**SUPPLEMENTARY:**

Additional Details about the Data, Analytics, and Modeling

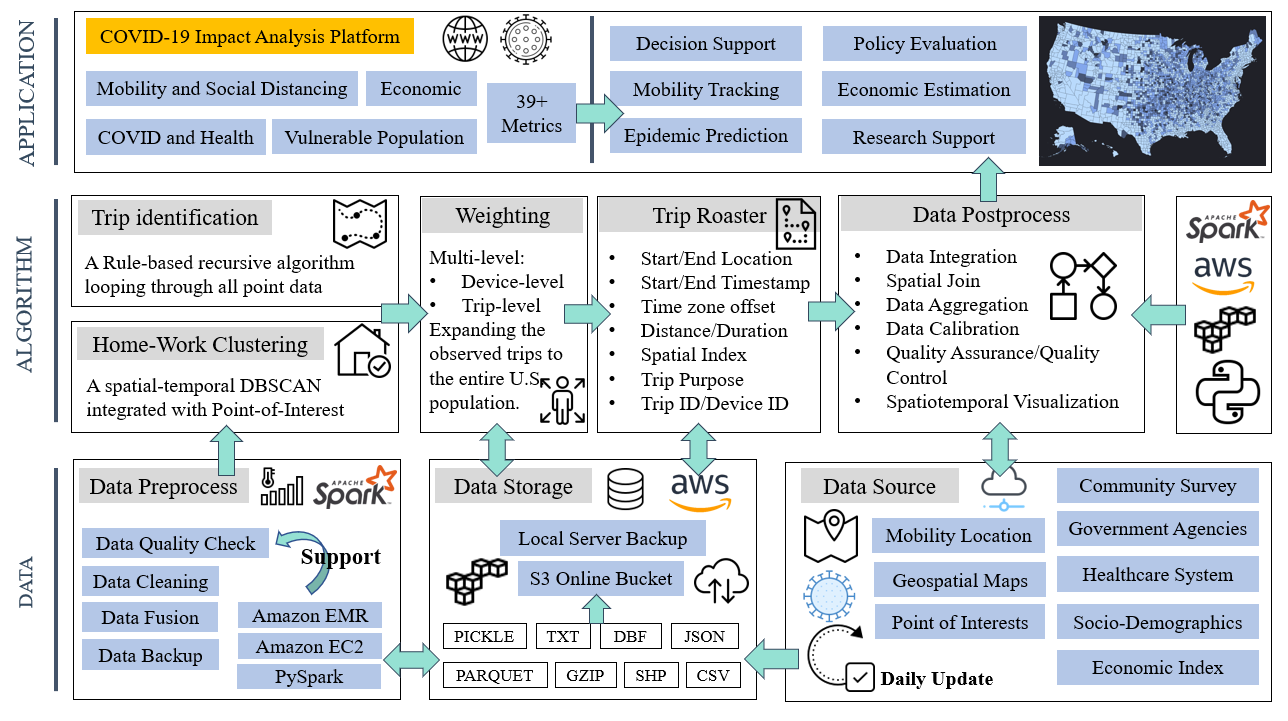
**Mobile device data reveals the dynamics in a positive relationship between human mobility and COVID-19 infections**

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**

***State-of-the-practice data processing***: Some common issues, such as unordered and duplicated records, need careful treatment before extracting any information from mobile device location data. The state-of-the-practice methods for raw data cleaning and quality control often include identifying and merging duplicate device observations, removing outliers, and checking on the obvious data consistency issues (e.g., devices with unreasonably high-speed readings). Figure S2 shows a general data cleaning procedure for mobile device location data taken by the research team based on the four dimensions of data quality assessment: consistency, accuracy, completeness, and timeliness.



Figure 2 State-of-the-practice data cleaning procedure

The completeness dimension cannot be considered without prior knowledge of the actual individual movements and mobile device usage. The timeliness is addressed by using daily feeds of mobile device location data for our application. For the first two dimensions related to data cleaning, the consistency dimension defines certain semantic rules that a set of data items should obey. A common type of semantic rule is integrity constraints. For example, the latitude and longitude of a location observation should be within a reasonable range. According to the integrity constraints, the cleaning procedure first deletes records with invalid entries and duplicate records to reduce redundancy. Since one subject cannot be at more than one place at the same time, the procedure keeps only one location record per second (with the highest accuracy, if applicable). Another important dimension of data quality assessment is accuracy, including syntactic and semantic accuracy. The syntactic accuracy measures the closeness of a value to all the elements of its corresponding definition domain. The semantic accuracy measures the closeness of a value to its real-world value. For example, an accuracy of 10 meters in a location sighting indicates that the subject should be within a radius of 10 meters from the observed location with a certain confidence level, e.g., 95%.

Therefore, the cleaning procedure removes the noisy records with extremely poor accuracy, e.g., two miles.

Location data providers describe their sample sizes with statistics such as daily active users (DAU) and monthly active users (MAU). MAUs are devices that are observed at least once a month and DAUs are devices that are continuously observed throughout the month. Reported data coverage by major data providers ranges between 5% to 70%, depending on whether they report MAU or DAU and how they define active users. While the overall sample size is measured by daily and monthly active users, these measures do not take into consideration that some devices may provide many sightings every day while other devices may only provide a few sightings in a very small number of days. **Table 1** presents more information about the mobile device location dataset used in this research. The following definitions describes the variables presented in the table:

* Population coverage: number of devices with identifiable home census block group (CBG) divided by the population of the study area.
* Temporal consistency: average number of days a device is observed in the study period.
* Frequency: the average location observations per device per day.
* Geographical representativeness: variance of population coverage among different zones of the study area, measured by a Gini coefficient between 0 and 1, with 0 indicating equal sampling rate in all zones and 1 indicating that all observed devices are from a single zone.
* Device representativeness: a measure of the variance in the location point frequency among observed devices. This measure shows if observed devices are comparable in terms of their data frequency and are also measured by a Gini coefficient falling between 0 and 1. Raw data representativeness has a lower value if all observed devices have more consistent data frequency.
* Hourly and daily temporal coverage: a measure of the variation of the number of location point observations among different hours of the day and different days of the month, respectively. Lower values between 0 and 1 indicate a more equitable distribution.

Table 1 Data quality comparison among three commercial LBS datasets

|  |  |
| --- | --- |
| **Selected Raw Data Quality Metrics** | **Mobile Device Location Dataset** |
| Population coverage (%) | 23.92 |
| Geographical representativeness (0~1) | 0.09 |
| Frequency (observations per device per day) | 190 |
| Temporal consistency (days per device) | 14.67 |
| Device representativeness (0~1) | 0.67 |
| Hourly temporal coverage (0~1) | 0.249 |
| Daily temporal coverage (0~1) | 0.03 |

***Trip identification***: Trips are not initially included in any mobile device location data sources. Instead, location sightings are continuously generated while the sample device moves, stops, stays static, or starts a new trip. As a result, we developed a trip identification algorithm, which can detect which location sightings form a trip together. We first sort device observations by time. The algorithm assigns a random ID to each trip it identifies. Many location points in the dataset may belong to no trips. The algorithm assigns “0” to the trip ID of these locations to tag them as static points. For every location point, we calculate distance, time, and speed between the point and its immediate previous and next points, if exist. Three hyperparameters need to be set for the algorithm: distance threshold, time threshold, and speed threshold. The speed threshold us used no identify if a location point is recorded on the move. The distance and time threshold are used to identify stay locations and trip ends. At this step, the algorithm identifies the device’s first observation with *speed from≥speed threshold*. This identified location point is recorded on the move, so a hashed trip ID is generated and assigned to this point. All points recorded before this point, if exist, are set to have “0” as their trip ID. Next, a recursive algorithm identifies if the next points are on the same trip and should have the same trip ID.

Then, a recursive algorithm has been developed to check every point to identify if they belong to the same trip as their previous point. If they do, they are assigned the same trip ID. If they do not, they are either assigned a new hashed trip id (when their *speed from≥speed threshold*) or their trip ID is set to “0” (when their *speed from<speed threshold*). Identifying if a point belongs to the same trip as its previous point is based on the point’s “*speed to*”, “*distance to*” and “*time to*” attributes. If a device is seen in a point with *distance to≥distance threshold* but is not observed to move there (*speed to<speed threshold*), the point does not belong to the same trip as its previous point. When the device is on the move at a point (*speed to≥speed threshold*), the point belongs to the same trip as its previous point; but when the device stops, the algorithm checks the radius and dwell time to identify if the previous trip has ended. If the device stays at the stop (points should be closer than the distance threshold) for a period of time shorter than the time threshold, the points still belong to the previous trip. When the dwell time reaches above the time threshold, the trip ends, and the next points no longer belong to the same trip. The algorithm does this by updating “time from” to be measured from the first observation in the stop, not the point’s previous point. The algorithm may identify a local movement as a trip if the device moves within a stay location. To filter out such trips, all trips that are shorter than 300 meters are removed.

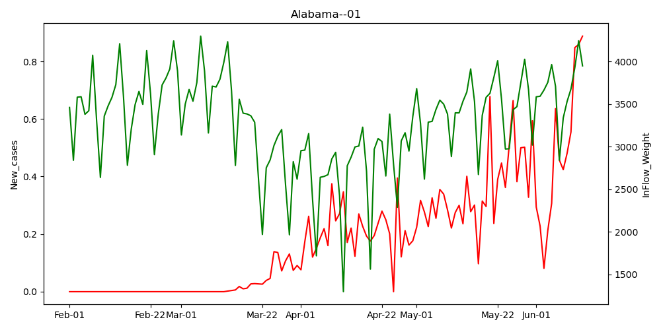
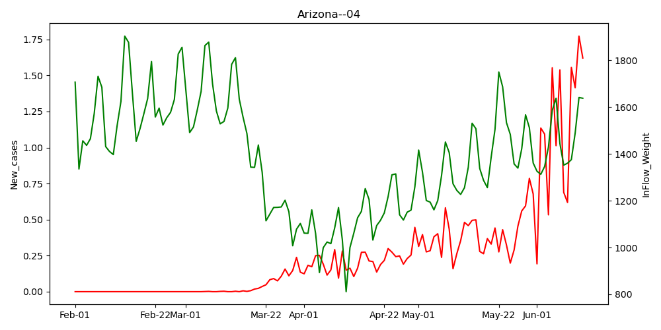
## *Activity Clustering:* The algorithm first applies HDBSCAN clustering algorithm to clusters all device observations into activity locations. This step takes the cleaned multi-day location data as input and applies an iterative algorithm until no cluster has a radius larger than two miles. The iterative algorithm consists of two parts: HDBSCAN based on a minimum number of point parameters and filtering non-static clusters based on time and speed checks. After finalizing the potential stay clusters, the algorithm combines nearby clusters to avoid splitting a single activity.

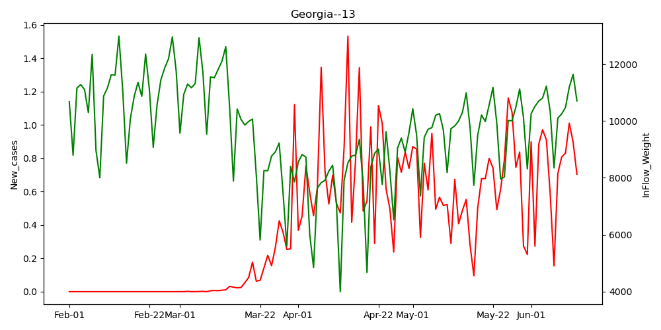
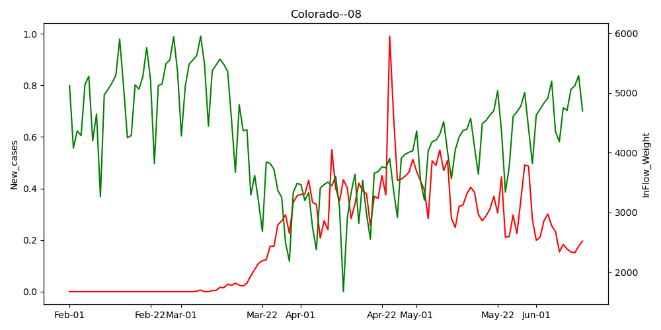
***Home and Work Identification:*** A typical methodology for identifying home and work clusters is to identify the most frequently visited clusters during the night and during the day. Here, instead of setting a fixed time period for each type, e.g., 8pm to 8am as the study period for home CBG identification and the other half day for work CBG identification, the framework examines both temporal and spatial features for the entire activity location list. The benefits are two-fold: the results for workers with flexible or opposite work schedules would be more accurate and the employment type for each device could be detected simultaneously.

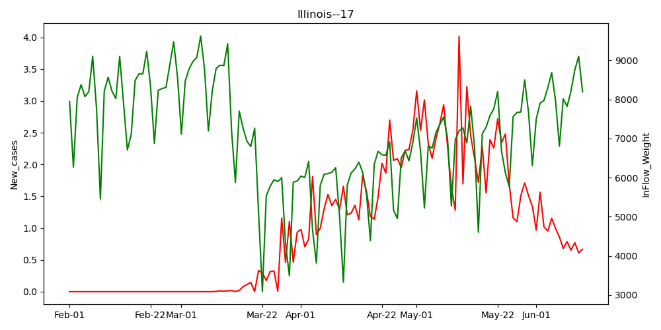
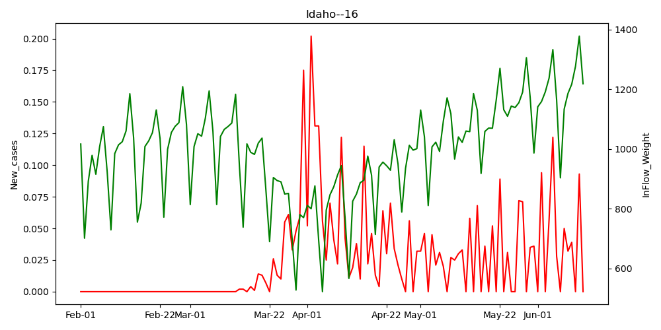
***Weighting*:** The sample data needs to be expanded to produce population-level statistics. The devices available in our dataset represent a sample of the population, so device-level weights are needed to expand the device sample. Also, for an observed device, only a sample of all trips may be recorded, so trip-level weights are needed as well. For the sake of timeliness, we have applied simple weighting methods to obtain county-level device weights and state-level trip weights. In order to obtain device-level weights, we have used the home county, obtained from the imputed home CBG information. The weight for each device is equal to the number of devices observed in the device’s imputed home county divided by the population of the county, so all devices residing in a county would have the same device-level weight. For instance, if our sample includes 100 devices in a county with a population of 2,000, each device would be assigned a weight of 20. For trip-level weights, we have calculated number of trips per person (trip rate) for each state during an average weekday in the first two weeks of February from our sample. We have also calculated this trip rate number for each state from the 2017 National Household Travel Survey. We have used a single trip rate for all trips generated from each state, equal to the NHTS trip rate divided by our observed trip rate.

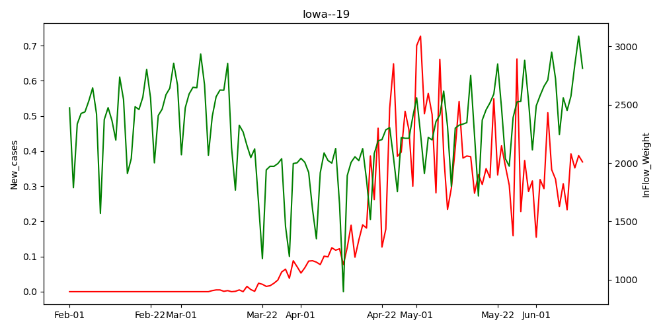
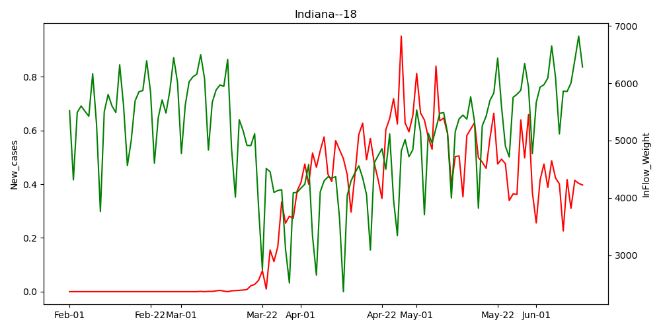
1. **Number of New cases and Inflow varying in reopen states**

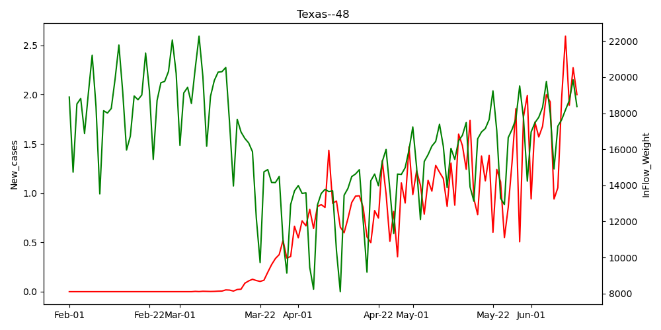
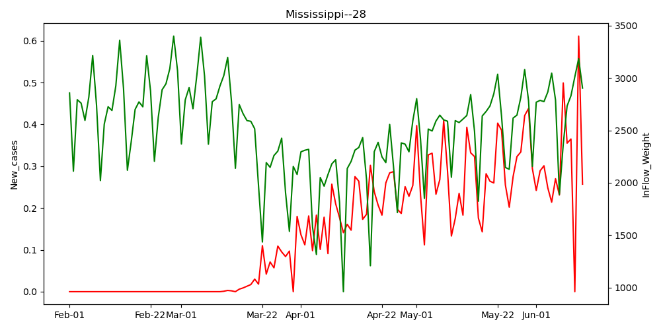
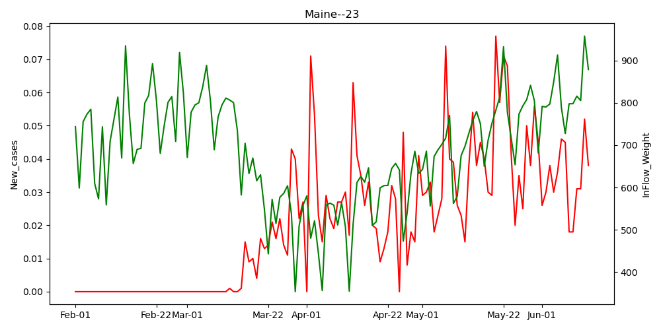
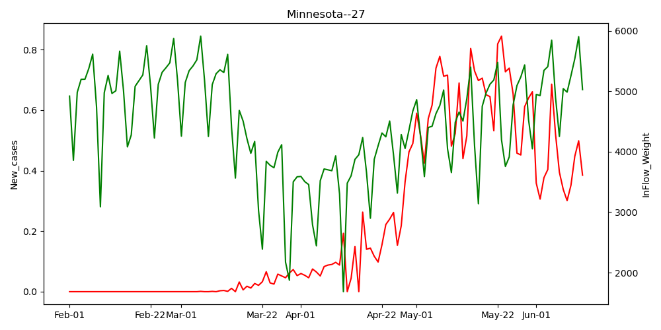
19 states are included in the group named reopen states, including Alabama, Arizona, Colorado, Georgia, Idaho, Illinois, Indiana, Iowa, Maine, Minnesota, Mississippi, New Mexico, North Dakota, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, and Utah, based on whether the partial reopen orders 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.

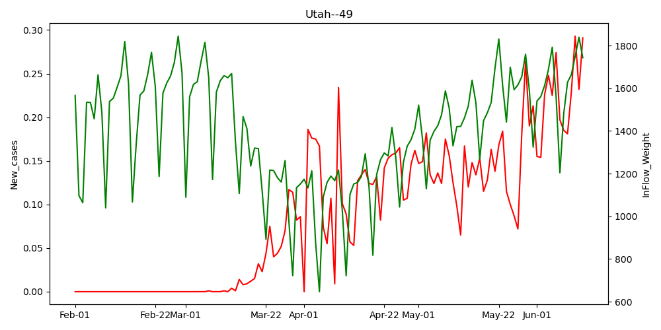
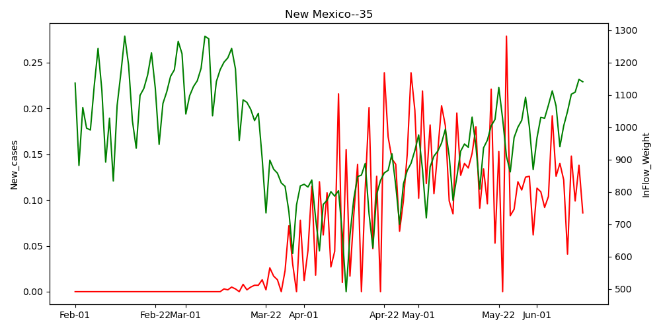
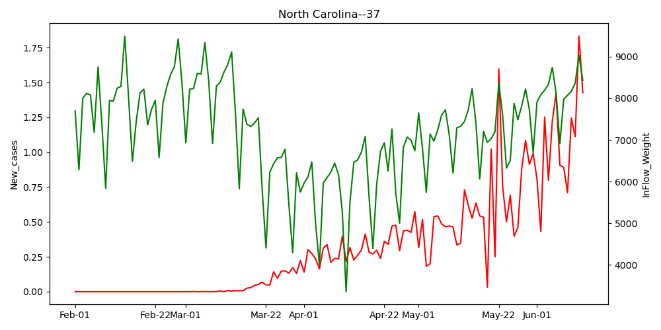
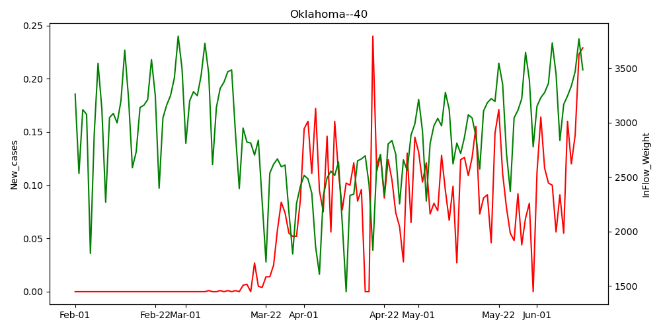
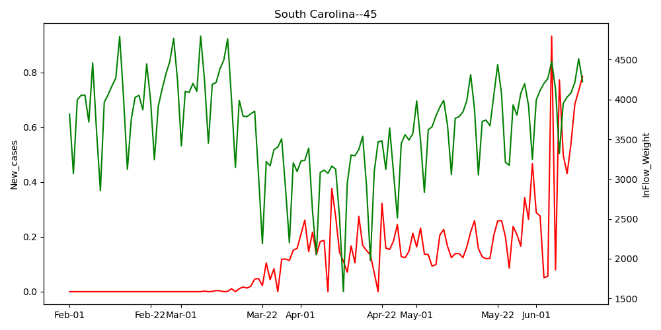
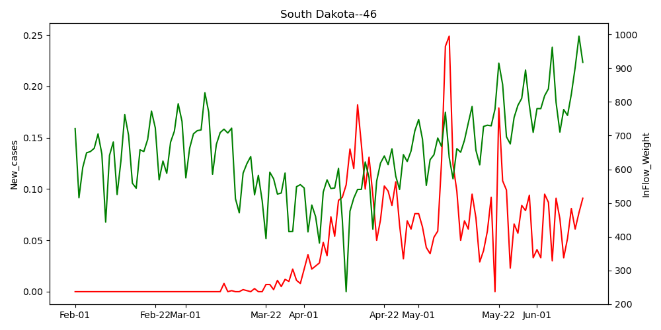
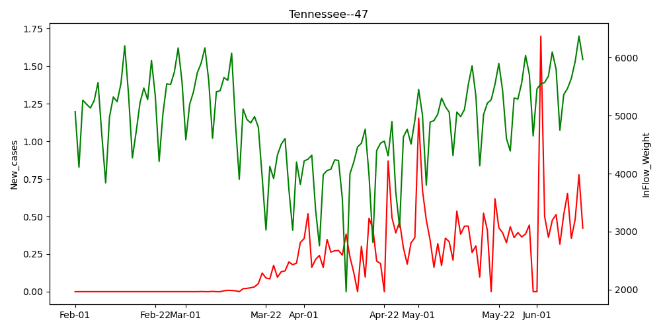
 







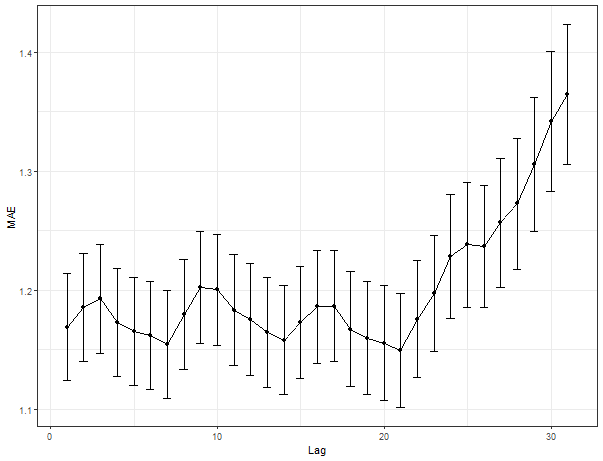




**Figure 3 The evolution of mobility inflow and number of daily confirmed cases in reopened states.**

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 4 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.

Table 2 ANOVA test between SEM with/without mobility feature

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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.

Table 3 Summary of model coefficients

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 5 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 6 BSTS results in New York County.**

A close up of a map

Description automatically generated

**Figure 7 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

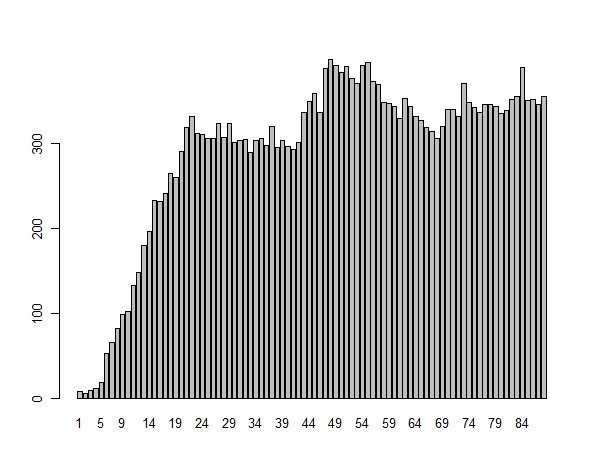
Description automatically generated

**Figure 8 The relationship between mobility inflow (weighted by cases) and confirmed cases in each county**

1. **Stationary Test**

As for stationary, we test it from two aspects. For the entire time-series, we test the stationary (Dickey–Fuller test) and found the series is significantly non-stationary (P-value = 0.5022, small P-values suggest that the data is stationary). However, when we take a lens to the time-series in each bandwidth (i.e. 7-day time-series), most of series become significantly stationary (90.64 % of time-series with P-value smaller than 0.05). The reason is that in a short interval, long-term trends and seasonal patterns (e.g. the weekly pattern, the monthly patterns) are largely eliminated. These arguments further confirm the reliability of dynamic coefficients across the entire study period and the robustness of the coefficient across each bandwidth.

We summarized the number of stationary time-series during different time periods and counties. We found the number of non-stationary time-series is small at first and then reach a plateau. The reason may because at first many counties do not have an explosive increase in number of cases. As for counties, we found counties with severe COVID-19 infections present more non-stationary time-series. For example, counites in New York like Rockland, New York, and Nassau present the greatest amount of non-stationary time-series across the nation.



**Figure 9 Number of non-stationary time-series across different time periods. X: Time index (Start from 2020/03/10); Y: Number of non-stationary time-series.**