**SUPPLEMENTARY**

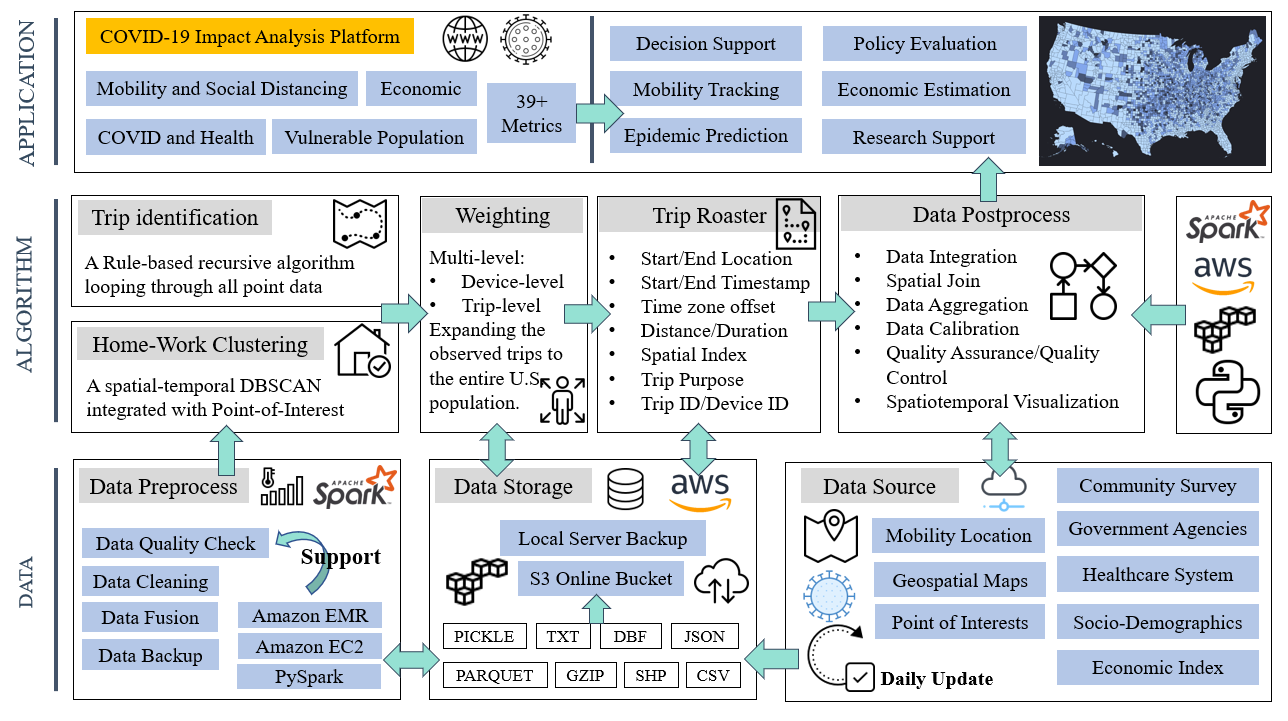
**Mobile device data reveals the dynamics in a positive relationship between human mobility and COVID-19 infections**

Chenfeng Xiong a,b,1, Songhua Hu a, Mofeng Yang a, Weiyu Luo a, and Lei Zhang a,1

1. Maryland Transportation Institute (MTI), Department of Civil and Environmental Engineering, 1173 Glenn Martin Hall, University of Maryland, College Park MD 20742.
2. Shock Trauma and Anesthesiology Research (STAR) Center, School of Medicine, 685 W Baltimore Street, Suite 600, University of Maryland, Baltimore MD 21201.
3. **Details of mobility metrics**

A data panel of emerging mobile device location data representing person movements for the entire U.S. is developed, incorporating over 100 million anonymous monthly active mobile devices. To fully capture all covariates, county attributes, public health measures, and other information are integrated into the data source. A set of previously developed and validated data analytics algorithms are integrated to identify activity locations, derive weighted trip rosters, and calculate and validate the human mobility metrics (Figure 1):

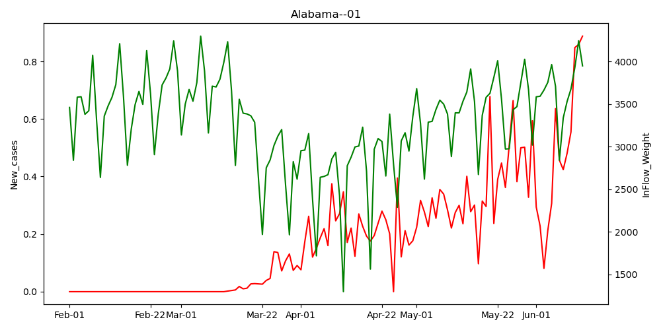
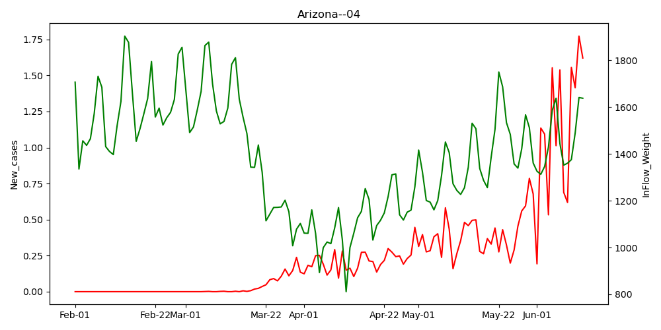
* First, a heuristic rule-based methodology is employed to identify activity locations and integrated with Point-of-Interest (POI) information. Sensitive locations such as the home and work are anonymized at the census block group level to protect privacy.
* Then, a rule-based recursive algorithm is used to identify trips from raw location points. This algorithm checks every point in the sequence sorted by devices and timestamps to identify if it belongs to the same trip as its previous point based on speed, time, and distance threshold.
* Next, a multi-level weighting procedure expands the observed trips to the entire U.S. population, using device-level and trip-level weights to ensure data representativeness in the total population.
* Finally, based on the weighted trip roster, various human mobility metrics are calculated via a post-processing step. The mobility informatics are analyzed daily at the national, state, and county levels in the U.S. and made available to the general public via the COVID-19 impact analysis platform (<https://data.covid.umd.edu/>).

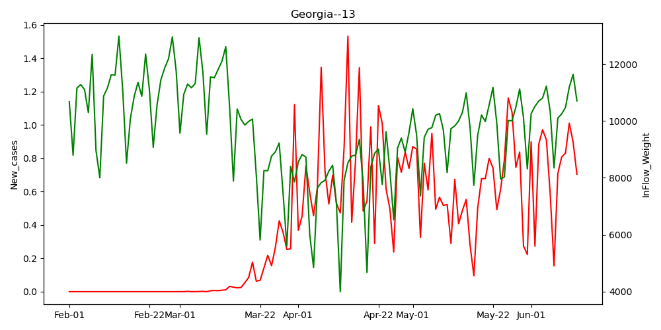
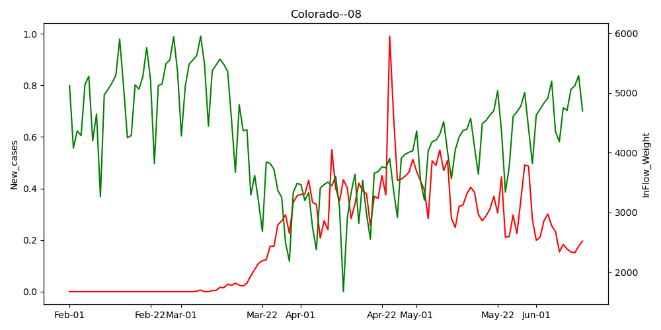


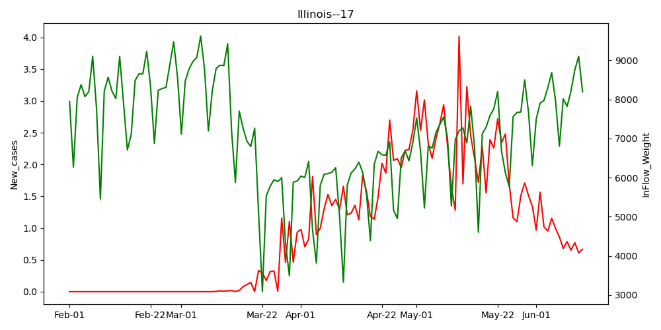
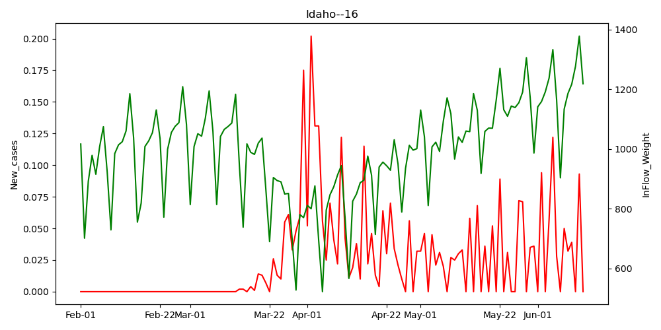
**Figure 1 A Big-Data Driven Analytical Framework for Understanding Human Mobility Trend and Policy Decision Support during COVID-19 Pandemic**

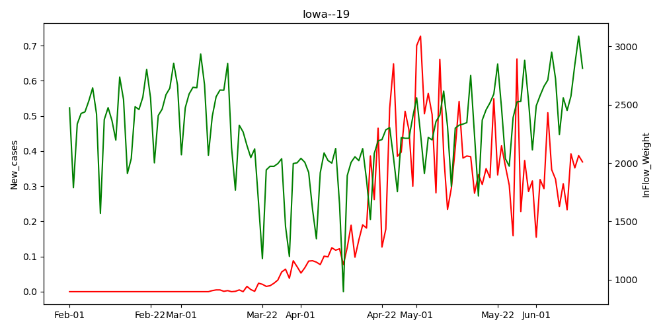
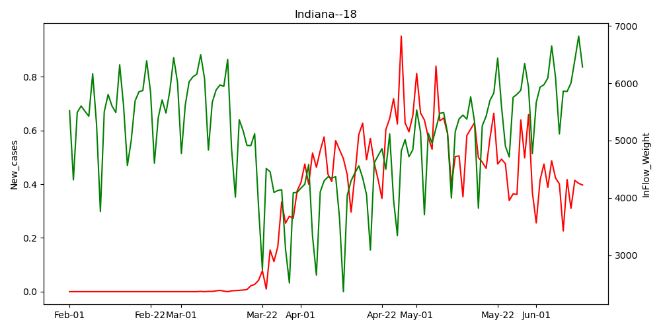
1. **Number of New cases and Inflow varying in reopen states**

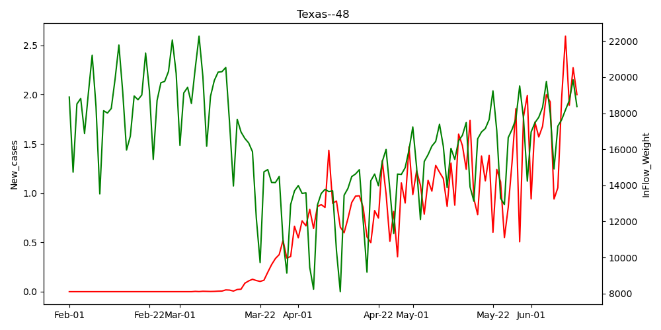
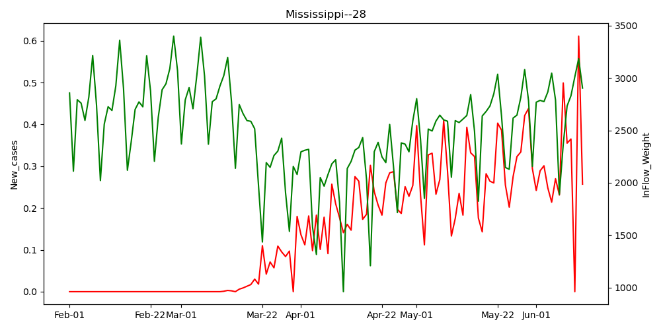
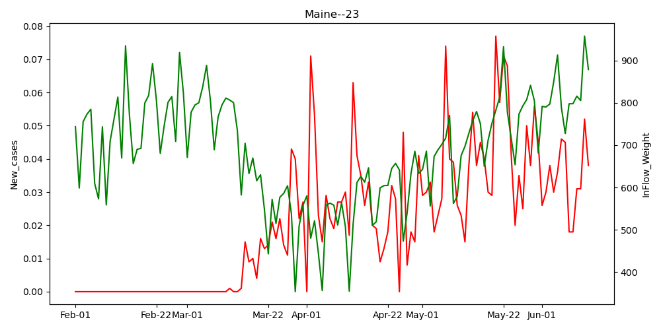
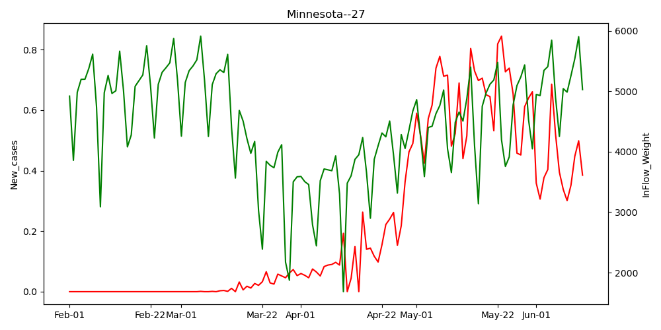
19 states are included in the group named reopen states, including Alabama, Arizona, Colorado, Georgia, Idaho, Illinois, Indiana, Iowa, Maine, Minnesota, Mississippi, New Mexico, North Dakota, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, and Utah, based on whether the partial reopen orders issued before 2020/05/01. For each state, the daily new cases and the daily inflow are plotted and stored in Fold “\Mobility\_COVID19\_PNAS\Figure-State Level”. As shown, most reopen states present a sharp increase in both inflow and the number of new cases after May 1st, 2020. Besides, compared with locked-down states, most reopen states present a smaller number of cases in their early stages, which may be the main reasons why these states first withdraw their stay-at-home orders.

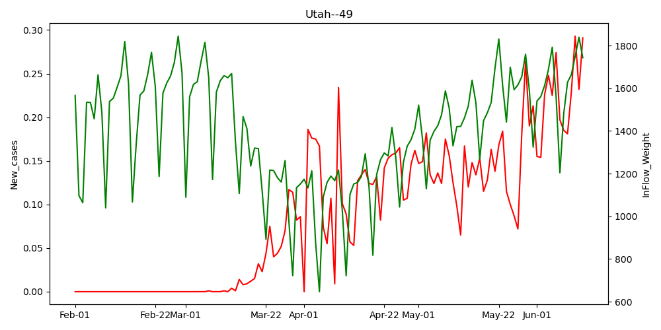
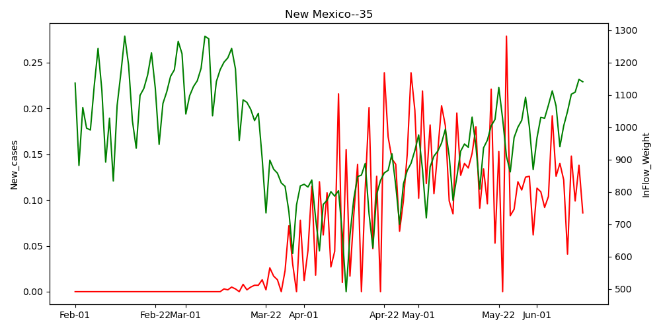
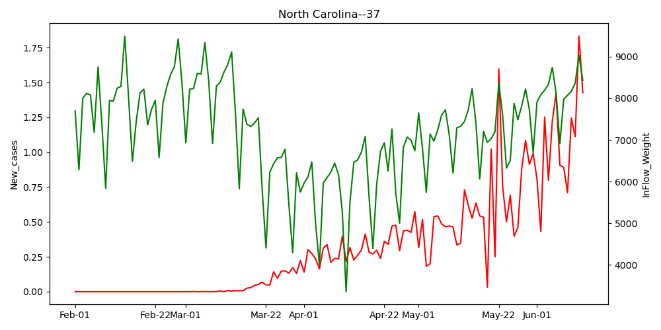
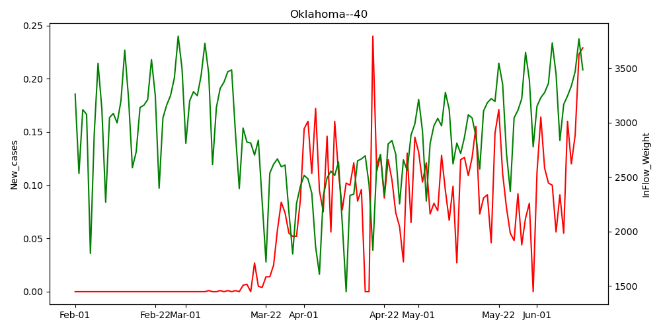
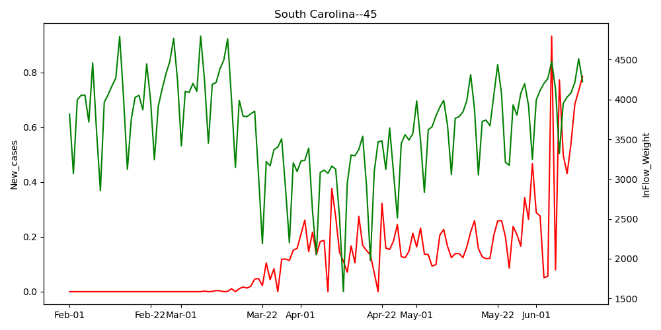
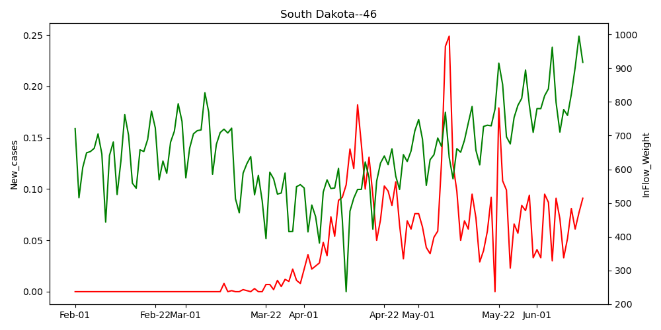
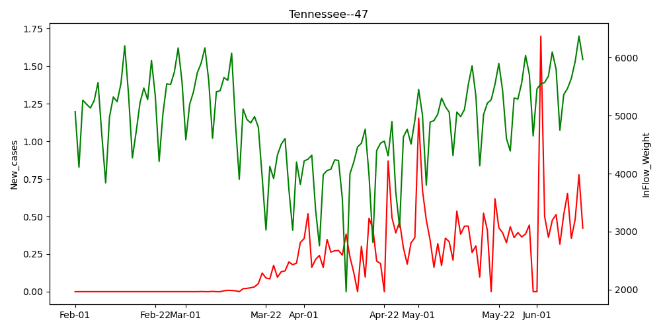
 







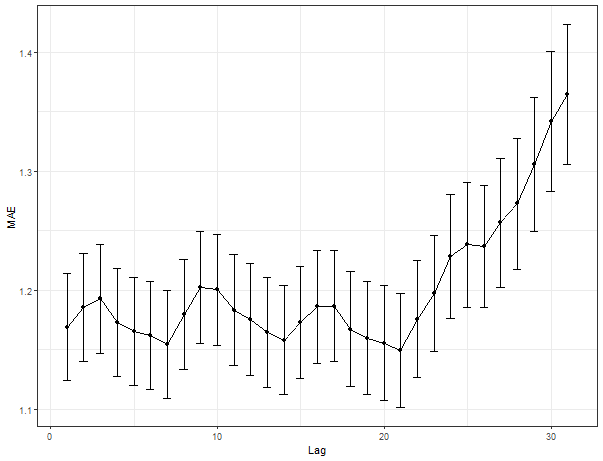




**Figure 2 The evolution of mobility inflow and number of daily confirmed cases in reopened states.**

1. **Optimal Lag**

To find the optimal lag, we change the lag (i.e. the lag presented in Equation 1) from 1 to 30 days and calculate the mean absolute error (MAE) for each model. A lower MAE indicates the model performs better under that lag. Results show the MAE is lower when the lag is smaller than 21 days, after that, the MAE sharply increases. This indicates the inflow can affect the number of cases in a 21-days interval, which is consistent with the incubation period of COVID-19. We finally choose 7 days as the optimal lag.

****

**Figure 3 The evolution of model MAE over different lags.**

1. **Model performance**

From a global SEM aspect, including mobility variables present a significant improvement in model goodness-of-fit. We conduct an ANOVA test between the model with/without inflow as an independent variable and report the results in the following table. As shown, the RMSEA (Root Mean Square Error of Approximation) of the SEM decreases by 42.69% and Chisq. decreases by 66.90% after including the inflow as a predictor, indicating the inflow significantly enhance the model performance.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | chisq | df | pvalue | cfi | tli | aic | bic | rmsea | srmr |
| With Inflow | 1182.441 | 8.000 | 0.000 | 1.000 | 0.999 | 84145.397 | 84351.021 | 0.051 | 0.004 |
| Without Inflow | 3568.216 | 8.000 | 0.000 | 0.999 | 0.997 | 86531.170 | 86736.795 | 0.089 | 0.011 |

1. **Model Interpretation**

The summary of the model coefficients across all time windows is presented in the following table. All values are calculated based on time windows with statistical significance (i.e. P-value < 0.1). We expect about 2.34% ((1.1) ^ 0.243=1.01762) increase in number of new cases when inflow increases by 10%. The effect of inflow is time-varying, however, from a minimal of 1.45% to a maximal of 2.96%.

Besides the mobility variables, we found the AR (1) term presents the greatest positive relationship with the number of new cases. Also, the population density presents significant positive relationships with new cases over most time windows.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Response | Predictor | Mean | St.d. | Min | Max | No. of sig. Time windows |
| Lag7\_Weighted\_Inflow | (Intercept) | 5.5561 | 8.0514 | -6.0903 | 34.1223 | 83.0000 |
| Employment density | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 5.0000 |
| Is Weekend | -0.2938 | 0.1074 | -0.6287 | -0.1665 | 87.0000 |
| Lag7\_Log\_National\_Cases | -0.3080 | 0.4652 | -1.9566 | 0.4000 | 83.0000 |
| Lag7\_Precipitation | 0.0052 | 0.0053 | -0.0066 | 0.0171 | 60.0000 |
| Lag7\_Temperature | -0.0005 | 0.0023 | -0.0068 | 0.0041 | 60.0000 |
| AR (1) | 0.9849 | 0.0089 | 0.9636 | 0.9974 | 87.0000 |
| Median Income | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 51.0000 |
| Pct\_Age\_0\_24 | 0.0080 | 0.2333 | -0.2581 | 0.5424 | 25.0000 |
| Pct\_Age\_25\_40 | 0.1385 | 0.2230 | -0.4054 | 0.3638 | 13.0000 |
| Pct\_Age\_40\_65 | -0.1058 | 0.3459 | -0.5985 | 0.5220 | 18.0000 |
| Black | -0.0216 | 0.0746 | -0.1209 | 0.1199 | 26.0000 |
| White | -0.0657 | 0.0510 | -0.1354 | 0.0746 | 27.0000 |
| Population density | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 12.0000 |
| New Cases | (Intercept) | -2.2167 | 0.4619 | -3.1537 | -1.1019 | 78.0000 |
| Is Weekend | -0.0766 | 0.0189 | -0.1285 | -0.0455 | 43.0000 |
| AR (1) | 0.6684 | 0.0494 | 0.5875 | 0.8060 | 87.0000 |
| Lag7\_Weighted\_Inflow | 0.2425 | 0.0371 | 0.1518 | 0.3061 | 87.0000 |
| Median Income | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 14.0000 |
| Pct\_Age\_0\_24 | 1.0616 | 0.2201 | 0.7715 | 1.5377 | 40.0000 |
| Pct\_Age\_25\_40 | -0.4739 | 1.5779 | -2.6894 | 1.1759 | 9.0000 |
| Pct\_Age\_40\_65 | -0.7817 | 2.0695 | -3.7165 | 2.4613 | 28.0000 |
| Population density | 0.0001 | 0.0000 | 0.0000 | 0.0001 | 80.0000 |

Some of the factors raised in this comment, such as the facemask usage, social distancing compliance, could not be directly measured and incorporated in the model. However, the effect of these factors that are not directly observable could be inferred via the model’s time-varying intercept coefficients (visualized below). Figure 4 visualizes the time-varying intercept coefficients for Equation [1] for the daily average number of confirmed new cases. A negative and statistically significant intercept is estimated, indicating an overall negative effect on the number of cases from all unobservable factors. More interestingly, this negative effect stays stable for the “locked-down” counties but got dampened for the “reopened” counties.

A close up of a map

Description automatically generated

**Figure 4 The time-varying intercepts in the new case equations.**

1. **Other methods: BSTS**

We also test the Bayesian structural time series (BSTS) to quantify the causal effect of stay-at-home orders on mobility inflow. We use the trip per person change (compared with January 2020) as the dependent variable and set the date when the stay-at-home order issued as the intervention time.

Figure 5 presents the results of using New York county as a case study. We include a weekly seasonality and various regressors such as the number of cases and weather in the BSTS model. Results show the total policy effect is positive (with absolute effect of 0.1 (CI: -0.95, 1.2)) and not significant (P-Value = 0.44096). We also expand the BSTS models across all counties in the U.S. Results show that stay-at-home orders only present significant causal impact (i.e. P-value <= 0.1) in 247 counties, with an average effect of -0.094.

Considering the limited number of effective counties and the weak capacity in handling panel data with spatial heterogeneity, we finally choose the SEM with time-varying effects to specify the relationship between the number of cases and inflow. Then, the causal impact that the BSTS could bring in can later be extended to in the SEM studies via a Granger causality analysis.

A close up of a map

Description automatically generated

**Figure 5 BSTS results in New York County.**

A close up of a map

Description automatically generated

**Figure 6 Significant Causal impact of stay-at-home orders across the nation.**

1. **Other variables: the risked inflow**

We also consider weighing the inflow by the number of new cases. We found the pattern is similar to directly using inflow as independent variables, although the coefficients are higher. We finally keep using inflow as independent variables since using the number of cases to weight the inflow may lead to data leakage considering the number of cases is also used as a dependent variable.

A close up of a map

Description automatically generated

**Figure 7 The relationship between mobility inflow (weighted by cases) and confirmed cases in each county**