Causal Inference for COVID-19 Interventions

Vikas Ramachandra
Adjunct Professor, Data Science
University of California, Berkeley
Berkeley, CA
virama@berkeley.edu

Abstract

The exponential spread of the COVID-19 pandemic has caused countries to impose drastic measures on the public including social distancing, movement restrictions and lockdowns. These government interventions have led to different mobility patterns for the populations. We propose a method of causal inference using community mobility datasets to determine the treatment effects of government interventions on population mobility related outcomes. We first identify the changepoint based on the data of government interventions. We also perform changepoint detection to verify that there is indeed a changepoint at the time of intervention. Then we estimate the mobility trends using a Bayesian structural causal model and project the counterfactual. This is compared to the actual values after interventions to give the treatment effect of interventions. As a specific example, we analyze mobility trends in India before and after interventions. Our analysis shows that there are significant changes in mobility due to government interventions. Our paper aims to provide insights into changes in response to government measures and we hope that it is helpful to those making critical decisions to combat COVID-19.

1. Introduction and background

The COVID-19 has been declared by the World Health Organization as a pandemic on March 11, 2020. The exponential nature of the spread of the disease has led many countries to impose drastic measures on the public including social distancing, movement restrictions, public health measures, social and economic measures, and lockdowns. We would like to study the effects of these government interventions on population mobility in areas including retail & recreation, grocery & pharmacy, parks, transit stations, workplaces, and residential life. Our aim is to quantify the effect of the government interventions on the above mentioned outcomes. For this, we first identify changepoints based on the COVID-19 Government Measures Dataset by ACAPS [1]. We also run changepoint detection to verify that there is indeed a changepoint at the time of intervention. Then we run a Bayesian structural causal model on the COVID-19 Community Mobility Reports by Google [2] to project the counterfactual after the changepoint.

The projected counterfactual is then compared to the actual value after the interventions to find out the treatment effect i.e. the difference in the outcomes given an intervention and the estimated outcome value without the intervention. As a specific example, we run our proposed method on community mobility data of India. Our analysis aims to provide insights into changes in response to government measures and we hope that it is helpful to those making critical decisions to combat COVID-19.

This paper is organized as follows. Section 2 provides details about the dataset used, followed by Section 3, which provides details about each step in the algorithm (changepoint analysis and time series causal inference algorithms). Concluding remarks are provided in Section 4.

2. Details of the Datasets

For our analysis, we use the ACAPS COVID-19 Government Measures Dataset and the Google COVID-19 Community Mobility Reports. The ACAPS COVID-19 Government Measures Dataset provides comprehensive information of government interventions in response to the COVID-19 pandemic in the following five categories: social distancing, movement restrictions, public health measures, social and economic measures, and lockdowns. Government measures of different severity are included. The Google COVID-19 Community Mobility Reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. Specifically, the reports show how visits and length of stay at different places change compared to a baseline. Changes for each day are compared to a baseline value for that day of the week. The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3-Feb 6, 2020. We use India's government interventions and community mobility reports for our example. Among the government interventions, we use the social distancing and movement restrictions measures to identify the changepoint. And we perform our analysis across all categories of places.

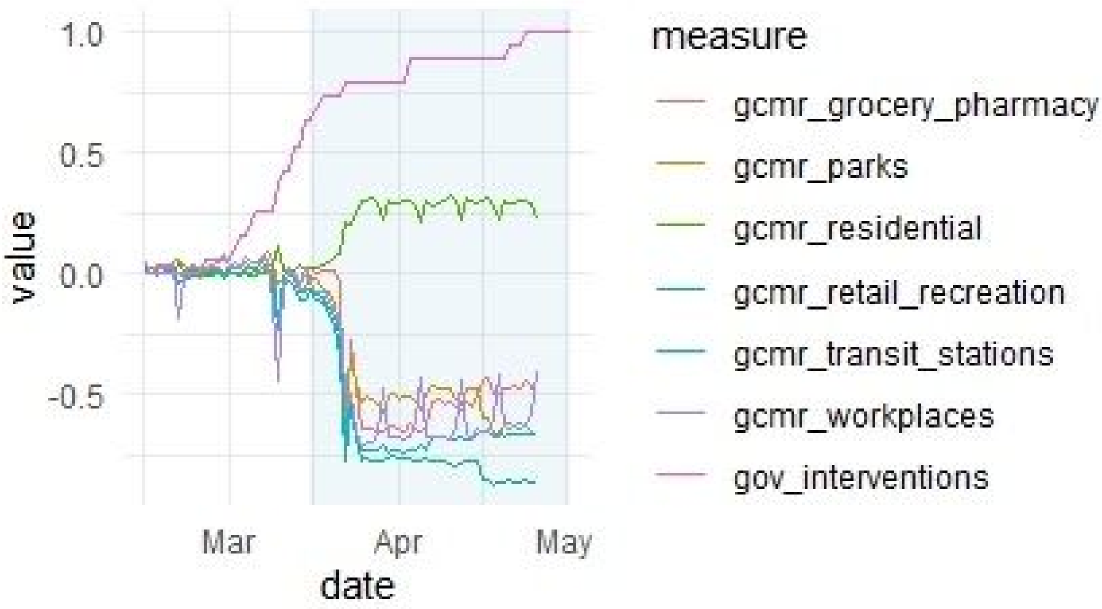

Figure 1: Plot of datasets

3. Proposed algorithm details

The proposed algorithm consists of the following steps:

A. Changepoint detection
B. Time series causal inference for population mobility trends

A. Changepoint detection

Our goal is to check whether the changepoint detection algorithm can identify the changepoint from the time series of mobility trends in Figure 1, thus verifying that the population movement patterns before and after the interventions are indeed different, and the difference is statistically significant.

We use the mean and variance based changepoint detection algorithm in the changepoint R package; the reader can find more details in the related documentation [3]. We recap some of the details below.

Time series changepoint detection using the mean and variance can be described as follows. Changepoint detection is the name given to the problem of estimating the point at which the statistical properties of a sequence of observations change.

Let us assume we have an ordered sequence of data: $y_{1:n} = (y_1, \dots, y_n).$

Changepoint occurs within this set when there exists a time: $\tau \in \{1, \dots, n-1\},$

such that the statistical properties of $\{y_1, \dots, y_\tau\}$ and $\{y_{\tau+1}, \dots, y_n\}$

are different in some way.

Single changepoint detection:

We briefly recap the likelihood based framework for changepoint detection. Instead of considering the more general problem of identifying $\boldsymbol{\tau}_{1:m}$ changepoint positions, we only consider the identification of a single changepoint here, for simplicity. The detection of a single changepoint can be posed as a hypothesis test. The null hypothesis, H_0, corresponds to no changepoint (m = 0) and the alternative hypothesis, H_1, is a single changepoint (m = 1). We now recap the general likelihood ratio based approach to test this hypothesis.

A test statistic can be constructed which we will use to decide whether a change has occurred. The likelihood ratio method requires the calculation of the maximum log-likelihood under both the null and alternative hypotheses. For the null hypothesis the maximum log-likelihood is log $p(y_{1:n}|\Theta\text{^})$, where p(.) is the probability density function associated with the distribution of the data and Θ^ is the maximum likelihood estimate of the parameters.

Under the alternative hypothesis, consider a model with a changepoint at $\boldsymbol{\tau}_1$ with $\boldsymbol{\tau}_1$ ε {1,2,..., n-1} Then the maximum log likelihood for a given $\boldsymbol{\tau}_1$ is:

$$ML(\tau_1) = \log p(y_{1:\tau_1}|\hat{\theta}_1) + \log p(y_{(\tau_1+1):n}|\hat{\theta}_2).$$

Given the discrete nature of the changepoint location, the maximum log-likelihood value under the alternative is simply $\max_{\boldsymbol{\tau}1}$ ML($\boldsymbol{\tau}_1$), where the maximum is taken over all possible changepoint locations. The test statistic is thus

$$\lambda = 2\left[\max_{\tau_1} ML(\tau_1) - \log p(y_{1:n}|\hat{\theta})\right].$$

The test involves choosing a threshold, c, such that we reject the null hypothesis if $\lambda > c$. If we reject the null hypothesis, i.e., detect a changepoint, then we estimate its position as $\boldsymbol{\tau}_{1,}$ the value of $\boldsymbol{\tau}$ that maximizes ML($\boldsymbol{\tau}_1$).

The test statistics used here are the mean and the variance.

As shown in Figure 2-4 below, a changepoint is detected for the retail & recreation, grocery & pharmacy, parks, transit stations, workplaces, and residential data around March 20, which is close to the time of lockdown in India, March 16. Alternatively, an expert could also pick the changepoint manually by visual inspection of the mobility data.

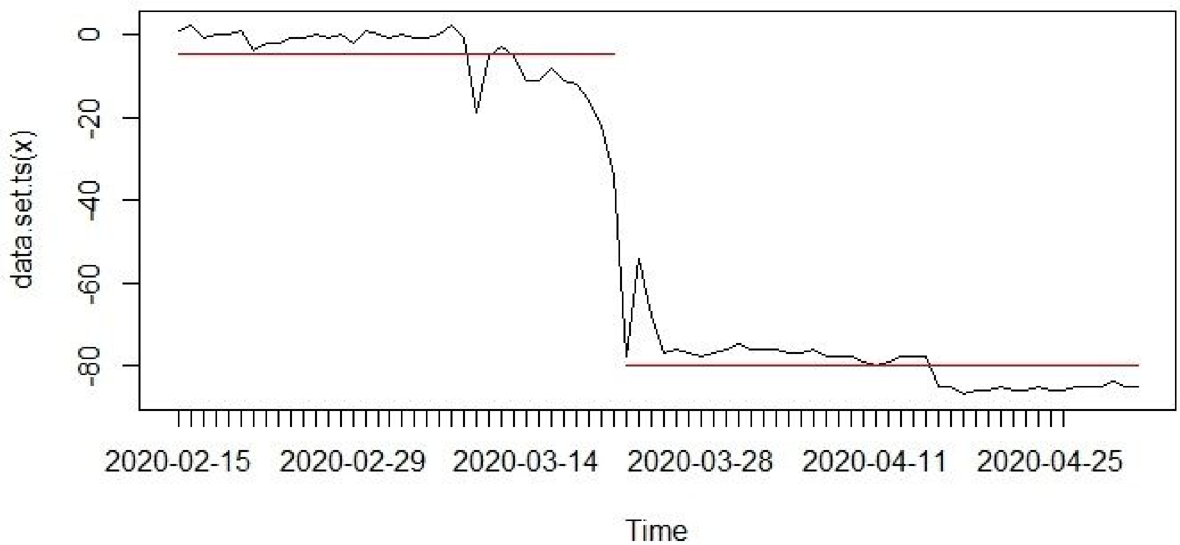

Figure 2: retail & recreation changepoint

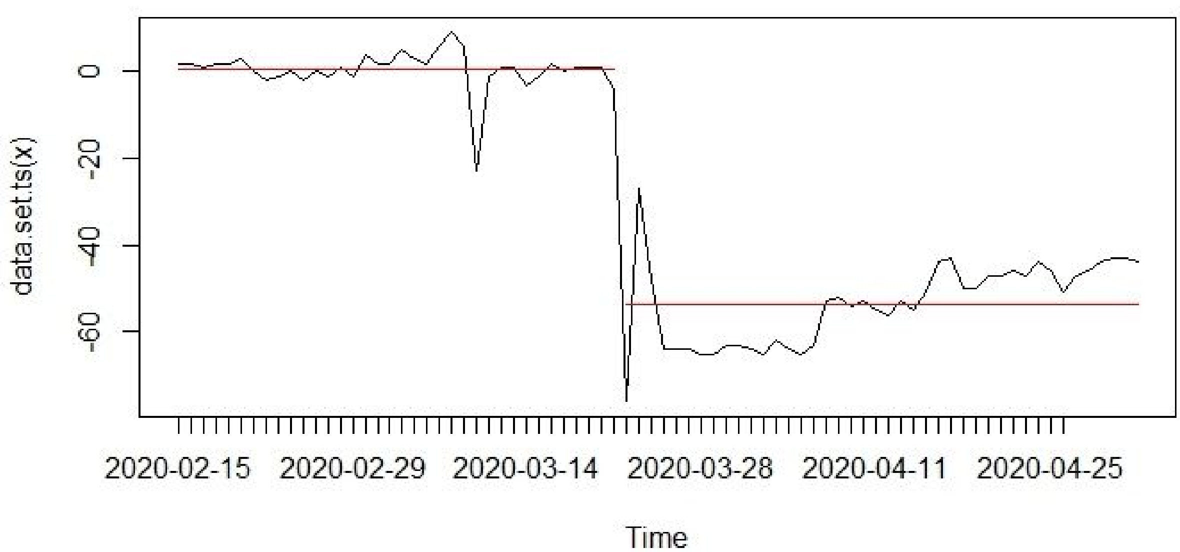

Figure 3: grocery & pharmacy changepoint

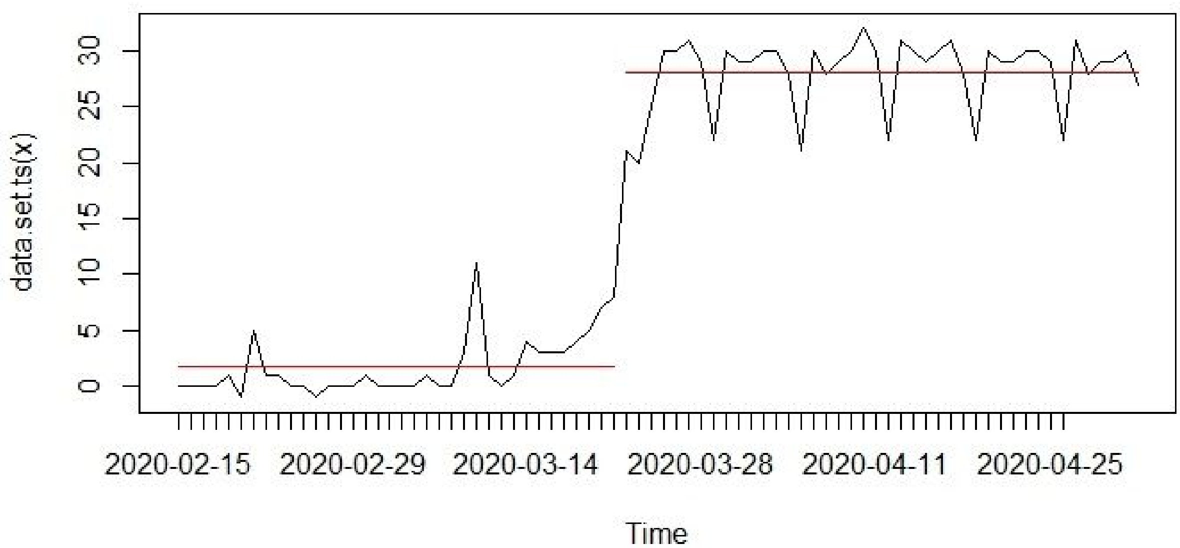

Figure 4: residential changepoint

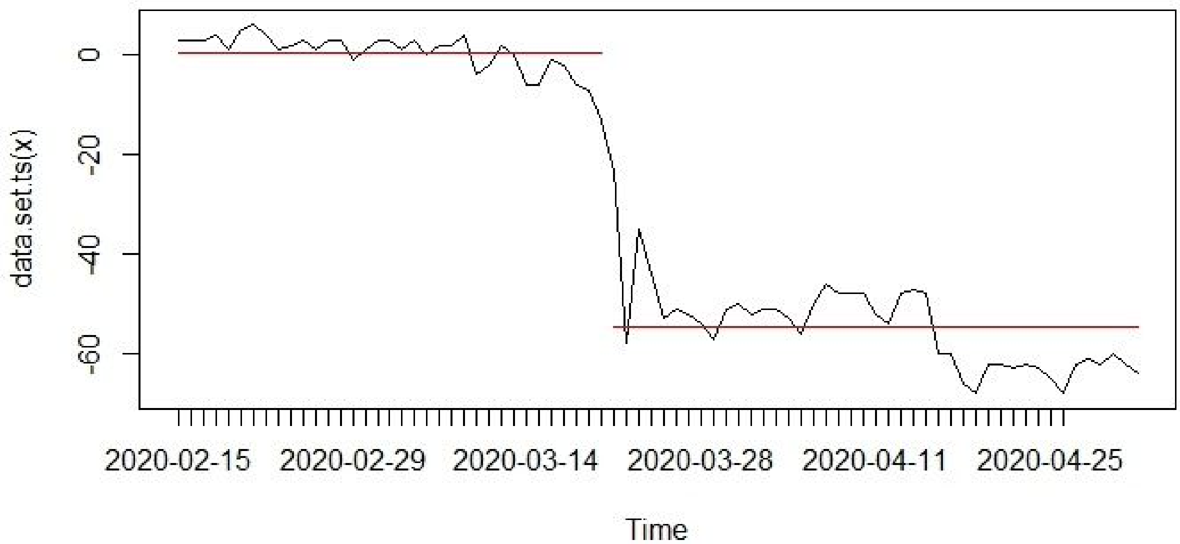

Figure 5: Parks changepoint

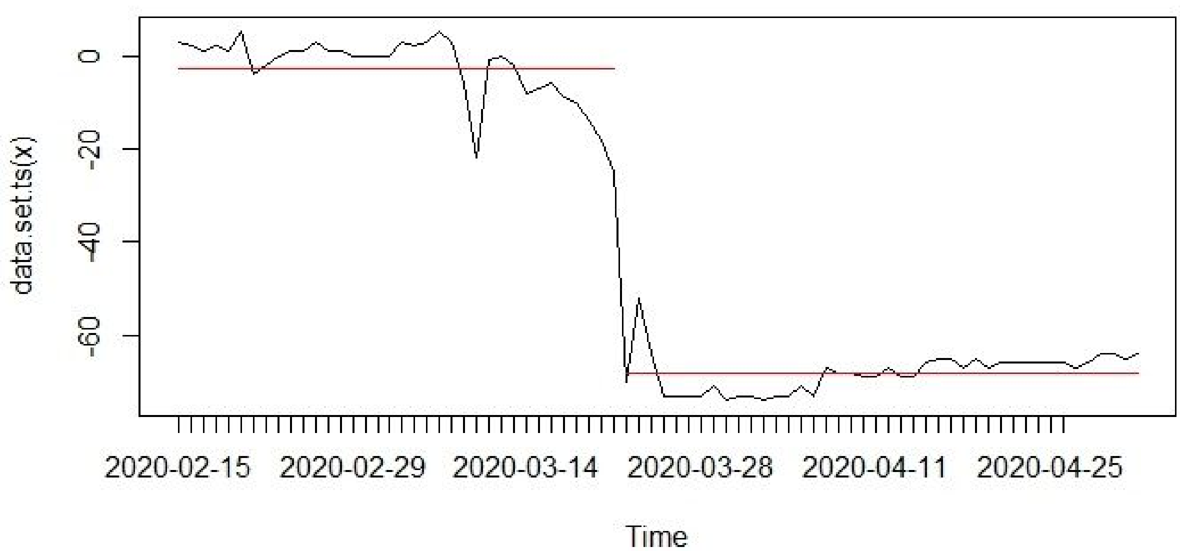

Figure 6: Transit stations changepoint

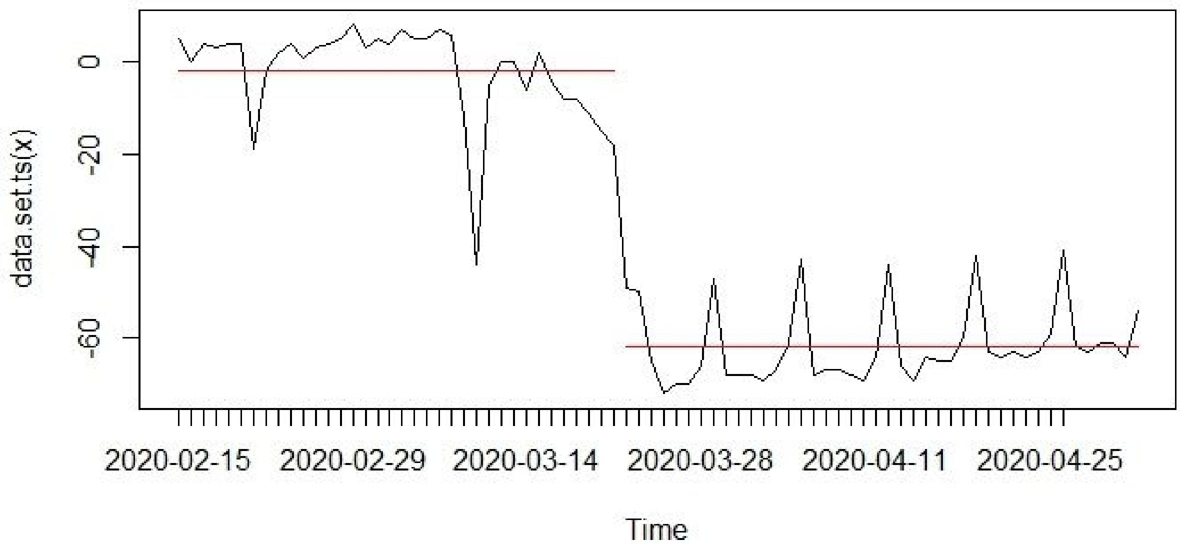

Figure 7: Workplaces changepoint

B. Time series causal inference for population mobility trends

For estimating treatment effect of the intervention, first we need to fit a model and forecast the counterfactual using the time series before the intervention and then compare that forecast to the actual time series recorded after the intervention. For this, we use the CausalImpact R package [4]. This package is based on causal inference for time series data using Bayesian structural models, we recap some of the details below. The original paper has a much more comprehensive discussion about the methods used, and the reader is referred to it for more information [4].

The above method uses state space models and flexible Bayesian priors to fit a time series model pre-intervention, and forecast/predict the counterfactual based on the fit model.

Structural time-series models are state-space models for time-series data. They can be defined in terms of a pair of equations:

$$y_t = Z_t^{\mathrm{T}} \alpha_t + \varepsilon_t,$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t,$$

where εt ∼ N (0,σ2t) and ηt ∼ N (0,Qt) are independent of all other unknowns. The first equation above is the observation equation; it links the observed data yt to a latent d-dimensional state vector αt . The second equation above is the state equation; it governs the evolution of the state vector αt through time. Here, yt is a scalar observation, Zt is a d-dimensional output vector, Tt is a $d \times d$ transition matrix, Rt is a $d \times q$ control matrix, εt is a scalar observation error with noise variance σt, and ηt is a q-dimensional system error with a $q \times q$ state-diffusion matrix Qt, where $q \leq d$.

The above parameters are learned using a Bayesian framework. Posterior inference is based on the Markov chain Monte Carlo (MCMC) technique, and a Gibbs sampler. Subtracting the predicted from the observed response during the post-intervention period gives a semiparametric Bayesian posterior distribution for the causal effect.

The input features are as derived in step A (tree cover percentage calculation). and described above. Formally, the covariates 'x' are location variables, and the output variable 'y' is the forest tree cover percentage at each given time point, and the treatment/ intervention variable is the location of the changepoint from step B.

For our example, we perform time series causal inference on all 6 categories of places: retail & recreation, grocery & pharmacy, parks, transit stations, workplaces, and residential. The results of this step are shown in Figure 8-13 below.

In Figure 8-13 below, the top panel shows the counterfactual estimate (dashed line) and confidence intervals for it, and the solid line is the actual observed values (beyond March 16, 2020, the intervention event of lockdown). The middle panel shows the estimated treatment effect, here the difference in the movement data between no intervention (projected counterfactual) and actual observed movement data. E.g.: on Figure 8, on April 1, the movement data (visits and length of stay) of retail & recreation is -75% (reduction) as compared to the projected no-intervention or 'before' values, this shows that government interventions did cause an significant decrease of activities in retail & recreation areas.

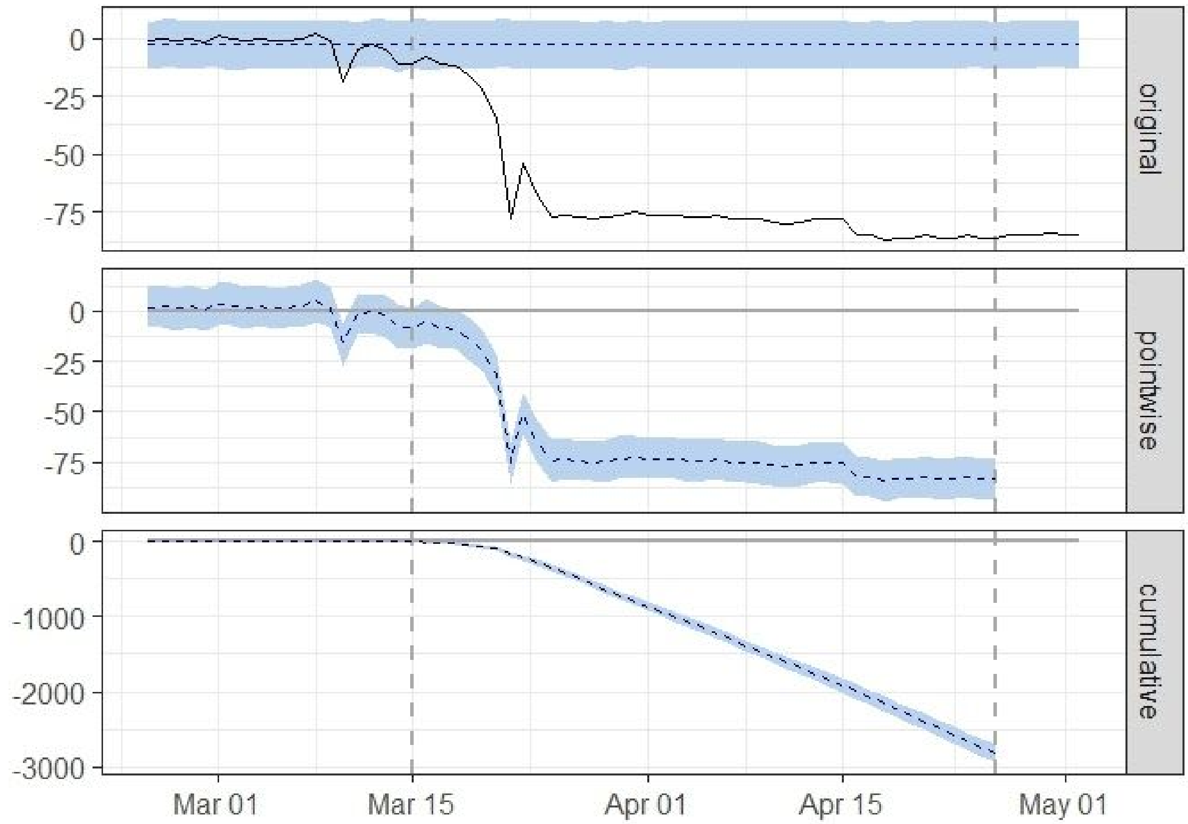

Figure 8: Retail & Recreation

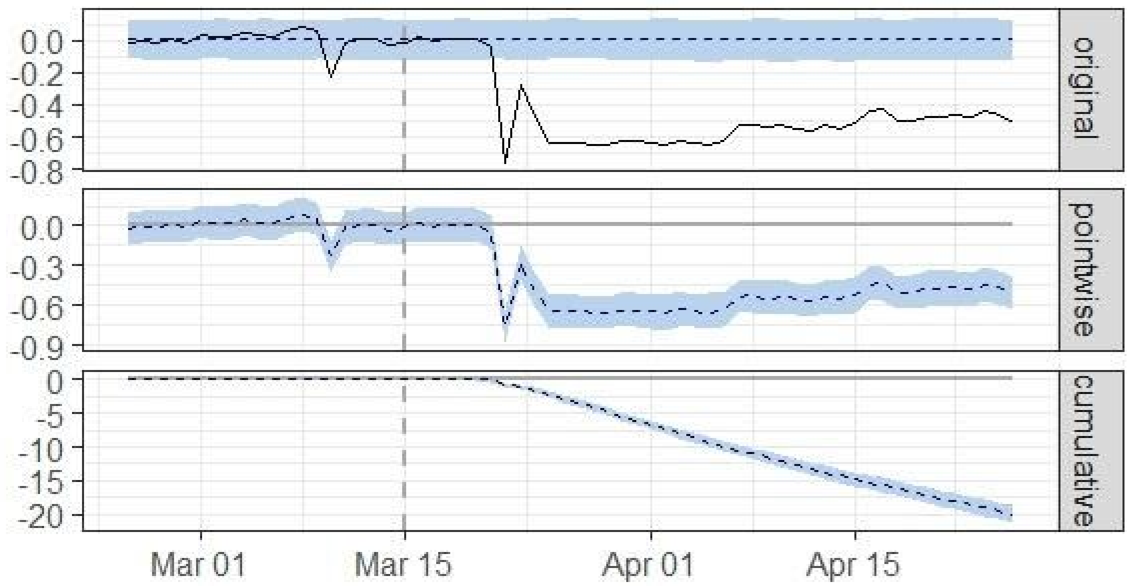

Figure 9: Grocery & Pharmacy

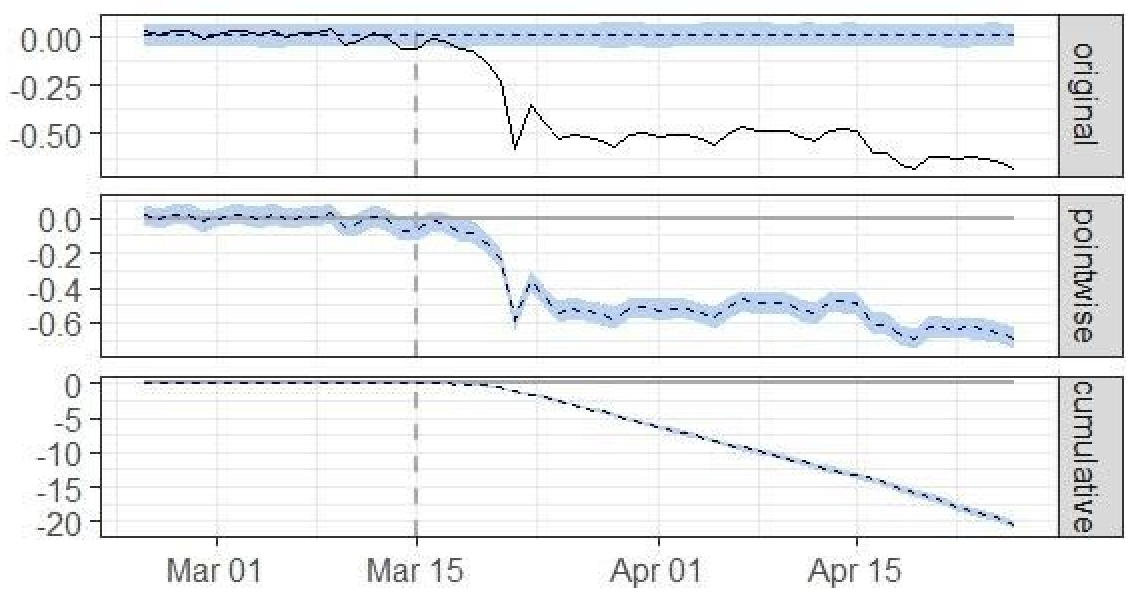

Figure 10: Parks

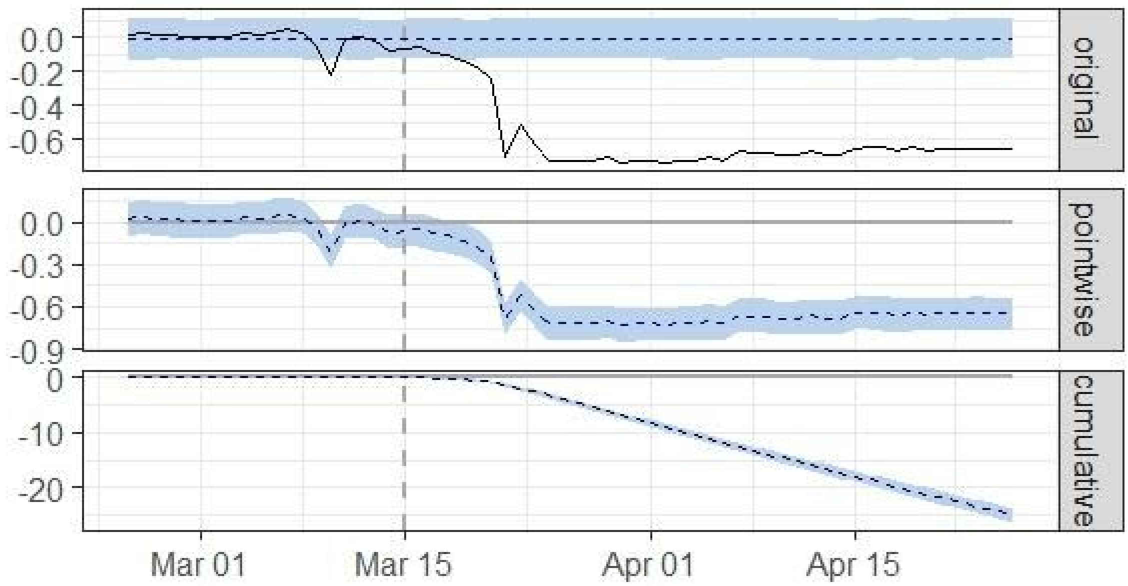

Figure 11: Transit stations

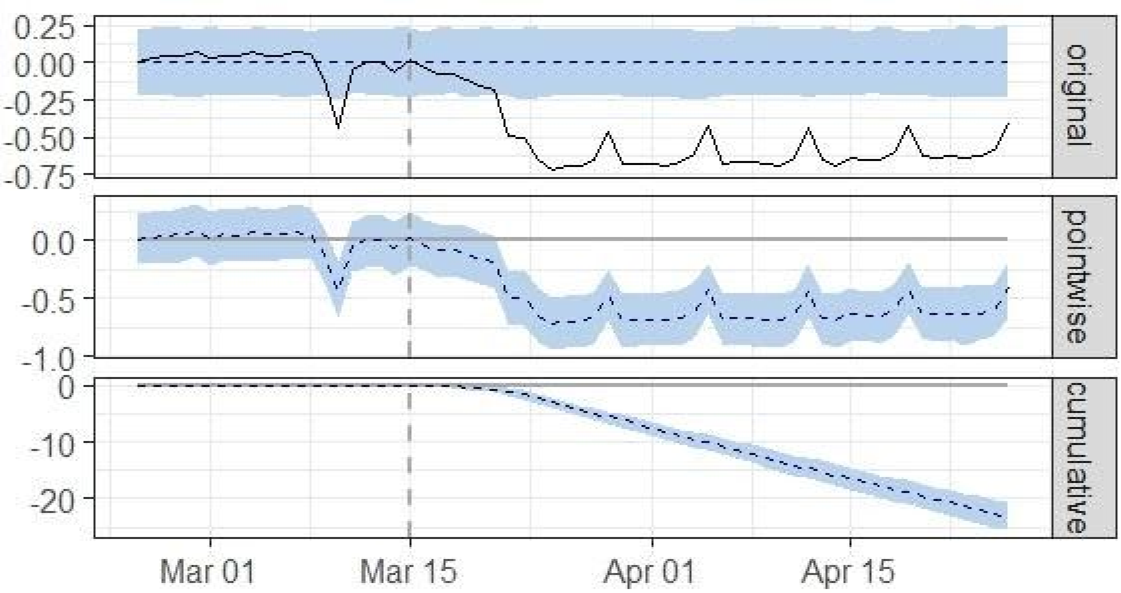

Figure 12: Workplaces

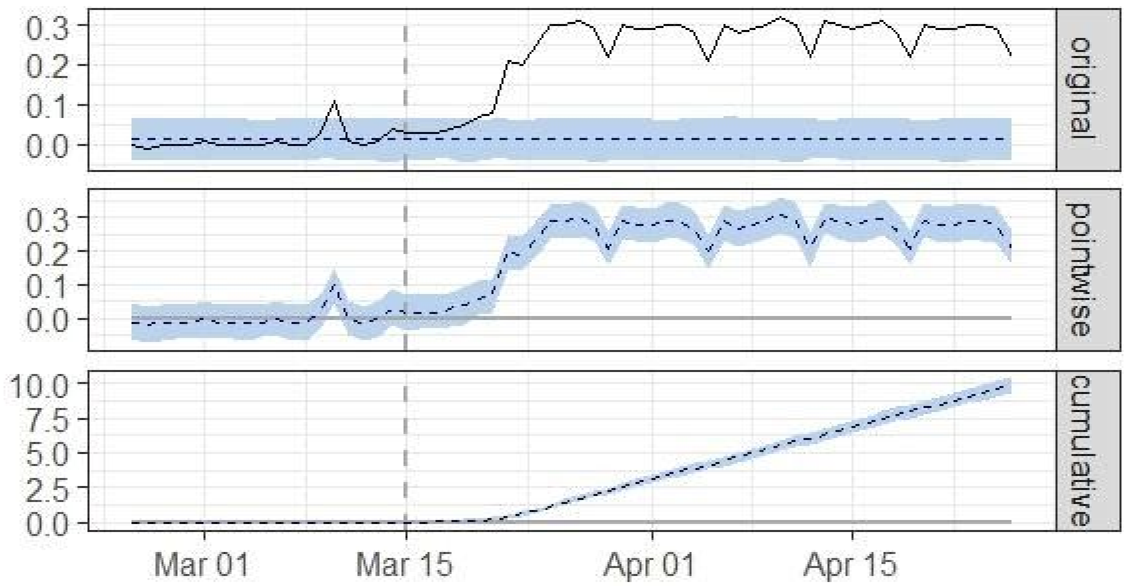

Figure 13: Residential

Top panel: shows the counterfactual estimate (dashed line) and confidence intervals for it, and the solid line is the actual observed values (beyond March 16, 2020, the intervention event of lockdown). Middle panel: Shows the estimated treatment effect, here the difference in the movement data between no intervention (projected counterfactual) and actual observed movement data. E.g.: on Figure 5, on April 1, 2020, the movement data (visits and length of stay) of retail & recreation is -75% (reduction) as compared to the projected no-intervention or 'before' values, this shows that government interventions did cause an significant decrease of activities in retail & recreation areas.

4. Conclusion

In this paper, an algorithm for causal inference using population movement data has been proposed, for COVID-19 interventions. A changepoint detection method is applied to the time series movement data to identify the time of intervention. Alternatively, we can use the recorded implementation time of government policies such as lockdown as the changepoint. Then we use a Bayesian structural causal model to forecast the values beyond the intervention point in the counterfactual scenario. This forecast is then compared to the actual observed value, and their difference gives the treatment effects of government measures. We apply the above algorithm to India's mobility data in retail & recreation, grocery & pharmacy, parks, transit stations, workplaces, and residential, and we estimate that the movements in these areas are much lower than before the intervention, except for residential areas where the movements are much higher, due to social distancing and movement restrictions policies. We hope that our analysis gives insights into how government interventions change population movements and is helpful for those making critical decisions combating the COVID-19 pandemic.

References

[1] ACAPS COVID-19 Government Measures Dataset

https://www.acaps.org/covid19-government-measures-dataset

[2] Google COVID-19 Community Mobility Reports

https://www.google.com/covid19/mobility/

[3] Changepoint package in R, 2016

https://cran.r-project.org/web/packages/changepoint/changepoint.pdf

[4] Bayesian structural time series models in R: CausalImpact package, 2015

https://cran.r-project.org/web/packages/CausalImpact/vignettes/CausalImpact.html