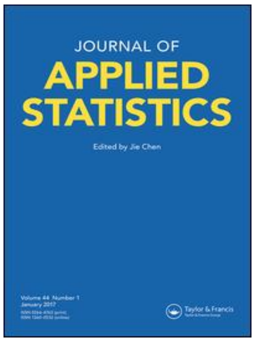


Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Comparing and integrating US COVID-19 data from multiple sources with anomaly detection and repairing

Guannan Wang, Zhiling Gu, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, Lei Gao & Li Wang

To cite this article: Guannan Wang, Zhiling Gu, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, Lei Gao & Li Wang (2021): Comparing and integrating US COVID-19 data from multiple sources with anomaly detection and repairing, Journal of Applied Statistics, DOI: 10.1080/02664763.2021.1928016

To link to this article: https://doi.org/10.1080/02664763.2021.1928016

Published online: 23 May 2021.

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2021.1928016

Taylor & Francis
Taylor & Francis Group

Comparing and integrating US COVID-19 data from multiple sources with anomaly detection and repairing

Guannan Wang [a], Zhiling Gu [b], Xinyi Li [c], Shan Yu [d], Myungjin Kim [b], Yueying Wang [b], Lei Gao [b] and Li Wang [b]

[a]College of William and Mary, Williamsburg, VA, USA; [b]Iowa State University, Ames, IA, USA; [c]Clemson University, Clemson, SC, USA; [d]University of Virginia, Charlottesville, VA, USA

ABSTRACT

Over the past few months, the outbreak of Coronavirus disease (COVID-19) has been expanding over the world. A reliable and accurate dataset of the cases is vital for scientists to conduct related research and policy-makers to make better decisions. We collect the United States COVID-19 daily reported data from four open sources: the New York Times, the COVID-19 Data Repository by Johns Hopkins University, the COVID Tracking Project at the Atlantic, and the USAFacts, then compare the similarities and differences among them. To obtain reliable data for further analysis, we first examine the cyclical pattern and the following anomalies, which frequently occur in the reported cases: (1) the order dependencies violation, (2) the point or period anomalies, and (3) the issue of reporting delay. To address these detected issues, we propose the corresponding repairing methods and procedures if corrections are necessary. In addition, we integrate the COVID-19 reported cases with the county-level auxiliary information of the local features from official sources, such as health infrastructure, demographic, socioeconomic, and environmental information, which are also essential for understanding the spread of the virus.

ARTICLE HISTORY
Received 12 August 2020
Accepted 1 May 2021

KEYWORDS
Anomaly detection; Coronavirus; count time series; data comparison; data integration; outlier correction

1. Introduction

Since the first infected case reported in December 2019, the outbreak of Coronavirus disease (COVID-19) has unfolded across the globe. In the US, coronavirus has infected more than five million people and killed over 160,000 people, as of the time of writing. While essential public health, economic and social science research in measuring and modeling COVID-19 and its effects is underway, reliable and accurate datasets are vital for scientists to conduct related research and for governments to make better decisions [5]. Unfortunately, errors could occur in the data collection process, especially under such a pandemic. In this work, we focus on data collection, comparison, data inconsistency detection, and corresponding curating.

CONTACT Li Wang ✉ lilywang@iastate.edu

Living through unprecedented times, governments must rely on timely, reliable data to make decisions to mitigate harm and support their citizens. Every day, several volunteer groups and organizations work very hard on collecting data on COVID-19 from all the counties and states in the US. There are four primary sources, including (1) the New York Times (NYT) [11], (2) the COVID Tracking Project at the Atlantic (Atlantic) [16], (3) the data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) [18], and (4) USAFacts [27]. Although these sources usually obtain their confirmed infectious and death cases data from the government agencies, the counts still vary due to the time of their collection as well as several other issues. However, these differences can be critical for real-time analysis. In this work, we first collect and compare the COVID-19 daily reported data from the above four open resources.

We observe a 7-day cyclical pattern for daily new cases and new deaths at the state and national level in the US. To test if the observed patterns are not accidental, we conduct a seasonality hypothesis test at the county, state and national level for the infected and death count time series from 15 March to 25 July 2020.

The COVID-19 data pose unique data quality challenges due to its spatiotemporal nature, and the problem of delayed-reporting and under-reporting. In this paper, we provide some anomaly and outlier detection techniques in the context of time series. After the anomaly detection, we explore various methods to repair the problematic data. To be more specific, the entire data cleaning procedure has been divided into two categories: (1) manual cleaning, and (2) automatic cleaning. On the one hand, manual cleaning has very high accuracy; on the other hand, it is challenging to implement due to the high cost in time and effort. In this paper, we propose some data repairing methods to address the aforementioned issues. We summarize the background of these methods and give details on the implementation of the repairing procedure for COVID-19 reported data with manual and/or automatic cleaning methods. Although other researchers also mentioned some similar data problems for COVID-19 in the literature, to the best of our knowledge, our work is the first one that focuses on how to address these issues and repair the COVID-19 data.

Furthermore, it has been observed that the local characteristics, such as socioeconomic inequity, may also contribute to the spread of epidemic [1,15]. For example, the intrinsic local community characteristics might influence and shape the spread of COVID-19, such as mobility, demographics, and socioeconomic status. The availability of census data thus leads us to include all the epidemic data, control measures, and local information while modeling the infections, deaths, and recoveries. To facilitate research in identifying the significant factors that affect the disease spread pattern and predict future infections and deaths, we also collect and combine local auxiliary information at the county level in the US from reliable sources.

To help users better visualize the epidemic data, we developed multiple `R` shiny apps embedded into a COVID-19 dashboard launched on 27 March 2020. Currently, we provide both infectious and death maps and time series of the US. Moreover, we provide a short-term (7-day) forecast [32] (updated daily) and a long-term (4-month) projection [33] (updated weekly) of the COVID-19 infected and death count at both the county level and state level. For public usage, a Github repository (https://github.com/covid19-dashboard-us/cdcar) is established to provide daily updated and cleaned data. An `R` package `cdcar` is also created for anomaly detection and repairing. In summary, we expect the proposed methods to have the following scientific merits. (i) Before choosing and integrating the data sources for analysis, it is important to understand how the data were collected and preprocessed. Therefore, in this article, we first investigate the similarities and dissimilarities among multiple data sources. (ii) Noticing the anomalies in the epidemic data, we develop several anomaly and outlier detection techniques in the context of count time series. Meanwhile, we further discover the reasons for these anomalies. (iii) After the anomaly detection, we introduce several methods to repair the problematic data and the corresponding historical data. Then, we obtain our database by integrating the cured data with many local characteristics. (iv) The proposed methods and the data are built into an R package, which is publicly available through GitHub.

The rest of the paper is organized as follows. Section 2 introduces the data related to the study of COVID-19, including a detailed description of the epidemic data, policy data, demographic characteristics, healthcare infrastructure, socioeconomic status, environmental factor and mobility data. Section 3 discusses the comparison of the epidemic data from different sources. Section 4 describes the cyclical pattern, types of anomalies of the COVID-19 reported time series, and how to perform the anomaly detection. Section 5 outlines methods for data repairing. Section 6 describes how to implement the proposed data comparison, anomaly detection and repairing procedure, and provide the details of the usage notes. Section 7 concludes the paper with a discussion.

2. Data

We collect the epidemic data up to county level in the US along with control measures and other local information, such as socioeconomic status, demographic characteristics, healthcare infrastructure, and other essential factors to analyze the spatiotemporal dynamic pattern of the spread of COVID-19. Our data covers about 3200 county-equivalent areas from 50 US states and the District of Columbia. A live version of the data analysis will be continually updated on our dashboard (https://covid19.stat.iastate.edu) and our Github repository (https://github.com/covid19-dashboard-us/cdcar). The sources and introductions for these data are detailed in Table 1.

2.1. Epidemic data

The daily counts of cases and deaths of COVID-19 are crucial for understanding how this pandemic is spreading. Thanks to the contribution of the data science communities across the world, multiple sources are providing the COVID-19 data with different precision and focus. In our article, we consider the reported cases from the following four sources: the NYT [11], the Atlantic [16], the JHU [18], and the USAFacts [27]. To clean the data, we first fetch data from the above four sources and compile them into the same format for further comparison. Then, we use the algorithms discussed in Section 4 to detect the anomalies in the data sources and choose the one with the least anomalies for further repair.

Table 1. Sources of datasets.

Data type	Source
COVID-19 Related Time-series	
Infections Data	[11,16,18,27]
Fatality Data	[11,16,18,27]
Recovery Data	[16]
Dates of COVID-19 Related Policies	
Declarations of State Emergency	[13]
Shelter-in-place or Stay-at-home Order	[9]
Mobility Data	
Bureau of Transportation Statistics	[17]
American Community Survey (ACS) Data	
2010–2018 Demographic and Housing Estimates	[22]
2005–2009 ACS 5-year Estimates	[20]
2012 Economic Census	[23]
2010 US Decennial Census	[21]
Homeland Infrastructure Foundation-level Data	[26]
USA Counties Database	[25]
US Census Bureau Gazetteer Files	[24]

2.2. Other factors

When analyzing the reported cases of COVID-19, many other factors may also contribute to the temporal or spatial patterns; see the discussions in [29]. For example, local features, like socioeconomic and demographic factors, can dramatically influence the course of the epidemic, and thus, the spread of the disease could vary dramatically across different geographical regions. Therefore, these datasets are also supplemented with the population information at the county level in our repository. We further classify these factors into the following six groups.

2.2.1. Policy data

In a race to stunt the spread of COVID-19, federal, state and local governments have issued various executive orders. Government declarations are used to identify the dates that different jurisdictions implemented various social distancing policies (emergency declarations, school closures, bans on large gatherings, limits on bars, restaurants and other public places, the deployment of severe travel restrictions, and ‘stay-at-home’ or ‘shelter-in-place’ orders). For example, former President Trump declared a state of emergency on 13 March 2020, to enhance the federal government response to confront COVID-19. Later in the past spring, at least 316 million people in at least 42 states, the District of Columbia and Puerto Rico were urged to stay home.

Since the late April 2020, all 50 states in the US began to reopen successively, due to the immense pressures of the crippled economy and anxious public. A state is categorized as ‘reopening’ once its stay-at-home order lifts, or once reopening is permitted in at least one primary sector (restaurants, retail stores, personal care businesses), or once reopening is permitted in a combination of smaller sectors. We compiled the dates of executive orders by checking national and state governmental websites, news articles, and press releases.

2.2.2. Demographic characteristics

In the demographic characteristics category, we consider the factors describing racial, ethnic, sexual, and age structures. These variables are extracted from the 2010 Census

[21], and 2010–2018 American Community Survey (ACS) Demographic and Housing Estimates [22].

2.2.3. Healthcare infrastructure

We also incorporate several features related to the healthcare infrastructure at the county level in the datasets, including the percent of persons under 65 years without health insurance, the local government expenditures for health per capita, and total bed counts per 1000 population.

2.2.4. Socioeconomic status

We consider diverse socioeconomic factors in the county level datasets. All of these factors collected from 2005–2009 ACS 5-year estimates [20].

2.2.5. Environmental factor

We also collect environmental factors that might affect the spread of epidemics significantly, such as the urban rate and crime rate.

2.2.6. Mobility

Another category of factors in the literature that affects the spread of infectious diseases significantly is the mobility; for example, movements of people from neighborhoods. We collect the mobility data from the Bureau of Transportation Statistics.

2.3. Geographic information

The longitude and latitude of the geographic center for each county in the US are available in Gazetteer Files [24].

3. Comparison of the epidemic data

In this subsection, we assess the similarities and differences of the reported infection and death cases from the previously mentioned four sources. The data collection sources and release times are indicated for each of the sources to help determine which factors may have an effect on the outcome of the assessment. The NYT released daily data at the national, state, and county levels at noon of the following day before July 8, 2020, after which the release time changed to midnight. The Atlantic releases daily state-level data along with testing, hospitalization, and recovery information, updated in the afternoon of the following day. Since March 7, 2021, the Atlantic has stopped the daily data updating. The COVID-19 Data Repository by the CSSE at JHU provides both state and county-level data daily. JHU released data at midnight on and before April 22 and then changed the release time to the early morning of the following day. USAFacts collects the county-level data in the evening and releases them in the early morning of the following day (by 9 a.m. PST) [28]. Table 2 summarizes the differences among the four sources of data based on how the data are collected and compiled.

Let K be the number of all available sources in the comparison. For the county level comparison, $K = 3$ since the Atlantic does not provide county level data, while for the state level, $K = 4$. Let T be the number of days observed, or the length of each time

Table 2. A summary of the comparison among four sources.

Source	NYT	Atlantic	JHU	USAFacts
Infected & death*	1,2,3	1,2	1,2,3	1,2,3
Recovered	0	1,2	1,2,3**	0
Tested	0	1,2	1,2	0
Hospitalized	0	1,2	1,2	0
Islands***	2,3	2	2,3	0
Unallocated****	3	0	3	3
Place of infection#	*r*, *p*	*r*, *p*	Unknown	Unknown
Place of fatality	$r, r+p, p$	Unknown	Unknown	Unknown
Probable infected##	*y*	*y*	*y*	Unknown
Probable death	*y*	*y*###	*y*	Unknown

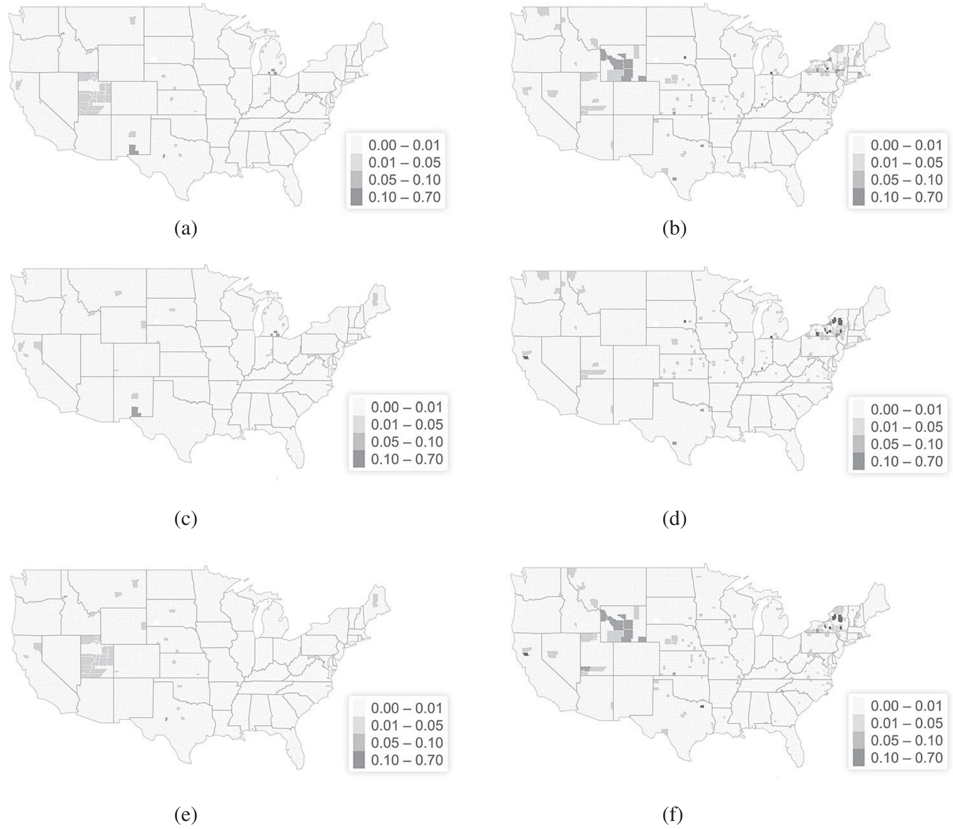
Note: *: Country Level = 1, State Level = 2, County Level = 3. USAFacts only provides county-level data for downloading. **: JHU pulls the number of people recovered data in the state-level from the Atlantic. ***: Whether the source includes Puerto Rico, American Samoa, Guam, Northern Mariana Islands, Virgin Islands. ****: Whether the dataset has unallocated/unassigned information, which is useful to match state-level and county-level data. #: How does the dataset assign the cases to a place. *p* indicates that the source assigns the counts according to the place of infection/fatality. $r+p$ indicates the source assigns both the deaths occur in the specific location, and the residents' deaths that occur outside the location. Specifically, this is related to New York City death data, see details in Section 3. $r, r+p, p$ indicates multiple standards exist. *unknown* indicates the information is not found. ##: Whether the dataset includes both confirmed and probable cases when probable data is available. *y* means yes. NYT releases daily live data for probable and confirmed cases separately, but historical data is unavailable. ###: Colorado started to report the number of deaths where COVID-19 is listed as a contributing cause on the death certificates since May 16. This number is significantly lower than deaths among infected. The Atlantic uses deaths where COVID-19 is listed on death certificates, while the other three sources use deaths among infected. This unveils different definitions of probable deaths applied by the four sources.

series. Let n be the number of counties or states. For source k, $k = 1, \ldots, K$, let $Y_{it}^{(k)}$ be the cumulative number of the reported cases of location i on day t, where $i = 1, \ldots, n$, $t = 1, \ldots, T$. In the following, we define a dissimilarity measure to assess the difference between two time series: $\mathbf{Y}_i^{(k)} = \{Y_{it}^{(k)}\}_{t=1}^{T}$ and $\mathbf{Y}_i^{(k')} = \{Y_{it}^{(k')}\}_{t=1}^{T}$, for any $1 \leq k \neq k' \leq K$. Let $\overline{Y}_{it} = K^{-1} \sum_{k=1}^{K} Y_{it}^{(k)}$, then the difference between $\mathbf{Y}_i^{(k)}$ and $\mathbf{Y}_i^{(k')}$ is defined as

$$d(k, k') \equiv d(\mathbf{Y}_i^{(k)}, \mathbf{Y}_i^{(k')}) := \begin{cases} \dfrac{1}{T} \|\mathbf{Y}_i^{(k)} - \mathbf{Y}_i^{(k')}\| / \overline{Y}_{iT}, & \overline{Y}_{iT} > 0, \\ 0, & \overline{Y}_{iT} = 0, \end{cases} \tag{1}$$

where $\overline{Y}_{iT}$ is used to mitigate the variability of the currently observed counts. Equation (1) provides a measurement that effectively detects the counties and states with the most discrepancy between each pair of sources and is meaningful in the comparison between different locations. In Figure 1, we present the county map for infected and death counts collected from three data sources. In Figure 2, we present the state map for infected and death counts collected from four data sources. Areas in dark shade in these two figures are determined to be different between the corresponding pair of two sources. In the rest of this section, we look further into the underlying reasons for dissimilarity at the county and state levels.

We list the ten most dissimilar counties in a pairwise comparison of the three sources, in terms of infection and death counts, in Tables 3 and 4, respectively. The specific reasons for these dissimilarities vary over locations. For the state of New York, the difference

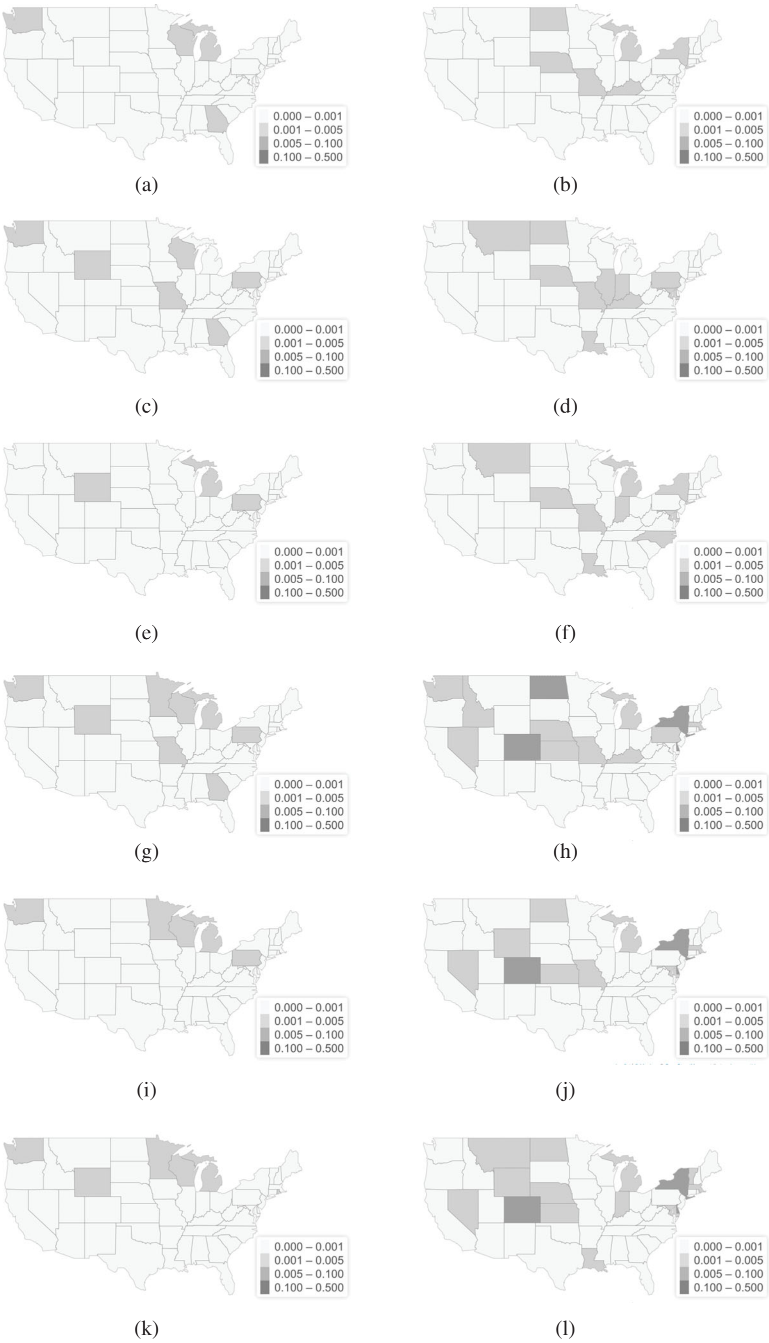


(a) (b) (c) (d) (e) (f)

Figure 1. County maps of the dissimilarity measure as of 25 July 2020. (a) Infection (NYT vs JHU), (b) Death (NYT vs JHU), (c) Infection (NYT vs USAFacts), (d) Death (NYT vs USAFacts), (e) Infection (JHU vs USAFacts) and (f) Death (JHU vs USAFacts).

between sources is caused by different geographical assignments. NYT and JHU combine Kings, Queens, Bronx, and Richmond counties with New York City while USAFacts does not use that combination. For the state of Utah, JHU combines counties to jurisdictions to be consistent with the official state source, while NYT and USAFacts provide county-level data. For Guam, NYT includes the data reported from USS Theodore Roosevelt, while JHU and USAFacts do not. In Michigan, NYT considers federal and state prison inmates' data when reporting at the county level, while the other two sources do not. For the state of Alaska, NYT and JHU include non-resident cases while USAFacts does not. For some states, such as Kentucky, Texas, Pennsylvania, Washington, Georgia and Tennessee, the official county-level data is subject to frequent adjustments, which can lead to discrepancies when one source corrects the errors while other sources do not. In summary, the county-level dissimilarities between data sources are mostly caused by different geographical rules, non-resident data, prison inmates data, and differed efforts in correcting the historical data.

Next, we look into the state-level comparison. According to our measure, using data up until 25 July 2020, states that show dissimilar infection data are illustrated in Figure 2. Here, we list out a few examples about how the dissimilarities arise. On the one hand,



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Figure 2. State maps of the dissimilarity measure as of 25 July 2020. (a) Infection (NYT vs JHU), (b) Death (NYT vs JHU), (c) Infection (NYT vs USAFacts), (d) Death (NYT vs USAFacts), (e) Infection (JHU vs USAFacts), (f) Death (JHU vs USAFacts), (g) Infection (NYT vs Atlantic), (h) Death (NYT vs Atlantic), (i) Infection (JHU vs Atlantic), (j) Death (JHU vs Atlantic), (k) Infection (USAFacts vs Atlantic) and (l) Death (USAFacts vs Atlantic).

Table 3. Top 10 counties with the largest value of the dissimilarity measure of the infectious counts between pairs of sources (as of 25 July 2020).

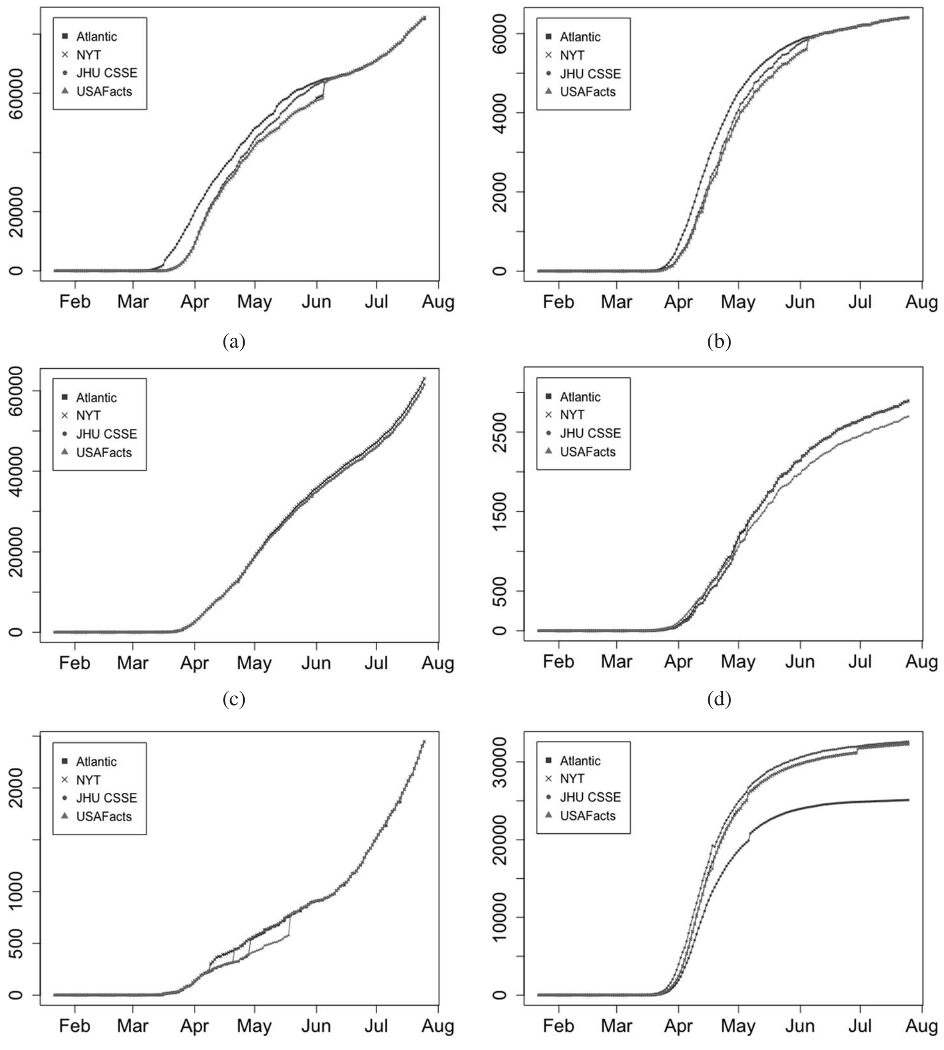
NYT vs JHU	NYT vs USAFacts	JHU vs USAFacts
BristolBay, AK	BristolBay, AK	Lewis, ID
Dillingham, AK	Dillingham, AK	Dukes, MA
Lewis, ID	Branch, MI	Bronx, NY
Dukes, MA	Jackson, MI	Kings, NY
Branch, MI	Otero, NM	New York, NY
Jackson, MI	Bronx, NY	Queens, NY
Lenawee, MI	Kings, NY	Richmond, NY
Otero, NM	New York, NY	Sterling, TX
Sterling, TX	Queens, NY	Emery, UT
Piute, UT	Richmond, NY	Piute, UT

Table 4. Top 10 counties with the largest value of dissimilarity measure of the death counts between pairs of sources (as of 25 July 2020).

NYT vs JHU	NYT vs USAFacts	JHU vs USAFacts
Crawford, IN	Glenn, CA	Glenn, CA
McLean, KY	Crawford, IN	Hamilton, NY
Branch, MI	Bronx, NY	Bronx, NY
Oswego, NY	Cortland, NY	Cortland, NY
Delaware, NY	Kings, NY	Lewis, NY
Seneca, NY	Lewis, NY	Queens, NY
Tompkins, NY	Schoharie, NY	Richmond, NY
Davison, SD	Seneca, NY	Schoharie, NY
Hopkins, TX	Tompkins, NY	Seneca, NY
Teton, WY	Davison, SD	Kings, NY

different responses to the change of probable cases reporting mechanisms in infections and/or deaths lead to discrepancies in the reported cases between the four sources. For instance, Wyoming started to include probable cases in their infected cases reporting during the week of April 6th. Each of the four sources responded to the change at different dates, as indicated by the jumps in data shown in Figure 3, with NYT being the first source to respond to the change; see Figure 3(e). Similarly, Michigan started reporting probable cases and deaths after April 5. This resulted in higher infection counts in the Atlantic for the following 3 months, probably due to a correction for probable cases; see Figure 3(a,b). As demonstrated in the time series plots of Indiana in Figure 3(c,d), inclusion or exclusion of probable cases also caused differences in reported cases of both infections and deaths among the four sources. Another difference among sources is caused by whether cases are reported according to the residence or the place of infection/death. For example, when reporting the death cases of New York, the Atlantic uses New York State reported deaths, while the other three sources use New York City reported deaths, which also reports deaths of residents that occur outside New York City. Starting August 6, NYT switched to reporting deaths by residence to make New York State death data consistent with the other states, which led to discontinuity on August 6. To summarize, the state-level dissimilarities are mainly caused by different report mechanisms to probable cases and varied choices of the geographical assignment.

Based on these examples, it is safe to conclude that the differences in reported cases do not indicate the inferiority or superiority of the source per se. No matter which source we use, we need to be clear and careful about the processing behind it. Generally speaking,



(a) (b)

(c) (d)

(e) (f)

Figure 3. Examples of infection and death time series up until 25 July 2020. (a) Michigan infection, (b) Michigan death, (c) Indiana infection, (d) Indiana death, (e) Wyoming infection and (f) New York death.

despite the geographical rules, USAFacts tends to be more conservative because it reports confirmed counts instead of the sum of confirmed and probable counts in several states, and NYT tends to report higher county-level counts by including non-resident data and prison inmates data.

4. COVID-19 time series: features and anomaly detection

4.1. A 7-day cycle in infection and death cases

We observe a 7-day cycle in (i) reported COVID-19 new cases, and in (ii) reported COVID-19 new deaths at the national and state level. To rigorously test the 7-day cycle, we conduct the hypothesis test using the R package seastests (function isSeasonal) [12]. By default, it implements the 'WO-test', an overall seasonality test scheme proposed in [35]. Given a set of various seasonality tests, the WO-test first conducts a recursive feature elimination algorithm in conditional random forests to identify the most informative candidate tests. Then the p-values from the selected tests are used again as predictors to grow a single conditional inference tree. The candidate tests may include a variety of seasonality tests tailored to particular manifestations, such as the modified QS test, the Friedman test, the Kruskal–Wallis test, the periodogram test, and the Welch test.

We first conduct seasonality tests on the time series of national confirmed cases and deaths. Both time series show a 7-day seasonal behavior with p-values less than 10^{-7} from a variety of tests, including the QS test, the Friedman test, and the Kruskal–Wallis test. Then, we compare the different tests on state-level infected cases and deaths. The results are summarized in Table 5, where we use a checkmark to signify a significant test result at significance level of 0.05. For the infected cases, all the tests suggest a significant 7-day cyclical pattern in 22 states. Meanwhile, we observe 27 states for which all the tests suggest a significant 7-day cyclical pattern on the death time series.

The cyclical pattern is less evident at the county level. After applying the Friedman test to more than 3000 counties, 701 counties show the 7-day cyclical pattern in infected cases with p-value smaller than 0.05. For the deaths, only 200 counties exhibit a cyclical behavior with p-value smaller than 0.05.

The stacked column bar plots in Figure 4 show the day on which the most infection/death cases are reported for each week in each state. From Figure 4(a), we can observe that in the states, such as Connecticut, Missouri Oklahoma, and Washington, Tuesdays usually report the most infections. States such as Illinois and Wisconsin, report more infection cases on Fridays. For death cases, many states reach their peak on Tuesdays; see Figure 4(b) for more details.

The stacked column bar plots in Figure 5 illustrate the day on which the highest infection/death cases are reported for each week from 15 March to July 25, 2020. From Figure 5(a), one can see that in the early stage of the pandemic (before April), Tuesdays usually had the highest number of infections. As seen in Figure 5(b), Sundays and Mondays typically reported a smaller number of death counts than the other days. Meanwhile, the peaks often occurred on either Tuesdays or Wednesdays. In time series analysis, for a short-term forecast, a few approaches can be considered to remove the weekly cycle: simple differencing ($Y_t - Y_{t-7}$); 7-day moving average; using dummy variables control for day-to-day variation; harmonic function (series of sine/cosine functions).

4.2. Anomaly detection

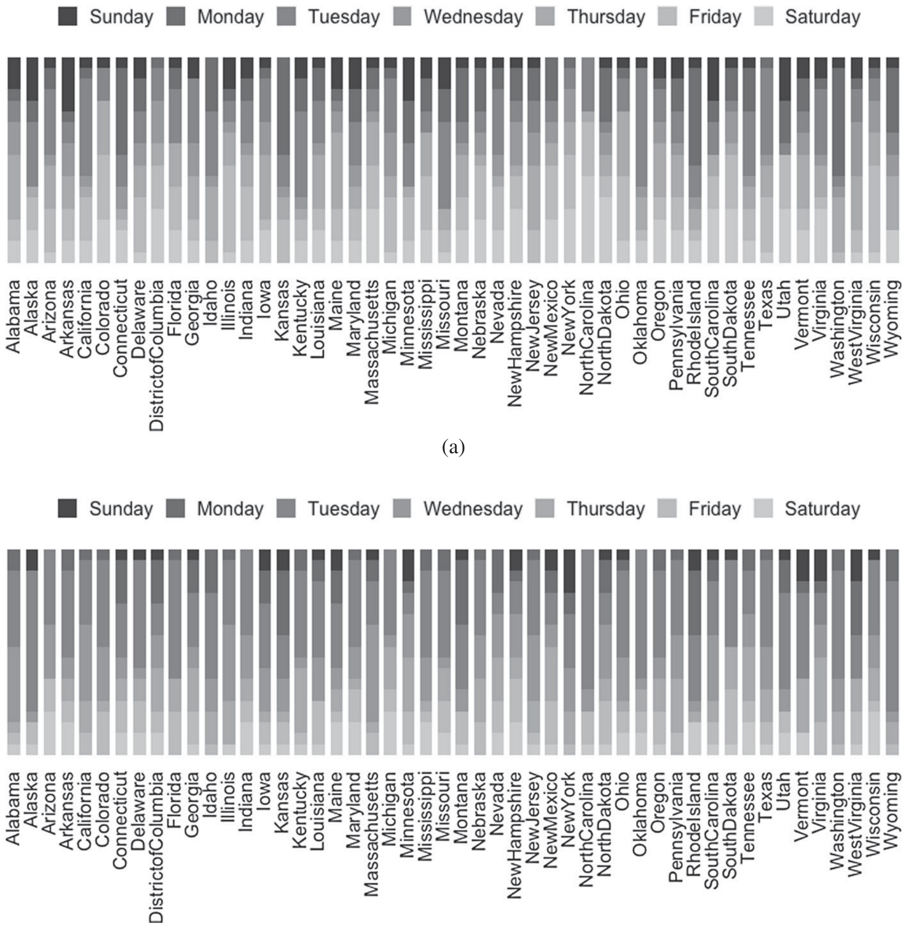
In addition to the exciting findings aforementioned in the raw data comparison, we observe two major types of anomalies in the data: (I) order dependencies violation, and (II) point anomalies. Examples of these two issues are illustrated in Figure 6. Before conducting any analysis of the epidemic data, one might need to account for these issues. In this section, we use the epidemic data from NYT as an illustration, but all four data sources exhibit similar issues.

Table 5. The detection of the 7-day cycle on the state-level times series using different tests, including the modified QS test (QS), the Friedman test (Fried), the Kruskal–Wallis test (KW), the Welch test (Welch), and the WO-test (WO) ('√' indicates that the p-value is less than 0.05).

	Infection					Death				
State	QS	Fried	KW	Welch	WO	QS	Fried	KW	Welch	WO
Alabama	√				√	√	√	√	√	√
Arizona	√	√	√	√	√	√	√	√	√	√
Arkansas	√		√	√		√		√	√	√
California	√	√	√	√	√	√	√	√	√	√
Colorado		√	√	√	√	√	√	√	√	√
Connecticut	√				√	√	√	√		√
Delaware							√			
Florida	√	√	√	√	√	√	√	√	√	√
Georgia	√	√	√	√	√	√	√	√	√	√
Idaho	√	√	√	√	√	√		√	√	√
Illinois		√	√	√	√	√	√	√	√	√
Indiana	√	√	√	√	√	√	√	√	√	√
Iowa	√	√	√	√	√	√	√	√	√	√
Kansas	√	√	√	√	√	√				
Kentucky		√	√		√	√	√	√	√	√
Louisiana		√	√			√	√	√	√	√
Maine	√	√	√	√	√		√	√	√	
Maryland	√	√	√	√	√	√	√	√	√	√
Massachusetts	√	√	√	√	√	√	√	√	√	√
Michigan		√	√	√	√	√	√	√	√	√
Minnesota	√	√	√	√	√	√	√	√	√	√
Mississippi		√	√	√	√	√	√	√	√	√
Missouri	√		√	√	√	√	√	√	√	√
Montana	√				√		√			
Nebraska	√	√	√	√	√	√	√	√	√	√
Nevada		√	√	√		√	√	√	√	√
New Hampshire		√	√	√	√	√	√	√	√	√
New Jersey							√	√	√	√
New Mexico		√	√	√	√		√	√	√	√
New York	√	√	√	√	√					
North Carolina	√	√	√	√	√	√	√	√	√	√
North Dakota		√	√	√	√	√	√	√		
Ohio	√	√	√	√	√	√	√	√	√	√
Oklahoma	√	√	√	√	√	√	√	√	√	√
Oregon		√				√	√	√	√	√
Pennsylvania	√	√	√	√	√	√	√	√	√	√
Rhode Island	√			√	√	√				√
South Carolina		√	√	√	√		√	√	√	√
South Dakota	√	√	√	√	√		√	√	√	√
Tennessee	√			√	√	√	√	√	√	√
Texas	√	√	√	√	√	√	√	√	√	√
Utah	√	√	√	√	√	√		√	√	√
Vermont							√			
Virginia	√	√	√			√	√	√	√	√
Washington	√	√	√	√	√		√	√	√	√
West Virginia	√	√	√		√					
Wisconsin	√	√	√	√	√	√	√	√	√	√
Wyoming			√							
District of Columbia	√							√		

4.2.1. Order dependencies violation

Order dependency (OD) is widely used in the relational database. In this project, we incorporate this concept into the anomaly detection and data repairing process of cumulative



(a)

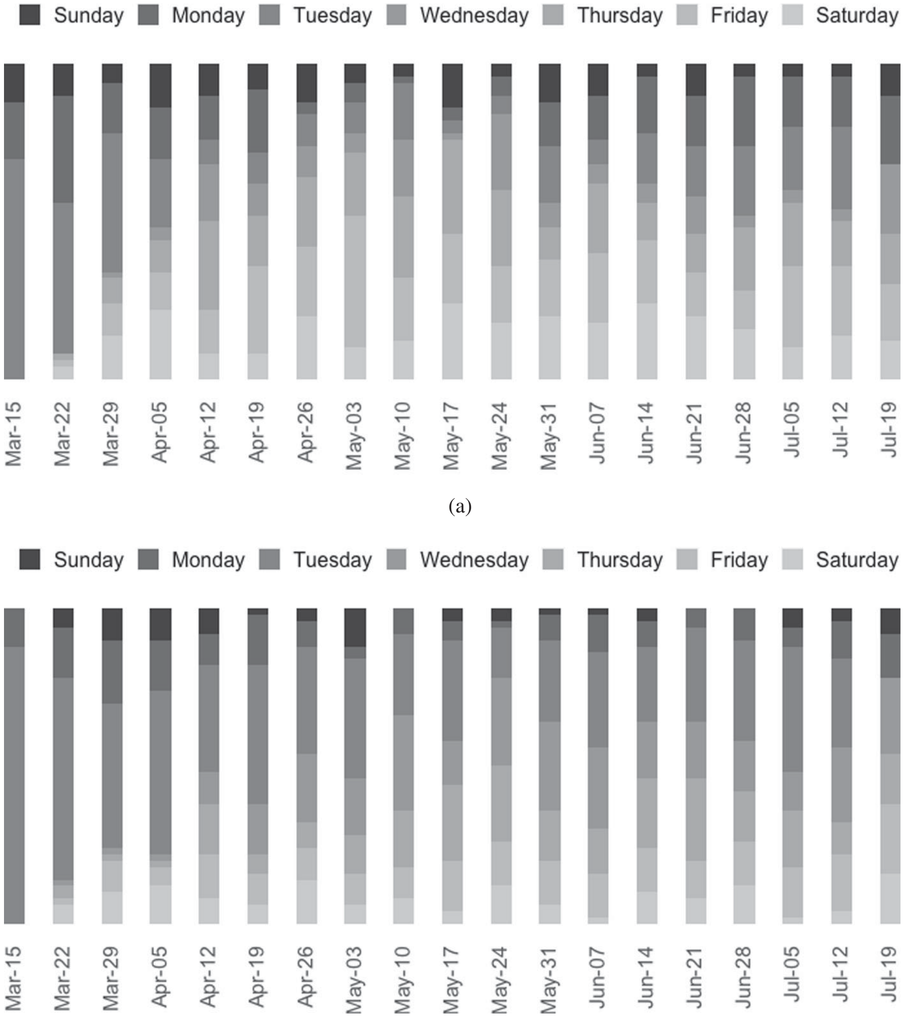
(b)

Figure 4. The 100% stacked column bar plot of the number of weeks that reaches the weekly maximum of the infection or death counts across days of the week in different states. (a) Infection and (b) Death.

time series. To be more specific, the OD for the cumulative time series can be defined as follows: for any two time points, t_1 and t_2, if $t_1 < t_2$, then $Y_{t_1} \leq Y_{t_2}$, where Y_t represents the cumulative infection/death count on day t. Obviously, the time series in Figure 6(a) violates the OD.

4.2.2. *Point anomalies*

A point anomaly refers to the situation where there is 1 day of an abrupt increase in the cumulative or daily new time series. This can also be considered as a violation of speed constraints. The anomaly can be caused by a number of factors, including (1) the result of a large batch of tests was released and (2) the change of reporting standard, such as some states starting to report probable cases from a specific date. For example, the anomaly in Figure 6(c) is due to New Jersey reporting 1854 probable deaths that may date back to earlier in the outbreak. The anomaly in Figure 6(d) might be related to the reported cases in the Texas Department of Criminal Justice (TDCJ). On May 31, 2020, at least 82 of the



(a)

(b)

Figure 5. The 100% stacked column bar plot of the number of states that reaches the weekly maximum of the infection or death counts across days of the week in different weeks. (a) Infection and (b) Death.

active cases in the county were TDCJ-related. To detect this type of anomaly, we exam the increasing speed of the time series. Specifically, if the time series Y_{t_i} satisfies some speed constraint (SC), then,

$$(Y_{t_2} - Y_{t_1})/(t_2 - t_1) < SC_1 \quad \text{and} \quad \frac{(Y_{t_1+1} - Y_{t_1})}{(Y_{t_2} - Y_{t_1})/(t_2 - t_1)} < SC_2,$$

where SC_1 and SC_2 are two predetermined thresholds.

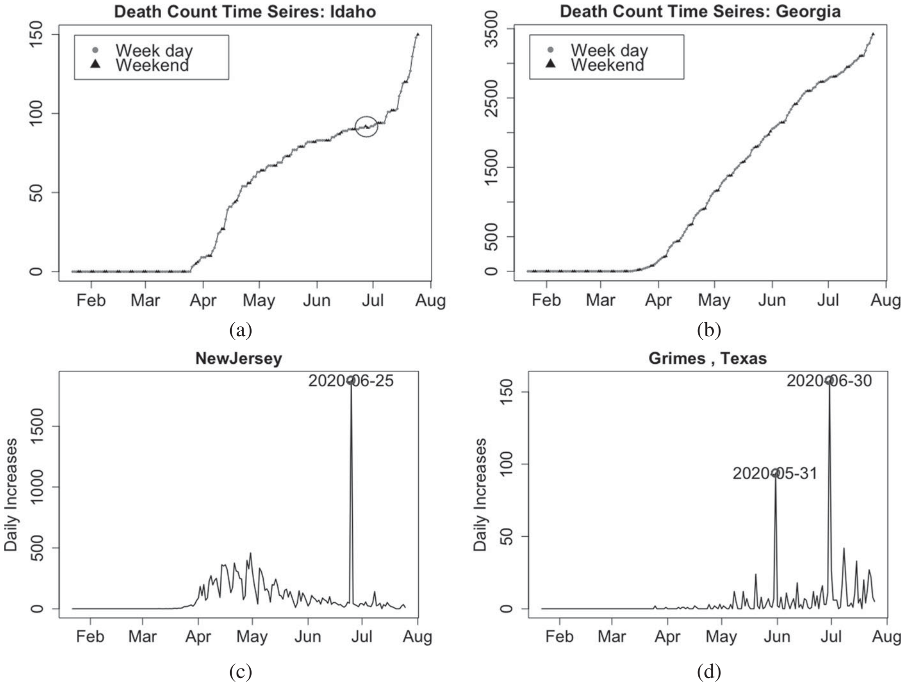
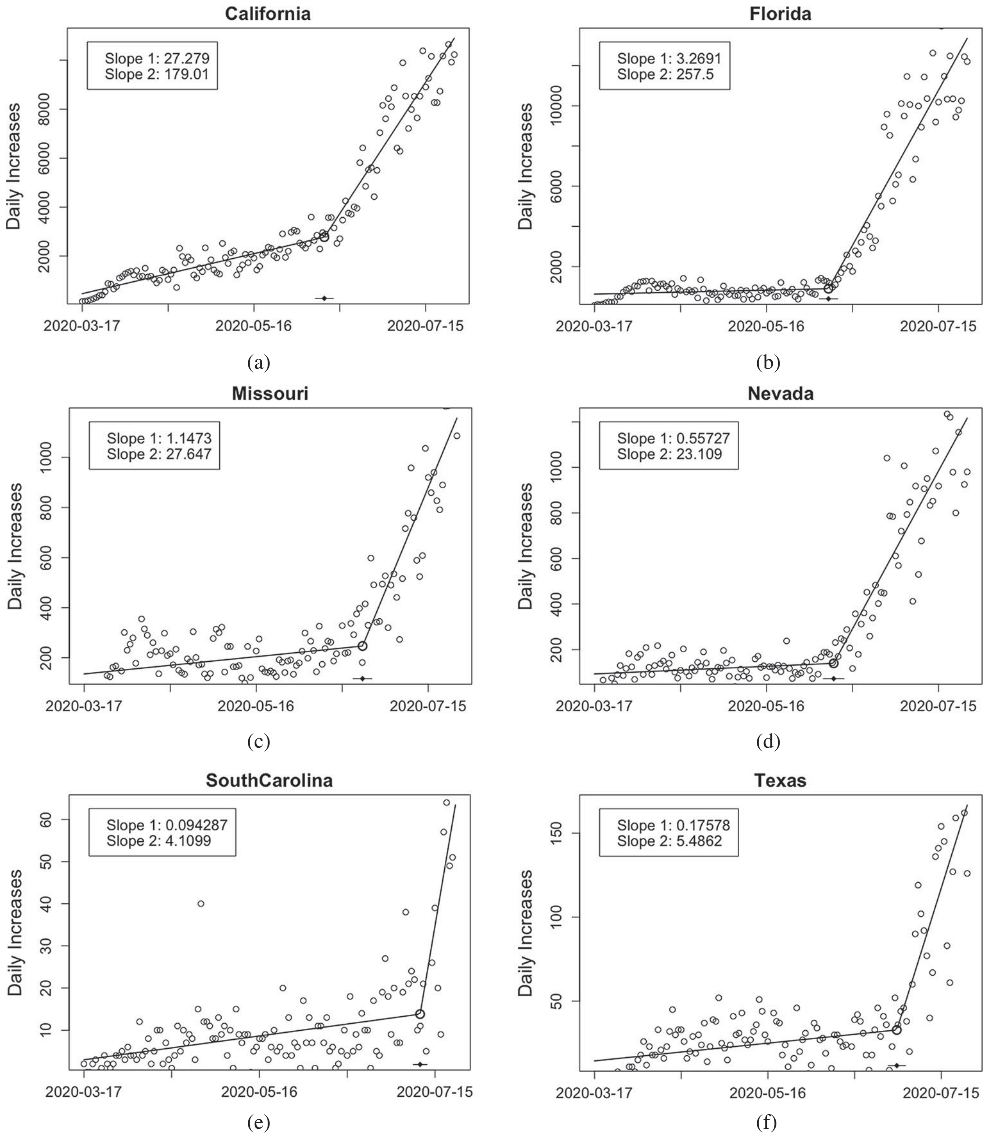


Figure 6. An illustration of different types of anomalies. (a) Order dependency; (b) weekend/holiday delay-reported issue; (c) single point anomaly; (d) two-point anomalies.

4.3. Change points in time series

Sometimes, we may experience a pattern change in the time series, which can be referred to as the period when the increasing speed is significantly different from the previous period. We apply the function `segmented` in `R` package `segmented` [10] to detect the change points. This function implements the segmented models in which the relationship between response and covariate(s) is modeled as piecewise linear segments connected at some joint points (or change points). Once a change point is detected, we will provide a warning message to let the user decide whether any repair is necessary.

For simplicity, we consider a segmented relationship between the response and the time. Let $\mu_i = E(Y_{t_i})$, and the variable (time) t_i is modeled by $g(\mu_i) = \beta_1 t_i + \beta_2 (t_i - \phi)_+$, where $g(\cdot)$ is the link function, $(t_i - \phi)_+ = (t_i - \phi) I(t_i > \phi)$, and $I(\cdot)$ is the indicator function. Here, β_1 illustrates the slope of left line segment and β_2 represents the difference-in-slopes. The main idea lies in testing whether $|\beta_2| > 0$. If a break-point does not exist, the difference-in-slopes parameter has to be zero. Table 6 presents the states in which change points are detected in daily new infections and deaths, and the dates of the change points are identified. Figure 7 visualizes the identified change points together with the time series of daily new infections and deaths. Based on the change point analysis, most of the changes occurred in June and July with sudden increases in incident cases and death.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7. Time series plots of the states with change points identified, where the small circle represents the daily observed value, and the big circle indicates the change point detected, and the segment shows the linear regression line before and after the change point. (a) New infection cases in California; (b) new infection cases in Florida; (c) new infection cases in Missouri; (d) new infection cases in Nevada; (e) new deaths in South Carolina; (f) new deaths in Texas.

5. Data repairing

Once raw data is collected, we start with the OD violation detection and repairing. Next, we check for the point anomalies, and let the user decide whether to repair it. Last, we investigate the weekly cycles and pattern changes in the time series. In this section, we propose several data repairing methods to handle the issues mentioned in Section 4. To

Table 6. A list states with change-point identified in the daily new infected cases and deaths.

Infection		Death	
State	Change Point	State	Change Point
California	2020-06-10	South Carolina	2020-07-13
Florida	2020-06-07	Texas	2020-07-01
Missouri	2020-06-23		
Nevada	2020-06-09		

resolve these issues, we focus on the daily new infected/death cases instead of the cumulative infected/death cases. In the following, we let $Z_t = Y_t - Y_{t-1}$ be the increase at time point t.

5.1. Anomaly repairing

First of all, the daily reported infected/death count could be considered as a count time series by nature. Therefore, when repairing a count time series, we need to take into account that the observations are nonnegative integers, and we should utilize the dependence structure among observations. Furthermore, in the study of the infectious disease, the population is usually assigned to compartments such as Susceptible (S), Infectious (I), or Recovered (R), and people may progress between compartments. Therefore, different compartments are usually considered as an entire system and are studied together; see for example, the SIR models [3,6,14]. Third, the spread of the disease also has a spatial pattern. In general, once a point anomaly is detected, we let $\mathcal{A} = \{t \in \mathcal{T} : Z_t \text{ is identified as a point anomaly}\}$. For $t \in \mathcal{A}$, the user can decide whether a correction is necessary. If so, the user can choose from the following methods to obtain a more reliable value, $\widehat{Z}_t$, to replace the point anomaly Z_t.

5.1.1. Time series model for count data

One of the conventional methods to deal with these challenges is the generalized linear model (GLM), which models the observations conditionally on past information. In this project, we consider both Poisson and Negative Binomial as the conditional distribution. The second important class for analyzing count time series is the integer autoregressive moving average models, and a comprehensive review is given by [36]. The state-space is another type of count time series models. Comparing with the GLM, it allows a more flexible data generating process. However, it requires a more complicated model specification. Due to the explicit formulation, the GLM-based models yield a more convenient way to make predictions. Thus, in this project, we focus on the GLM-based method.

To repair the dataset, we model the conditional mean $\mu_t = \mathrm{E}(Z_t | Z_{t-1}, \mu_{t-1})$ in the following form:

$$\nu_t = \beta_0 + \sum_{k=1}^{p} \beta_k Z_{t-k} + \sum_{l=1}^{q} \alpha_l \nu_{t-l},$$

where $\nu_t = \log(\mu_t)$.

For this type of data repairing, we use the R package tscount [7], which conducts a model estimation by the quasi-conditional maximum likelihood method (function tsglm).

5.1.2. Combined linear and exponential predictors

The second method we consider is similar to the combined linear and exponential predictors models proposed in [2], which assembles the following three different models.

(1) An individual county-/state-level exponential predictor: model (2) uses a series of separate predictors for each county to capture the reported exponential growth of COVID-19 infected and death counts, and we assume

$$\log\{\mathrm{E}(Z_t|t)\} = \beta_0 + \beta_1 t, \tag{2}$$

where the parameters β_0 and β_1 are the coefficients in the generalized linear model (GLM) using glm function in R with a log link function.

(2) An individual county-/state-level linear predictor: model (3) fits a linear version of the separate county predictors; specifically, we assume that

$$\mathrm{E}(Z_t|t) = \beta_0 + \beta_1 t. \tag{3}$$

(3) An individual county-/state-level exponential epidemic predictor: model (4) uses a series of disease related factors to capture the reported exponential growth of COVID-19 infectious and death counts. We assume that

$$\log\{\mathrm{E}(Z_t|Z_{t-1})\} = \beta_0 + \beta_1 \log(Z_{t-1} + 1). \tag{4}$$

5.1.3. Spatio-temporal epidemic model

Based on the idea of the SIR models, this paper [34] proposes the discrete-time spatial epidemic model, which combines the susceptible state, infectious state, and removed state together. In the following, we denote I_{it}, D_{it}, and R_{it} the cumulative number of cases in infected, death and recovered states, respectively, in county i observed on day t. Let $\mu^I_{i,t}$, $\mu^D_{i,t}$, $\mu^R_{i,t}$ be the conditional mean value of daily new positive cases, deaths and recovery cases, respectively, which can be modeled via a link function g as follows:

$$\begin{aligned}
g(\mu^I_{i,t}) &= \beta^I_{0t}(\mathrm{lon}_i, \mathrm{lat}_i) + \beta^I_{1t}(\mathrm{lon}_i, \mathrm{lat}_i)\log(I_{i,t-1}),\\
g(\mu^D_{i,t}) &= \beta^D_{0t}(\mathrm{lon}_i, \mathrm{lat}_i) + \beta^D_{1t}(\mathrm{lon}_i, \mathrm{lat}_i)\log(I_{i,t-1}),\\
\mu^R_{i,t} &= \beta^R_{0t} + \beta^R_{1t} I_{i,t-1}.
\end{aligned}$$

In practice, one can use the bivariate spline over triangulation to approximate the spatially varying coefficient functions, $\beta_{0t}(\mathrm{lon}_i, \mathrm{lat}_i)$ and $\beta_{1t}(\mathrm{lon}_i, \mathrm{lat}_i)$. The triangulation can be obtained through various software packages; see for example, the Matlab code DistMesh, and the R package Triangulation [31]. Based on a triangulation, the bivariate spline basis can be generated via the R package BPST [30]. The entire estimation procedure is completed using a quasi-likelihood approach via the penalized spline approximation and an iteratively reweighted least-squares technique; see details in [34].

5.2. Outlier correction

The consequences of outliers may result in reduced forecast accuracy due to (1) bias in the estimates of model parameters and (2) a carry-over effect of the outlier on the prediction. Consequently, the reported data (i.e. infection, death, and recovery) should undergo a preprocessing step to lessen the impact of inaccurate data or anomalies. Therefore, outlier detection and correction are vital for time series analysis based on the reported COVID-19 data.

A simple solution to lessen the impact of an outlier is to replace the outlier with a more typical value before generating the forecasts. This process is often referred to as 'Outlier Correction'. Traditionally, an outlier usually occurs due to measurement variability, data entry, or experimental error. Once an outlier is detected, especially for the ones resulted from the experimental error, we sometimes exclude them from the dataset. However, the outliers in COVID-19 epidemic data typically occur for specific reasons, such as the release of a large batch of tests and the change of reporting standards. Simply excluding extreme values solely due to their extremeness can distort the data analysis. Therefore, the outlier correction procedure should be different from the traditional ones. In this subsection, we describe an automated procedure for 'correcting' the history before forecasting. For a count time series $\{Z_t\}_{t=1}^{T}$, we assume that M outliers, $\{Z_{t_m}\}_{m=1}^{M}$, have been detected, then we implement Algorithm 1 for the repairing procedure. Figure 8 illustrates some examples of the outlier correction.

```
Data: Count time series with outliers.
Result: Count time series with outliers repaired.
Step 0. Sort the $M$ outliers based on the time, i.e. $t_1 < t_2 < \cdots < t_M$.
for $m \leftarrow 1$ to $M$ do
    Step 1. Implement the data repairing methods discussed in Section 5.1 to
      obtain a reasonable estimate $\widehat{Z}_{t_m}$ of the observed data point $Z_{t_m}$.
    Step 2. if $|Z_{t_m} - \widehat{Z}_{t_m}| > \delta$ then
        (i) Manually investigate the causes and the problematic period caused by
          the point anomaly $Z_{t_m}$, denoted as $\mathcal{A}_m$.
        (ii) Distribute the residual, $Z_{t_m} - \widehat{Z}_{t_m}$, proportional to the value of the time
          series within the problematic period
              $Z_t^* = Z_t + (Z_{t_m} - \widehat{Z}_{t_m})Z_t / \sum_{t \in \mathcal{A}_m} Z_t,\ t \in \mathcal{A}_m$
    end
end
```

Algorithm 1: The outlier repairing algorithm.

6. Technical validation and usage notes

The entire detection and repairing procedure is illustrated in Figure 9. First of all, we obtained the data from all of the four data sources, and used the dissimilarity measure

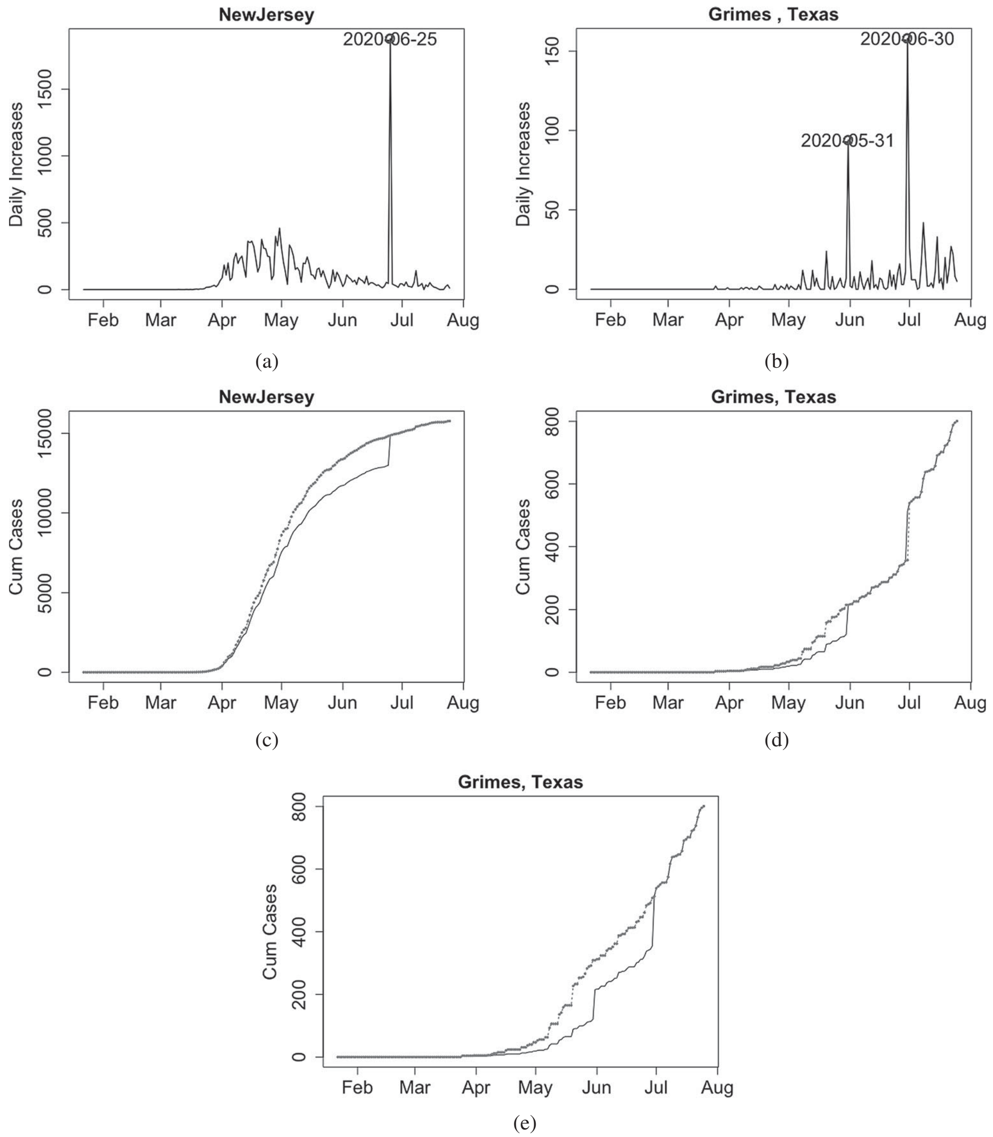


Figure 8. Point anomaly repairing. (a) New death count with one point anomaly; (b) new infected count with two individual point anomalies; (c) point anomalies repairing on the cumulative death count time series; (d) the first point anomaly repairing on the cumulative infected count time series; (e) the second point anomaly repairing on the cumulative infected count time series.

proposed in the above to compare them. We visualize and check the difference at the state level among different data sources based on the comparison results. For the county-level data, we calculate the measure and report the top 10 counties, which are the most different pairwisely. Then, all the data are processed with all types of anomaly detection discussed in Section 4.2. Once an anomaly has been detected, a warning will be given automatically by `R` package `cdcar`. We handle different types of anomalies depending on the circumstances. For example, if an order dependency violation is detected, we will repair that point using our data repairing algorithms proposed in Section 5.1. If a point anomaly is detected,

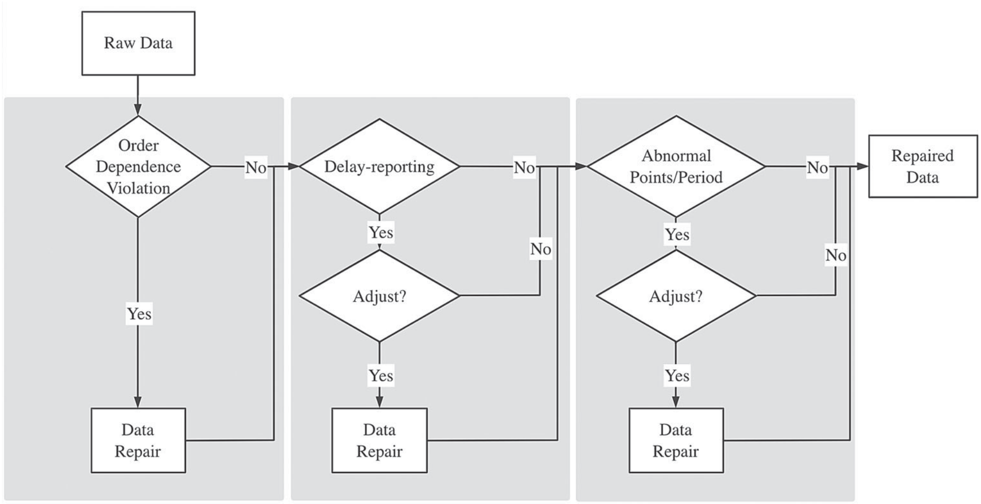


Figure 9. Data curation flowchart.

we first manually check possible legitimate reasons based on news and social media. If correction is necessary, we will repair the point anomalies using the proposed algorithm, see Algorithm 1.

The integrated data are openly available to assist researchers to investigate the spread of COVID-19 in the US. We will continue to provide the cleaned data as the pandemic progresses. Despite the fact we try our best in the data curation, there are two issues without a perfect solution now, which require attention from the users when they try to draw conclusive statements using the data.

(1) *Probable versus Confirmed Cases.* Excluding the population with symptoms but not confirmed by tests leads to the under-reported issue of infectious counts. On 5 April 2020, the Council of State and Territorial Epidemiologists released an interim [4] related to the COVID-19 reporting. It requires the local or state public health authority to submit a report of a condition to the Centers for Disease Control and Prevention (CDC) within 24 h, and CDC should publish data for both 'Confirmed' and 'Probable' cases in the CDC Print Criteria. Before the interim was released, most of the states primarily reported confirmed cases. For the states and counties which started to report probable cases, thereafter, the count of the cases would incur an unavoidable jump after including the probable cases.

(2) *Antibody Test versus Virus Test.* In general, there are two types of tests on infection; one is an antibody test, and the other is a virus test (also referred to as the PCR test). Unfortunately, in many datasets, the type of reported tests is not specified. Those tested positive for the virus are infected at the moment and suggested to be quarantined to avoid infecting others. Meanwhile, those tested positive on an antibody test must have been exposed to the virus, but there is no indication of whether they are still infectious or recovered [19]. In addition, antibody tests are known to be much less accurate. Mixing these two tests makes positive cases uninterpretable. Some states and counties have started to separate antibody tests from virus tests [8], while states such as Pennsylvania, Texas, Georgia, and Vermont did not specify the type of tests.

We will continue to keep close track of the data sources we depend on and update our datasets regularly. We strongly encourage users of our datasets to contact us if there is any anomaly or error. You can reach us either by submitting a request on the Github repository (https://github.com/covid19-dashboard-us/cdcar) or emailing the corresponding author.

7. Conclusion and discussion

The COVID-19 pandemic is generating enormous amounts of data. Open-access data with high quality are critical for the COVID-19 scientific research and response efforts. This paper compares and integrates different data sources and provides a semiotic-based framework for understanding the reported cases' data quality and the techniques for anomaly detection and anomaly repairing.

Correcting the history for severe outliers or anomalies will often improve the forecast; however, if the outlier is not genuinely severe, corrections might make the history smoother than it actually was, which will change the forecasts and narrow the confidence intervals. If the correction was not necessary, it might lead to poor forecasts and unrealistic confidence intervals. We suggest using a high threshold for anomaly detection. In addition, the detected outliers should ideally be individually reviewed by the forecaster, and the reasons for the outliers should be investigated to determine whether a correction is appropriate.

For public usage, all code regarding the proposed anomaly detection and repairing algorithms is built-in `R` package `cdcar`. The package and the cleaned data that are regularly updated can be found in the Github repository (https://github.com/covid19-dashboard-us/cdcar).

Some aspects of our data comparison and curation methods are constrained by the official information released, and we will continuously investigate them. First of all, some data are not retrievable at the county level, such as the recovery data and the face mask data. Second, the data reporting protocols are not consistent, especially for recoveries. Third, although we discover the cyclical pattern in the epidemic data, the related reason still needs to be examined. Fourth, the unassigned issue and under-reported issue might be some other important problems under the pandemic without detailed discussion. In the future, we plan to overcome the data sharing barrier and extend our US COVID-19 database to a worldwide database.

Acknowledgments

The authors are grateful to the referee, Associate Editor and Editor for careful reading of the paper and valuable suggestions and comments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Zhiling Gu and Li Wang's research was partially supported by the National Science Foundation award DMS-1916204. Shan Yu's research was partially supported by the Iowa State University

Plant Sciences Institute Scholars Program. Myungjin Kim's research was partially supported by the National Science Foundation award CCF-1934884 and Laurence H. Baker Center for Bioinformatics & Biological Statistics.

ORCID

Guannan Wang http://orcid.org/0000-0001-6551-4465
Zhiling Gu http://orcid.org/0000-0002-8052-7608
Xinyi Li http://orcid.org/0000-0003-0080-7034
Shan Yu http://orcid.org/0000-0002-0271-5726
Myungjin Kim http://orcid.org/0000-0001-7784-0516
Yueying Wang http://orcid.org/0000-0003-4861-2658
Lei Gao http://orcid.org/0000-0002-4707-0933
Li Wang http://orcid.org/0000-0001-8432-9986

References

[1] F. Ahmed, N. Ahmed, C. Pissarides, and J. Stiglitz, *Why inequality could spread COVID-19*, Lancet Public Health 5 (2020), p. e240.

[2] N. Altieri, R.L. Barter, J. Duncan, R. Dwivedi, K. Kumbier, X. Li, R. Netzorg, B. Park, C. Singh, Y.S. Tan, T. Tang, Y. Wang, C. Zhang, and B. Yu, *Curating a COVID-19 data repository and forecasting county-level death counts in the United States*, preprint (2020). Available at https://arxiv.org/abs/2005.07882.

[3] F. Brauer, P. van den Driessche and J. Wu, *Mathematical Epidemiology*, Vol. 1945, Springer, Berlin, 2008.

[4] Council of State and Territorial Epidemiologists, *Standardized surveillance case definition and national notification for 2019 novel coronavirus disease (COVID-19)*, preprint (2020). Available at https://cdn.ymaws.com/www.cste.org/resource/resmgr/2020ps/Interim-20-ID-01_COVID-19.pdf (accessed November 15, 2020).

[5] B.D. Killeen, J.Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, and M. Unberath, *A county-level dataset for informing the United States' response to COVID-19*, preprint (2020). Available at https://arxiv.org/abs/2004.00756v2.

[6] A.B. Lawson, S. Banerjee, R.P. Haining, and M.D. Ugarte, *Handbook of Spatial Epidemiology*, CRC Press, New York, 2016.

[7] T. Liboschik, K. Fokianos, and R. Fried, *tscount: Analysis of count time series*, R package version 1.4.3, 2020. Available at https://CRAN.R-project.org/package=tscount (last accessed November 15, 2020).

[8] A.C. Madrigal and R. Meyer, *How could the CDC make that mistake?*, preprint (2020). Available at https://www.theatlantic.com/health/archive/2020/05/cdc-and-states-are-misreporting-covid-19-test-data-pennsylvania-georgia-texas/611935 (last accessed November 15, 2020).

[9] S. Mervosh, D. Lu, and V. Swales, *See which states and cities have told residents to stay at home*, preprint (2020). Available at https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html (last accessed November 15, 2020).

[10] V.M. Muggeo, *segmented: Regression models with break-points/change-points estimation*, R package version 1.3-0 (2020). Available at https://CRAN.R-project.org/package=segmented (last accessed November 15, 2020).

[11] NYT, *Coronavirus (Covid-19) data in the United States*, preprint (2020). Available at https://github.com/nytimes/covid-19-data (last accessed November 15, 2020).

[12] D. Ollech, *seastests: Seasonality tests*, R Package version 0.14.2, 2019. Available at https://cran.r-project.org/web/packages/seastests (last accessed November 15, 2020).

[13] R. Perper, E. Cranley, and S. Al-Arshani, *Almost all US states have declared states of emergency to fight coronavirus–here's what it means for them*, preprint (2020). Available at https://www.businessinsider.com/california-washington-state-of-emergency-coronavirus-what-it-means-2020-3 (last accessed November 15, 2020).

[14] D.U. Pfeiffer, T.P. Robinson, M. Stevenson, K.B. Stevens, D.J. Rogers, and A.C.A. Clements, *Spatial Analysis in Epidemiology*, Oxford University Press, New York, 2008.

[15] C. Silver, *Stiglitz: Pandemic exposed health inequality and flaws of market economy*, preprint (2020). Available at https://www.investopedia.com/nobel-winner-joseph-stiglitz-on-income-inequality-after-covid-19-4843052 (last accessed November 15, 2020).

[16] The Atlantic, *The COVID tracking project data*, preprint (2020). Available at https://covidtracking.com (last accessed November 15, 2020).

[17] The Bureau of Transportation Statistics, *Trips by distance*, preprint (2020). Available at https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv (last accessed November 15, 2020).

[18] The Center for Systems Science and Engineering at Johns Hopkins University, *COVID-19 data repository*, preprint (2020). Available at https://github.com/CSSEGISandData/COVID-19 (last accessed November 15, 2020).

[19] The U.S. Department of Health and Human Services, *Guidance on interpreting COVID-19 test results* preprint (2020). Available at https://www.whitehouse.gov/wp-content/uploads/2020/05/Testing-Guidance.pdf (last accessed November 15, 2020).

[20] U.S. Census Bureau, *2005–2009 American community survey 5-year estimates*, preprint (2020). Available at https://data.census.gov/cedsci/table?q=gini%20coefficient&hidePreview=false&tid=ACSDT1Y2018.B19083&vintage=2018 (last accessed November 15, 2020).

[21] U.S. Census Bureau, *2010 U.S. decennial census*, preprint (2020). Available at https://data.census.gov/cedsci/table?q=urban%20rate&hidePreview=false&tid=DECENNIALSF12010.H2&vintage=2010 (last accessed November 15, 2020).

[22] U.S. Census Bureau, *2010–2018 American community survey demographic and housing estimates*, preprint (2020). Available at https://data.census.gov/cedsci/table?q=population%20density&hidePreview=false&tid=ACSDP1Y2018.DP05&vintage=2018 (last accessed November 15, 2020).

[23] U.S. Census Bureau, *2012 economic census*, preprint (2020). Available at https://data.census.gov/cedsci/table?q=government%20expenditures%20for%20health&hidePreview=false&tid=ECNGRANT2012.EC1262SXSB3&t=Government%3AHealth&vintage=2012 (last accessed November 15, 2020).

[24] U.S. Census Bureau, *The U.S. gazetteer files*, preprint (2020). Available at https://www2.census.gov/geo/docs/maps-data/data/gazetteer/2019_Gazetteer (last accessed November 15, 2020).

[25] U.S. Census Bureau, *USA counties: 2011*, preprint (2020). Available at https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html (last accessed November 15, 2020).

[26] U.S. Department of Homeland Security, *Homeland infrastructure foundation-level data*, preprint (2020). Available at https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0/data (last accessed November 15, 2020).

[27] USAFacts, *Coronavirus locations: COVID-19 map by county and state*, preprint (2020). Available at https://usafacts.org/visualizations/coronavirus-covid-19-spread-map (last accessed November 15, 2020).

[28] USAFacts, *Detailed methodology and sources: Covid-19 data: Detailed methodology on how USAFacts collects COVID-19 data*, preprint (2020). Available at: https://usafacts.org/articles/detailed-methodology-covid-19-data/ (last accessed November 15, 2020).

[29] G. Wang, Z. Gu, X. Li, S. Yu, M. Kim, Y. Wang, L. Gao, and L. Wang, *Comparing and integrating US COVID-19 daily data from multiple sources: A county-level dataset with local characteristics*, preprint (2020). Available at https://arxiv.org/abs/2006.01333v3.

[30] G. Wang, L. Wang, M.J. Lai, M. Kim, X. Li, J. Mu, Y. Wang, and S. Yu, *BPST: bivariate spline over triangulation*, R package version 1.0, 2019. Available at https://github.com/funstatpackages/BPST (last accessed November 15, 2020).

[31] L. Wang and M.J. Lai, *Triangulation*, R package version 1.0, 2019. Available at https://github.com/funstatpackages/Triangulation (last accessed November 15, 2020).

[32] L. Wang, G. Wang, L. Gao, X. Li, S. Yu, M. Kim, and Y. Wang, *An R shiny app to visualize, track, and predict real-time infected cases of COVID-19 in the United States*, preprint (2020). Available at https://covid19.stat.iastate.edu (last accessed November 15, 2020).

[33] L. Wang, G. Wang, L. Gao, X. Li, S. Yu, M. Kim, Y. Wang, and Z. Gu, *An R shiny app to predict the infected and death cases of COVID-19 in the U.S. for the next four months*, preprint (2020). Available at https://covid19.stat.iastate.edu/longtermproj.html (last accessed November 15, 2020).

[34] L. Wang, G. Wang, L. Gao, X. Li, S. Yu, M. Kim, Y. Wang, and Z. Gu, *Spatiotemporal dynamics, nowcasting and forecasting of COVID-19 in the United States*, preprint (2020). Available at https://arxiv.org/abs/2004.14103.

[35] K. Webel and D. Ollech, *An Overall Seasonality Test Based on Recursive Feature Elimination in Conditional Random Forests*, Proceedings of the 5th International Conference on Time Series and Forecasting, 2018, pp. 20–31.

[36] C.H. Weiß, *Thinning operations for modeling time series of counts–a survey*, AStA Adv. Stat. Anal. 92 (2008), pp. 319–341.

Appendix. Data records

A.1 Epidemic data

Using the algorithm discussed in Section 5, we aggregate the reported COVID-19 infected, death, and recovered cases from 22 January 2020 from (1) the NYT [11], (2) the Atlantic [16], (3) the COVID-19 Data Repository from the JHU [18], and (4) the USAFacts [27]. These daily updated epidemic datasets are available on Github repository https://github.com/covid19-dashboard-us/cdcar.

In the state level epidemic data, we include the following variables. Among those variables, the variable **State** can be used as the key for data merge.

(1) **State**–Name of state. There are 48 mainland US states and the District of Columbia.
(2) **XYYYY.MM.DD**–Cumulative infection or death cases related to the date of **YYYY.MM.DD**. **YYYY**, **MM**, and **DD** represent year, month and day, respectively. It starts from **X2020.01.22**. For example, the variable **X2020.01.22** is either infection or death cases in a certain state (**State**) on 01/22/2020.

For county-level data, two more county-specific variables are included. As the key of this table, variable **ID** can be used for future data merge.

(1) **ID**–County-level Federal Information Processing System (FIPS) code, which uniquely identifies the geographic area. The number has five digits, of which the first two are the FIPS code of the state to which the county belongs.
(2) **County**–Name of county matched with **ID**. There are about 3200 counties and county-equivalents (e.g. independent cities, parishes, boroughs) in the US.
(3) **State**–Name of state matched with **ID**. There are 50 states and the District of Columbia in the US.
(4) **XYYYY.MM.DD**–Cumulative infection or death cases related to the date of **YYYY.MM.DD**. **YYYY**, **MM**, and **DD** represent year, month and day, respectively. It starts from **X2020.01.22**. For example, the variable **X2020.01.22** is either infection or death cases in a certain (**County**) on 01/22/2020.

A.2 Other factors

A.2.1 Policy data

We release two datasets for the 'stay-at-home/shelter-in-place' order and the declaration of 'state of emergency' from (1) Business Insider [13], (2) New York Times [9], and additional local news.

(1) **ID**–County-level Federal Information Processing System (FIPS) code, which uniquely identifies the geographic area. The number has five digits, of which the first two are the FIPS code of the state to which the county belongs.
(2) **County**–Name of county matched with **ID**. There are about 3200 counties and county-equivalents (e.g. independent cities, parishes, boroughs) in the US.
(3) **State**–Name of state matched with **ID**. There are 50 states and the District of Columbia in the US.
(4) **XYYYY.MM.DD**–Indicators for whether the policy is in effect on the date of **YYYY.MM.DD**, 1 indicates the policy is in effect, 0 otherwise. **YYYY**, **MM**, and **DD** represent year, month and day, respectively. It starts from **X2020.01.22**. For example, the variable **X2020.01.22** represents whether the policy is in effect in a certain (**County**) on 01/22/2020.

A.2.2 Demographic characteristics

In the demographic characteristics category, we consider the factors describing racial, ethnic, sexual, and age structures. Specifically, we include the following six variables. Among these six variables, **AA_PCT** and **HL_PCT** are obtained from the 2010 Census [21]. The other four variables are extracted from the 2010–2018 American Community Survey (ACS) Demographic and Housing Estimates [22].

(1) **AA_PCT**–The percent of the population who identify as African American;
(2) **HL_PCT**–The percent of the population who identify as Hispanic or Latino;
(3) **Old_PCT**–The percent of aged people (age $\geq$ 65 years);
(4) **Sex_ratio**–The ratio of male over female;
(5) **PD_log**–The logarithm of the population density per square mile of land area; **Pop_log**–The logarithm of local population;
(6) **Mortality**–The 5-year (1998–2002) average mortality rate, measured by the total counts of deaths per 100, 000 population in a county.

A.2.3 Healthcare infrastructure

We incorporated three features related to the healthcare infrastructure at the county level in the datasets. Among these variables, **NHIC_PCT** is available in the USA Counties Database [25], **EHPC** is obtained from Economic Census 2012 [23], and **TBed** is compiled from Homeland Infrastructure Foundation-level Data [26].

(1) **NHIC_PCT**–the percent of persons under 65 years without health insurance;
(2) **EHPC**–the local government expenditures for health per capita;
(3) **TBed**–total bed counts per 1000 population.

A.2.4 Socioeconomic status

We consider diverse socioeconomic factors in the county level datasets. All of these factors collected from 2005–2009 ACS 5-year estimates [20]. We also calculate the Gini coefficient based on the household income data from the 2005–2009 ACS [20] to measure the income inequality.

(1) **Affluence**–Social affluence generated by factor analysis from **HighIncome**, **HighEducation**, **WCEmployment** and **MedHU**;

(2) **HIncome_PCT**–The percent of families with annual incomes higher than $75,000;
(3) **HEducation_PCT**–The percent of the population aged 25 years or older with a bachelor's degree or higher;
(4) **MedHU**–The median value of owner-occupied housing units;
(5) **Disadvantage**–Concentrated disadvantage obtained by factor analysis from **HHD_PAI_PCT**, **HHD_F_PCT** and **Unemployment_PCT**;
(6) **HHD_PAI_PCT**–The percent of the households with public assistance income;
(7) **HHD_F_PCT**–The percent of households with female householders and no husband present;
(8) **Unemployment_PCT**–Civilian labor force unemployment rate;
(9) **Gini**–The Gini coefficient, a measure for income inequality and wealth distribution in economics.

A.2.5 Environmental factor

We also collect some environmental factors that might affect the spread of epidemics significantly, such as the urban rate and crime rate.

(1) **UrbanRate**–Urban rate [21];
(2) **ViolentCrime**–The total number of violent crimes per 1000 population [25];
(3) **PropertyCrime**–The total number of property crimes per 1000 population [25];
(4) **ResidStability**–The percent of the population residence in the same house for one year and over [21].

A.2.6 Mobility

The mobility data are collected from the US Department of Transportation, Bureau of Transportation Statistics. It describes the daily number of trips within each county, which are produced from an anonymized national panel of mobile device data from multiple sources. Trips are defined as movements that include a stay of longer than 10 min at an anonymized location away from home.

(1) **Number of trips X–XX**–Number of trips by residents greater than X miles and shorter than XX miles. There are 10 different trip ranges: '≤ 1' '1–3', '3–5', '5–10', '10–25', '25–50' '50–100', '100–250', '250–500', and '≥500'.
(2) **Population Stay at Home**–Number of residents staying at home, that is, persons who make no trips with a trip end more than one mile away from home.

A.3 Geographic information

The **longitude** and **latitude** of the geographic center for each county in the US are available in Gazetteer Files [24].