get the latent unobserved factors (time-varying) and uses these factors to estimate factor loadings (unit-specific intercepts) for the treated unit (Xu, 2017). This implies GSCM even relaxes the assumption of selection-on-variables to a great extent and permits the treatment status to be endogenous to unknown time-varying and unit-specific covariates.

We exploit this advantage in our paper as our treatment assignment of social distancing is not random. GSCM also allows for multiple treatment units, which is also the case in our paper. Xu and Liu (2020) shows implementation of the model in R[i]. We follow the codes in order to implement the model on the data for this paper. Also, we look at the matching quality between the treatment average and the synthetic twin in the pre-treatment period by eyeballing if their paths overlap. Any difference in the post-treatment period can be attributed to the effect of social distancing.

Results And Discussion

Table 1 shows the results. The outcome variable is Growth Rate of Confirmed Cases. In Table 1 column (1), we have state and as well as day fixed effects and we also control for the first day of infection. This is the main model of our interest. We find a statistically significant average treatment effect on the treated (ATT) of -1.67 (-167%). This means, social distancing is associated with lower covid-19 infection growth rate (by 167%). We are not claiming this is the exact impact of social distancing as no method can deterministically tell us what would have happened if the treatment states did not receive the treatment (we don’t have that parallel world). That said, this number unambiguously demonstrates the direction of the impact of social distancing.

Figure 4 shows how ATT evolves over time from the pre- to post-social-distancing era. We eyeball the match quality between the treatment and its counterfactual in the pre-social-distancing period. Their paths don’t perfectly overlap, but they follow each other very closely. It shows that we have a good match in the pre-treatment period, and we can take the treatment effect in the post-treatment period seriously. In our opinion, this is a striking result showing how effective social distancing can be in reducing the growth rate of the infection during a pandemic.

Columns 2 and 3 show the treatment effect with exclusive state and day fixed effects, respectively. The ATT with state fixed effect only is -37% and significant at 5% level. The ATT with day fixed effect only is -17% and not significant. Again, our main result is the full model with both state and day fixed effects in column 1 and we report columns 2 and 3 for comparison purposes. Figure 5 shows treatment and estimated counterfactual average growth rate in the pre- and post- social distancing era. The lower panel shows ATT time dynamics.

Table 1: Generalized Synthetic Control Results: Outcome variable is the Growth Rate of Confirmed cases

	(1)	(2)	(3)
Effect (Average Treatment Effect on Treated)[1]	-1.67*	-0.37**	-0.17
	(0.074)	(0.026)	(0.100)
First Day of infection	-0.042	-0.27***	-0.042
	(0.734)	(0.004)	(0.778)
State fixed effect	Yes	Yes	-
Day fixed effect	Yes	-	Yes
Number of Treatment states	24	24	24
Number of Control states	30	30	30

Notes: p-values are given in parentheses and calculated by a parametric bootstrap (1000) procedure. See Xu (2017, p 64) for the detail estimation procedure. ***, **, and * implies 1%, 5% and 10% statistical significance levels, respectively.

Please note that GSCM creates synthetic twin for each treatment state using information on the control pool of states. More precisely, GSCM channels control pool information through an interactive fixed effect model of which state and day fixed effects are special cases (Xu, 2017). Our treatment (social distancing by March 26) is not random and can very well be correlated with unknown and/or unobserved state and

time specific miscellaneousness. Thus, our main result is stated by both way fixed effect model in column 1, as it takes care of the correlation between treatment assignment and unknown factors in a more comprehensive way.

[1] In our first draft, we mistakenly included the cruise ship Diamond Princes—that reported the early cases of Covid-19 in the very early days of the pandemic—as a Province_State, as the data we were using treated it in that way. Here, we present the corrected results without Diamond Princes and we did not see any change in terms of significance and only small a decrease in the effect—making it a 167% decrease in the infection rate. Number of treatment states now becomes 24.

Conclusion

We investigate whether social distancing measures in the US worked when compared to a no policy counterfactual. In our case, the treatment status is not exogenous and possibly corelated with so many other factors. Thus, traditional DID like approaches of causal inference would not help to identify the impact of social distancing on infection growth rate. Thus, we used the Generalized Synthetic Control Method, which permits treatment status to be corelated with unknown factors that vary over time and across states and estimates the counterfactual for the treatment states by doing an out-of-sample (post-social distancing period) prediction (Xu, 2017). This is in the same spirit of predictive machine learning modeling. We find that social distancing is associated with lower covid-19 infection growth rate (by 167%) when compared to the no intervention counterfactual.

Declarations

Acknowledgement/Funding Information: No person or entity has provided funding for this research.

Conflicts of Interest: There is no conflict of interest known to the authors.

Ethics Statement: This research uses publicly available secondary data from Johns Hopkins University and no ethics approval was necessary for that purpose.

References

1. Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, *105*(490), 493-505.
2. Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57-76.

3. Glanz et al. (2020, April 02). Where America Didn't Stay Home Even as the Virus Spread. Retrieved from https://www.nytimes.com/interactive/2020/04/02/us/coronavirus-social-distancing.html
4. Churches, Tim. (2020, March 05). COVID-19 epidemiology with R. *R Views.* https://rviews.rstudio.com/2020/03/05/covid-19-epidemiology-with-r/
5. Xu, Y., & Liu, Licheng (2020, March 06). gsynth: Generalized Synthetic Control Method. https://yiqingxu.org/software/gsynth/gsynth_examples.html
6. Liu, Licheng and Yiqing Xu (2018). "panelView: an R package of visualizing panel data." Available at http://yiqingxu.org/software/panelView/panelView.html.
7. Data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). (Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). https://github.com/CSSEGISandData/COVID-19)
8. Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time." *The Lancet infectious diseases* 20, no. 5 (2020): 533-534.
9. Glanz, James, et al. *"Where America Didn't Stay Home Even as the Virus Spread." The New York Times, The New York Times, 2 Apr. 2020, www.nytimes.com/interactive/2020/04/02/us/coronavirus-social-distancing.html.*
10. Greenstone, Michael, and Vishan Nigam. "Does Social Distancing Matter?." *University of Chicago, Becker Friedman Institute for Economics Working Paper* 2020-26 (2020).
11. Reluga, Timothy C. "Game theory of social distancing in response to an epidemic." *PLoS computational biology* 6, no. 5 (2010).
12. Hatzius, Jan, et al. March 20, 2020. "US Daily: A Sudden Stop for the US Economy," *Goldman Sachs*.
13. Hilsenrath, Jon, and Armour, Stephanie. March 23, 2020. "As Economic Toll Mounts, Nation Ponders Trade-Offs," *The Wall Street Journal.*
14. Thunstrom, Linda and Newbold, Stephen and Finnoff, David and Ashworth, Madison and Shogren, Jason F., The Benefits and Costs of Using Social Distancing to Flatten the Curve for COVID-19 (April 14, 2020). Forthcoming Journal of Benefit-Cost Analysis. Available at SSRN: https://ssrn.com/abstract=3561934 or http://dx.doi.org/10.2139/ssrn.3561934
15. Glass, R. J., Glass, L. M., Beyeler, W. E., & Min, H. J. (2006). Targeted social distancing design for pandemic influenza. *Emerging infectious diseases*, *12*(11), 1671–1681. https://doi.org/10.3201/eid1211.060255

Figures

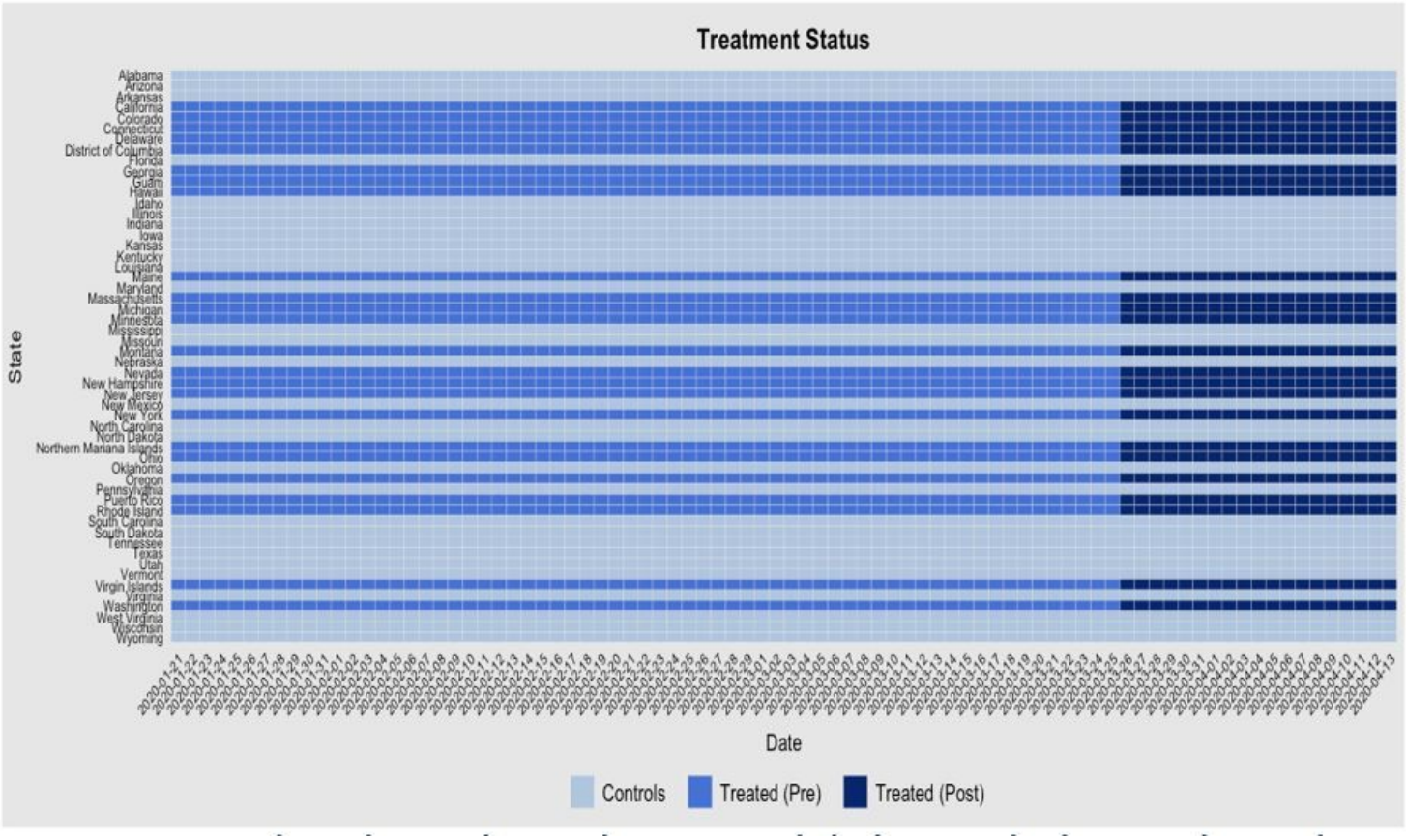

Figure 1

Treatment and Control states in the pre- and post-event period. This figure is produced using ‘panelView’ package in R , developed by Xu & Liu (2018)

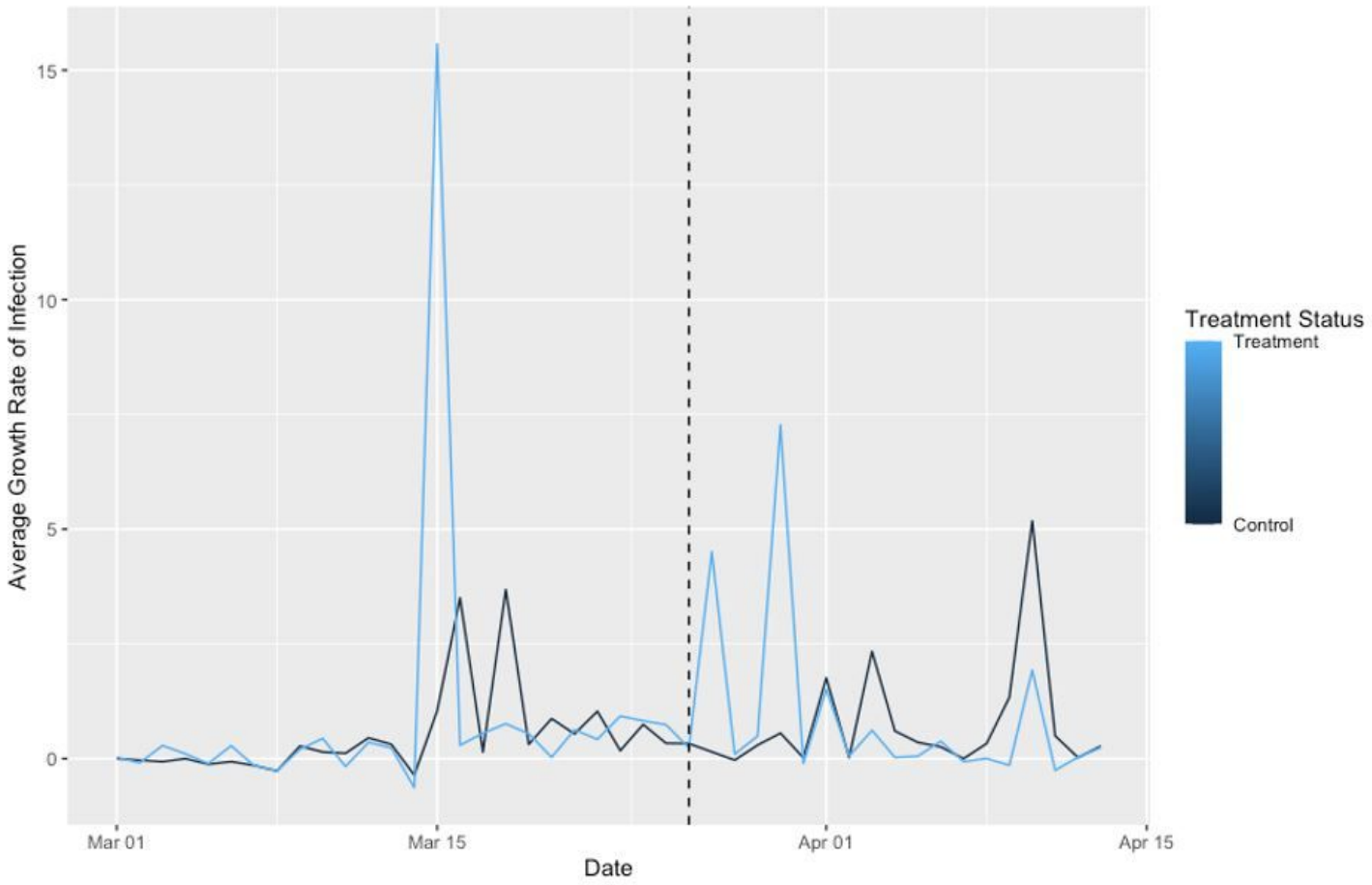

Figure 2

Average Growth Rate of Infections in the Sample Period by the Treatment Status

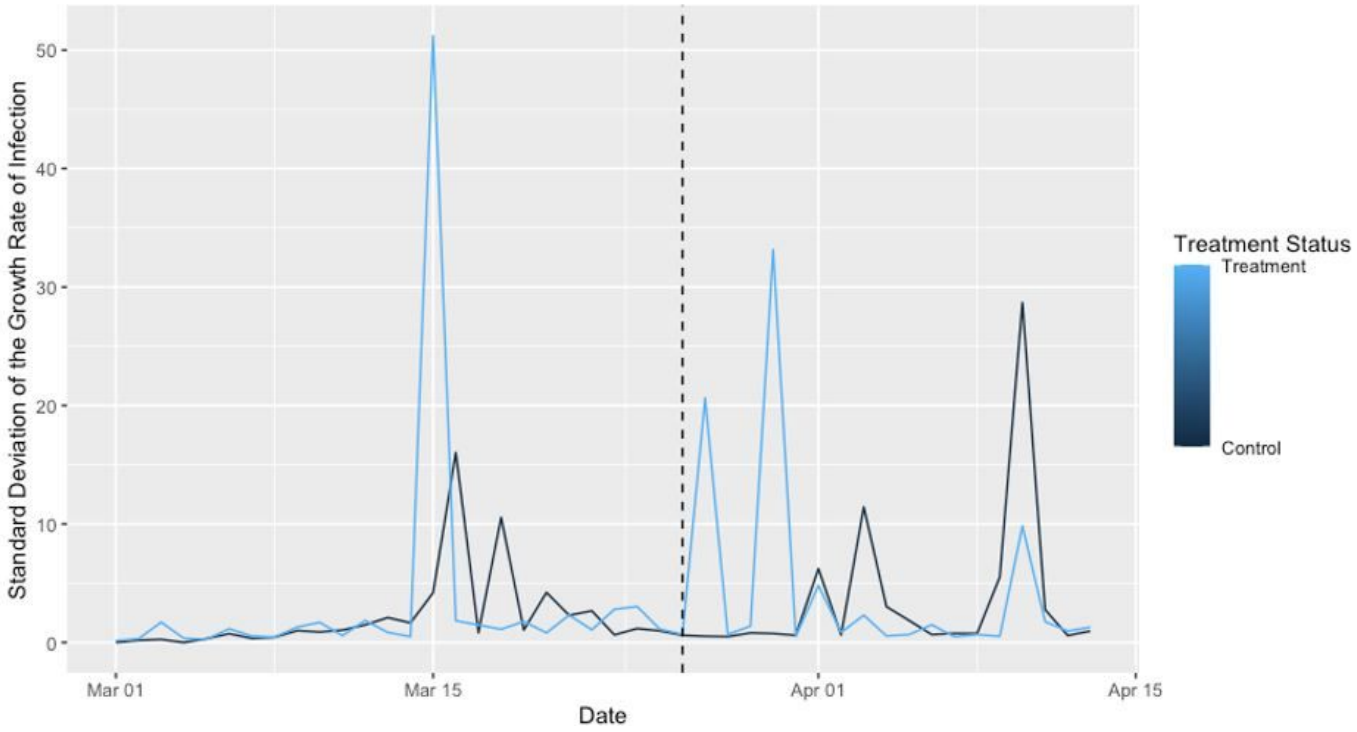

Figure 3

The Variance of Growth Rate between the Treated and the Control States before and after the Social Distancing

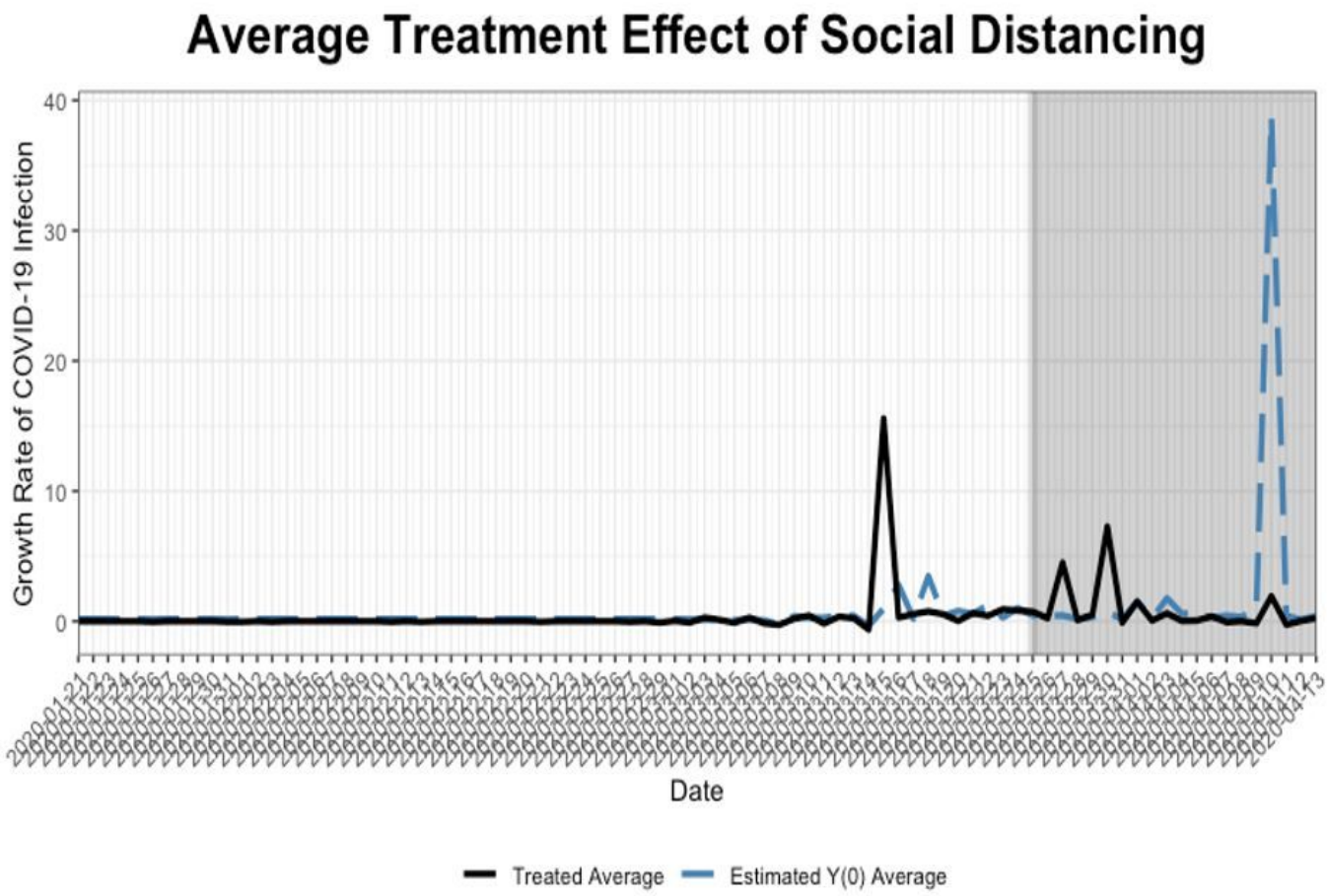

Figure 4

Average Treatment Effect of Social Distancing on Treated States

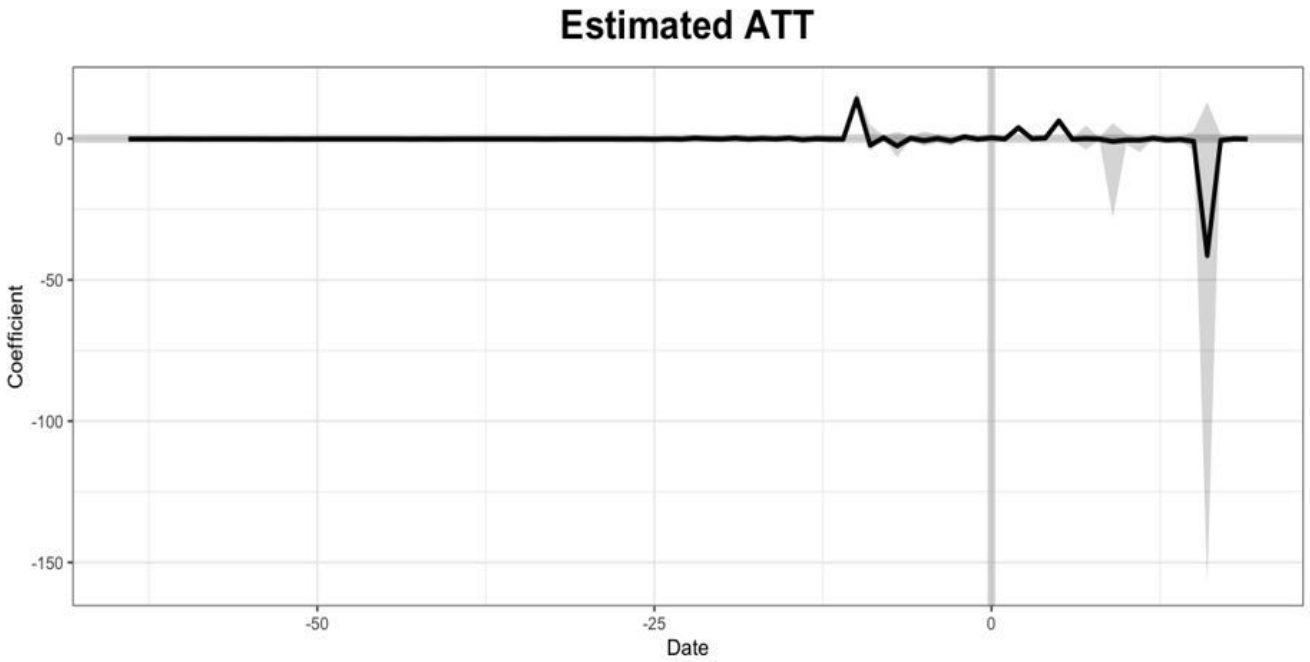

Figure 5

Upper panel: Treatment and estimated counterfactual average growth rate in the pre- and post- social distancing era. Lower panel: ATT time dynamics. Figures are produced using the ‘gsynth’ package in R, developed by Xu and Liu (Mar 6, 2020)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- socialdistancedatacleaning.r
- socialdistancingmodel.r
- v1apr13data.rds