Research Article **Open Access**

Tao Hu, Weihe Wendy Guan, Xinyan Zhu, Yuanzheng Shao, Lingbo Liu, Jing Du, Hongqiang Liu, Huan Zhou, Jialei Wang, Bing She, Luyao Zhang, Zhibin Li, Peixiao Wang, Yicheng Tang, Ruizhi Hou, Yun Li, Dexuan Sha, Yifan Yang, Ben Lewis, Devika Kakkar, Shuming Bao*

Building an Open Resources Repository for COVID-19 Research

https://doi.org/10.2478/dim-2020-0012
received May 1, 2020; accepted June 3, 2020.

Abstract: The COVID-19 outbreak is a global pandemic declared by the World Health Organization, with rapidly increasing cases in most countries. A wide range of research is urgently needed for understanding the COVID-19 pandemic, such as transmissibility, geographic spreading, risk factors for infections, and economic impacts. Reliable data archive and sharing are essential to jump-start innovative research to combat COVID-19. This research is a collaborative and innovative effort in building such an archive, including the collection of various data resources relevant to COVID-19 research, such as daily cases, social media, population mobility, health facilities, climate, socioeconomic data, research articles, policy and regulation, and global news. Due to the heterogeneity between data sources, our effort also includes processing and integrating different datasets based on GIS (Geographic Information System) base maps to make them relatable and comparable. To keep the data files permanent, we published all open data to the Harvard Dataverse (https://dataverse.harvard.edu/dataverse/2019ncov), an online data management and sharing platform with a permanent Digital Object Identifier number for each dataset. Finally, preliminary studies are conducted based on the shared COVID-19 datasets and revealed different spatial transmission patterns among mainland China, Italy, and the United States.

Keywords: COVID-19, open data, data repository, GIS, spatial data

***Corresponding author: Shuming Bao,** China Data Institute, Ann Arbor, MI, USA. E-mail: sbao@umich.edu
Tao Hu, Weihe Wendy Guan, Ben Lewis, Devika Kakka, Center for Geographic Analysis, Harvard University, Cambridge, MA, USA
Xinyan Zhu, Yuanzheng Shao, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei, China; Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan Hubei, China
Lingbo Liu, School of Urban Design, Wuhan University, Wuhan, Hubei, China
Jing Du, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei, China
Hongqiang Liu, College of Geomatics, Shandong University of Science and Technology, Qingdao, Shandong, China
Huan Zhou, Jialei Wang, School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei, China
Bing She, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
Luyao Zhang, School of Business Management, East China Normal University, Shanghai, China
Ruizhi Hou, School of Mathematical Sciences, East China Normal University, Shanghai, China
Zhibin Li, School of Government, Peking University, Beijing, China
Peixiao Wang, School of Government, Peking University, Beijing, China; The Academy of Digital China, Fuzhou University, Fuzhou, Fujian, China
Yicheng Tang, School of Management, Hefei University of Technology, Hefei, Anhui, China
Yun Li, Dexuan Sha, Department of Geography and GeoInformation Science, George Mason University, Fairfax, VA, USA
Yifan Yang, Department of Biological Sciences, University of California, San Diego, CA, USA

1 Introduction

As of May 29, 2020, the novel coronavirus (COVID-19) has spread to more than 216 countries and terrorism (Henrik, Byron, & Sergio, 2020) with a total of 5,704,736 confirmed cases and 357,736 deaths globally (WHO, 2020a). On March 12, 2020, World Health Organization (WHO) announced COVID-19 outbreak a pandemic due to the rapid increase of cases across the world and growing number of affected countries (WHO, 2020b). A coordinated global response is desperately needed to overcome this unprecedented challenge (Remuzzi & Remuzzi, 2020). Consistent recording of COVID-19-related information is critically important for understanding transmissibility, risk of geographic spreading, routes of transmission, and risk factors for infection (Xu et al., 2020). This article presents

the process of collecting, integrating, and publishing COVID-19-related resources, facilitating the investigation of COVID-19 pandemic by global researchers.

There are many open resources related to COVID-19 on daily infection case counts, Twitter data, population mobility data, research papers, and so on. However, those resources have several limitations: (1) the open data are deployed on different platforms without a central collection. It is hard for researchers to discover and integrate the data for their studies. (2) The data sources lack a standard for interoperability. Without a data interoperability standard, the data are heterogeneous, not comparable in their native form. (3) The open data are not deposited following the professional standard for permanent collections. Many data are published on GitHub, the world's leading software development platform without a professional management system for permanent data depository. Some data may only exist on temporary websites for a short period due to limited resource support. As a joint effort by scholars and professionals from the Center for Geographical Analysis at Harvard University, the Geo-Computation Center for Social Sciences at Wuhan University, the China Data Institute, the RMDS Lab, and some other institutions, an initiative on "Resources for COVID-19 Study" was sponsored by the China Data Lab project (https://projects.iq.harvard.edu/chinadatalab/resources-covid-19). The objectives of this project are: (1) to provide data support for the spatial study of COVID-19 at local, regional, and global levels with information collected and integrated from different sources; (2) to facilitate quantitative research on spatial spreading and impacts of COVID-19 with advanced methodology and technology; (3) to promote collaborative research on the spatial study of COVID-19 on the China Data Lab, Dataverse, and WorldMap platforms; and (4) to build research capacity for future collaborative projects. This article discusses the methodology and technology for collecting data from different sources, integrating data with base maps and other resources for social, economic, environmental, and health care data, publishing the standard data products on the Harvard Dataverse with unique and permanent Digital Object Identifiers (DOIs) for those data collections and datasets deposited on Dataverse. This article also presents some preliminary studies on the transmission patterns of the COVID-19 pandemic.

2 Overview of Related Work

2.1 Open Data Resources

The open data resources related to COVID-19 studies that the focus of this article can be classified into the following major categories: infection cases, policy, population mobility, social media, and research articles. A list of popular COVID-19 data resources is presented in Table 1, including data title, provider, covered region, and update frequency. The 2019-nCoV Time Series Infection Data Warehouse is one of the earliest open datasets after the COVID-19 breakout in China. It focuses on the reported cases in China and scrapes the data from DXY.cn using Python-based crawler with hourly updates (Lin, 2020). John Hopkins University Center for Systems Science and Engineering (JHU CSSE) published its coronavirus repository from the beginning of February 2020. It started from case counts in China and later expanded to case counts globally, with an interactive web-based dashboard for tracking COVID-19 in real time (Dong, Du, & Gardner, 2020). This dataset is widely used by many organizations. Since March 2020, the novel coronavirus outbreak in Europe became worse, with Italy at the forefront. As of 20 March, Italy is the world's center of active coronavirus cases with 42,681 active cases—more than double the number of active cases of any other countries and exceeding those of China and Iran combined (Worldometer, 2020). Thus, the Presidency of the Council of Ministers—Department of Civil Protection in Italy is monitoring the situation and publishing province and regional-level case counts in real time on GitHub. It is helpful for scientists to explore the spreading of the virus and make predictions more precisely. With the continuing increase of confirmed cases in the United States, there are several open data sources with daily updated case counts besides John Hopkins University (JHU), such as *New York Times* (Times, 2020), including confirmed and death cases at state and county levels. More countries also started to report COVID-19 case counts for their second-level administrative divisions, including Japan, South Korea, France, Germany, the United Kingdom, and so on (Merlière, 2020; Wu, Ge, Yu, & Hu, 2020).

Besides the COVID-19 case count datasets, there are several other related data collections provided by different organizations. The Oxford COVID-19 Government Response Tracker (OxCGRT) systematically collects information on policy responses by different governments, scores the stringency of such measures, and aggregates those scores into a common Stringency Index. It provides a systematic

Table 1
Open Resources Related to COVID-19 Study

ID	Type	Title	Provider	Region	Update frequency
1	Cases	COVID-19 Cases	JHU CSSE	Global	Hours
2		Geographic Distribution of COVID-19 Cases Worldwide	European Centre for Disease Prevention and Control	Global	Daily
3		COVID-19/2019-nCoV Time Series Infection Data Warehouse	Isaac Lin	China	Hours
4		nCov2019: An R Package with Real-time Data, Historical Data, and Shiny App	Guangchuang Yu, Xijin Ge, et al.	China, South Korea, USA, Japan, Iran, Italy, Germany, and UK	Hours
5		covid-19-data	*New York Times*	USA	Daily
6		Dati COVID-19 Italia	Presidency of the Council of Ministers—Department of Civil Protection	Italy	Days
7		OpenCOVID19-fr	Antoine Augusti, Colin Maudry, et al.	France	Hours
8	Policy	OXFORD COVID-19 Government Response Tracker	University of Oxford	Global	Daily
9		State Actions to Mitigate the Spread of COVID-19	Kaiser Family Foundation	USA	Daily
10	Mobility	Baidu Mobility Data	Baidu, Inc.	China	Daily
11		Mobility Changes in Response to COVID-19	Descartes Labs	USA	Daily
12	Social media	#COVID-19: The First Public Coronavirus Twitter Dataset	University of Southern California	Global	Daily
13	Publications	Database of Publications on Coronavirus Disease (COVID-19)	WHO	Global	Daily
14		COVID-19 Open Research Dataset (CORD-19)	Allen Institute for AI	Global	Weekly

cross-national, cross-temporal measure for understanding how government responses have evolved over the period of the disease outbreak (University of Oxford, 2020). Kaiser Family Foundation also published the state policy actions in the United States in categories of stay at home order, quarantine for travelers, business closures, large gatherings ban, school closures, bar/restaurant limits, primary election postponement, emergency declaration, and health and insurance policies (Foundation, 2020).

Mobility data track people's movement, which help to explore the spatial trend of virus spreading. Stephen published January and February mobility data at the University of Virginia Dataverse, which defines daily patterns and the connectivity of population movements at county and prefecture (city) levels across mainland China. The team used de-identified and aggregated mobile phone data, air passenger itinerary data, and case reports to assess spreading risk of the novel coronavirus within China (Lai et al., 2020). Michael and Samuel at Descartes Labs use anonymized mobile device locations to measure mobility (Warren & Skillman, 2020). The US state- and county-level mobility data have been made freely available in the GitHub repository.

Social media data can reveal misinformation and unverified rumors and can help to understand public fear, panic, and other social sentiments (Chen, Lerman, & Ferrara, 2020). The first coronavirus-related Twitter Dataset was published on GitHub, which includes tweets collected using Twitter API since January 22, 2020, based on key words such as coronavirus, koronavirus, Wuhanlockdown, covid-19, covid19, sars-cov-2, and so on. It also tracks Twitter accounts such as CoronaVirusInfo,

V2019N, CDCemergency, CDCgov, WHO, HHSGov, NIAIDNews, and so on. Although the dataset covers tweets by people from all around the world, it does not have geographic information. Thus, it is hard to tell where the tweets were posted from, even the country identity.

In response to the COVID-19 pandemic, WHO and the Allen Institute for AI distributed a free resource of scholarly articles about COVID-19 for use by global research community. The latter includes over 51,000 articles, covering the coronavirus family of viruses (WHO, 2020c; Scholar, 2020). Both datasets are intended to mobilize researchers to apply recent advances in natural language processing to generate new insights in support of the fight against this infectious disease. The corpus is updated weekly as new research is published in peer-reviewed publications and archival services such as bioRxiv, medRxiv, and others.

There are two more large open datasets for COVID-19 research. Google Cloud released a COVID-19 public dataset program to make data freely accessible for better public outcomes (BigQuery Public Datasets Program, 2020). Although the Google Cloud offers many valuable datasets, including COVID-19 case reports, health data, census data, open street map, and some other datasets, the program will remain in effect only until September 15, 2020. The World Bank also makes some COVID-19-related datasets open for public (Bank, T.W, 2020), including Living Standards Measurement Study Collection, Demographic and Health Surveys, UNICEF's Multiple Indicator Cluster Survey, WHO's Multi-Country Studies Programs, and Integrated Public Use Microdata Series. All of those datasets are historical country-level datasets. To help researchers, officials, and medical staffs to understand the virus and pandemic more, AMiner platform collects all kinds of open datasets about COVID-19, covering research, knowledge, media, policy, and pandemic-related topics and keeps updating daily (AMiner, 2020).

2.2 Published Studies

In rapid response to the worldwide crisis of novel coronavirus, researchers from multiple fields, including public health, computer science, data science, geographic information science, economics, etc., have collaborated each other for disease surveillance, early warning, risk assessment, and emergency response. In addition to medical scientists who are developing means for testing, treating, and immunization against novel coronavirus, researchers from other fields collect and share related data (such as state- and county-level confirmed/recovered/death case counts, population mobility, hospital facilities, and other socioeconomic data) for exploring spreading patterns and impact factors of coronavirus (e.g., policy, environmental, and socioeconomic factors), supporting decision-making and preparedness. Dong et al. (2020) developed an interactive web-based dashboard to track COVID-19 in real time, aiding visualization of the virus transmission in the spatiotemporal context. In the earth science domain, Luo et al. collected temperature data and humidity data to address the role of absolute humidity on transmission rate of the COVID-19 outbreak at province level and found that changes in weather alone does not necessarily lead to decline in COVID-19 case counts (Luo et al., 2020). Sajadi et al. (2020) predicted spreading and seasonality of COVID-19 using temperature and latitude and pointed out a zone at increased risk for COVID-19 spreading. Wang et al. investigated the influence of air temperature and relative humidity and found that high temperature and high humidity significantly reduce the transmission of COVID-19, from which they concluded that the arrival of summer and rainy seasons in the northern hemisphere can effectively reduce the transmission of the COVID-19 (Wang, Tang, Feng, & Lv, 2020). To compare the environmental changes before and during the pandemic in mainland China, Liu et al. (2020) investigated the spatial and temporal characteristics of Nighttime Light (NTL) radiance and Air Quality Index (AQI). In the public health domain, many researchers focus on comparing infected cases before and after the announcement of social distancing policies. Wells et al. estimated the impact of travel restrictions and border control measures and concluded that reduction in the rate of exportation could delay the importation of cases into cities unaffected by the COVID-19 outbreak and buy time for public health responses (Wells et al., 2020). Tian et al. (2020) evaluated travel restrictions of Wuhan City in response to the 2019 novel coronavirus outbreak and discovered that the travel ban slowed the dispersal of novel coronavirus from Wuhan to other cities in China by 2.91 days (95% CI: 2.54–3.29). Similarly, Chinazzi et al. found that the Wuhan travel quarantine delayed the epidemic progression in mainland China by only 3–5 days and had significant effect at the international scale (Chinazzi et al., 2020). Transportation data also play an important role in analyzing the transmission of COVID-19. Bogoch et al. (2020) evaluated the potential dissemination of COVID-19 across the world via commercial air travel at the early stage of epidemic and stated that major Asian hubs are the most probable sites of exportation.

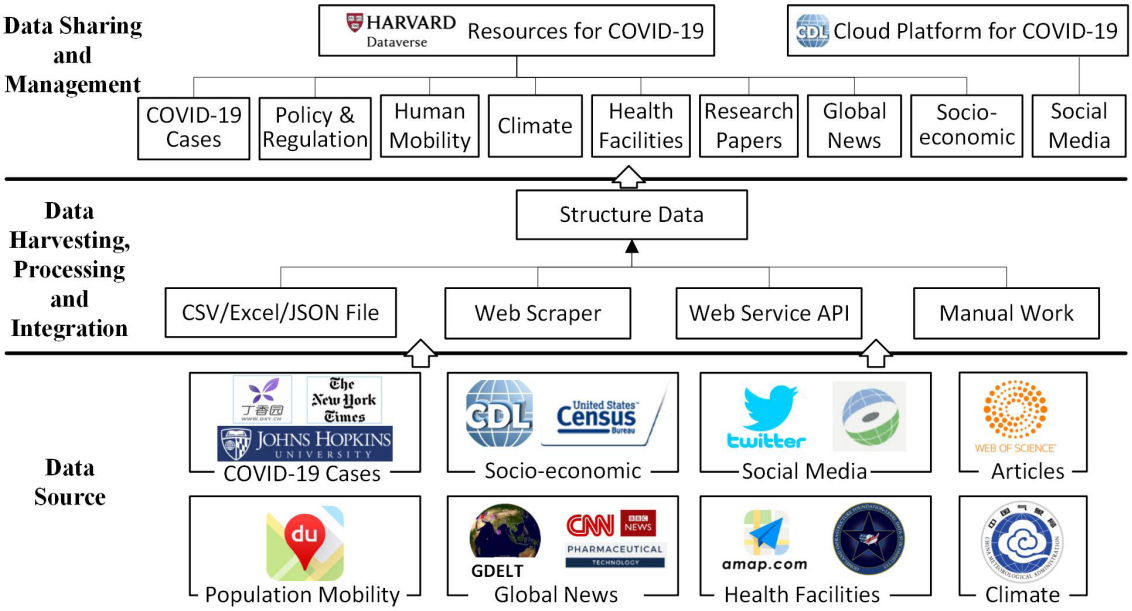

Figure 1. The flow chart of data collection and deployment.

3 Building an Open Resource Repository on COVID-19

Daily counts of COVID-19 confirmed, recovered, and death cases are among the most important data for decision-makers and researchers. Human mobility causes the spreading of contagious diseases. Meanwhile, the recovered and death cases are closely related to capacity at health facilities. This project also collected medical health facilities data (such as hospital) to support further analysis. Mobile social media, especially those with geolocations, contain valuable data for analyzing human behavior and events in the real world (Fujita, 2013). Monitoring public opinion and sentiment variations using social media data is also important for decision-makers to track people's reactions to the COVID-19 outbreak, especially what topics people care about the most (Hu, She, Duan, Yue, & Clunis, 2019). Several studies have shown that climate has close relations with the spreading of COVID-19. Considering the potential relevance of various data in COVID-19 research, we collected different data from various sources, including infection cases, socioeconomic data, social media, scholarly articles, population mobility, global news, health facilities, and climate. Figure 1 presents the framework for collecting and managing these data resources. According to the features of these resources, we applied different methodologies to collect the data. Some resources provide structured data in the format of csv, excel, and JSON, such as COVID-19 infection cases, climate, scholarly articles, and socioeconomic data; some resources provide web service API to access the data, such as Twitter or health facilities POI data; some resources are web pages that do not have data access interfaces, such as Baidu population mobility data, thus web scrapers are developed to collect the data. Policies and regulations in the global news are extracted manually based on the importance of events and a chronicle is built up and updated timely.

To facilitate the study of the COVID-19 outbreak, we deployed the data on the Harvard Dataverse, an online data repository for data sharing, preservation, citation, and exploration. Our data collection covers several categories, including COVID-19 cases, Policy and Regulation, Baidu Mobility, Research Papers, Health Facilities, Socioeconomic data, Global News, and Social Media.

3.1 Data Collection

3.1.1 COVID-19 Daily Cases

This dataset focuses on the COVID-19 daily confirmed, recovered, and deaths cases in China, the United States, and some other countries. As described in Section 2, JHU provides comprehensive COVID-19 cases data compiled from different sources. However, it only covers provincial level case counts in China without data for cities. We scraped the daily COVID-19 infection data from DXY.cn.

For US daily cases, we collected daily cases from *New York Times* compiled from the state and local governments and health departments. For the global cases, we used the JHU's near real-time time-series files.

These three daily case count resources provide different data formats. To standardize the format, we saved those daily cases into three time-series data files: confirmed cases, recovered cases, and death cases. Each file includes regional name, code, and case numbers by dates (t_1, t_2, ..., *tn*, where *tn* is the date). The regional names are the same as base maps provided by the China Data Institute. Section 3.2 describes the procedures of data integration.

3.1.2 Census and Socioeconomic Data

This project obtained China and US Census data and socioeconomic data from separate sources. The China data is provided by the China Data Institute, which is one of the primary data providers for China data services. We used 2000 and 2010 comparable administrative maps in shape file format at province and prefecture city levels as base maps, which can match the 2000 and 2010 China Census data. The GIS map layers include the boundary map (1:100 million) with data from 2000 and 2010 population Census, including general population, fertility, nationalities, marriage, age, education, occupation, housing, migration, and so on.

The US data came from the US Census Bureau (Bureau, 2020). The US data include population estimates in 2018, demographic and housing estimates, general economic characteristics, educational attainment, and selected social characteristics in 2017. The files are compiled at state and county levels.

3.1.3 Population Mobility Data

Mobility data tracks people's movement on the space which helps explore the spatial trend of virus spreading. Baidu offers location-based service (LBS), based on the global positioning system (GPS), IP addresses, locations of signaling towers, Wi-Fi, for online searching and mapping, and a large variety of apps and software on mobile devices (Lai et al., 2020). These data have been used to visualize population mobility around the Chinese New Year (Merlière, 2020). The mobility data is categorized as inflows and outflows by province and city in China. We built the daily mobility matrices for provinces and cities with Baidu inflow and outflow data.

Baidu mobility data are collected based on monitoring the features of HTTP requests to the data server. The request service is provided (https://huiyan.baidu.com/migration/historycurve.jsonp?dt=city&id=440100&type=move_in&callback=jsonpdf) and four parameters need to be identified, including *dt* whose default value is "city", *id* that indicates the city administrative code, *type* that is move in or move out, and *callback* that is the returned data type. After analyzing the responding JSON file, the inflow and outflow matrix is generated. Table 2 presents a sample of outflow matrix between 10 cities on March 28, 2020. The top row names are departure cities while left column names are designated cities. The sum of all values of each column is 100%. For example, the mobility index from Beijing to Tianjin is 8.74 and the index from Tianjin to Beijing is 8.78. Beijing has the largest outflow to Baoding (10.97) than that to any other cities. The index for intra city mobility is null.

3.1.4 Medical Health Facilities

This project collected China's hospital POI data via web services provided by AutoNavi (https://lbs.amap.com). In the data collection request, POI type of "medical service" and administrative region are filled out and web service returns POI's name, address, longitude, and latitude in the format of JSON. To be consistent with the medical health organization classification standard released by the Chinese National Bureau of Statistics, we transform the medical service POI data type to the standard first-level categories, including hospital, special hospital, clinic, emergency center, medical and health care service place, and disease prevention institute. In the second-level categories, the special hospital consists of plastic surgery, stomatology, ophthalmology, ENT, chest, orthopedic, tumor, maternity, psychiatric, and infectious disease hospitals. Each POI record contains hospital name, address, city name, city code, province name, province code, longitude, latitude, and first- and second-level medical service categories.

We also collected the US hospital data from Homeland Infrastructure Foundation-Level Data (HIFLD) (https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals), including 7,581 records in total. The dataset does not contain nursing homes or health centers. Different from China's health organization classification standard, the US hospitals are categorized into children, chronic disease, critical access, general acute care, long-term care, military, psychiatric, rehabilitation, special, and women hospitals. The metadata also contains

Table 2
A Sample of Outflow Matrix between 10 Cities on March 28, 2020 (%)

City_EN	Beijing	Tianjin	Shijiazhuang	Tangshan	Qinhuangdao	Handan	Xingtai	Baoding	Zhangjiakou
Beijing		8.78	2.33	7	3.5	4.82	3.04	16.07	18.41
Tianjin	8.74		2.4	31.21	6.97	7.04	2.15	4.91	4.79
Shijiazhuang	2.05	2.19		3.02	2.76	9.67	29.94	20.35	5.57
Tangshan	1.8	17.03	1.96		37.06	1.04	0.98	1.95	3.71
Qinhuangdao	0.47	1.35	0.64	15.19		0.16	0.2	0.46	0.66
Handan	1.35	3.07	7.23	0.78	0.51		19.34	1.96	0.71
Xingtai	0.77	1.16	21.77	0.72	0.51	19.44		2.26	0.78
Baoding	10.97	3.83	18.67	2.49	1.83	2.62	2.6		7.97
Zhangjiakou	2.77	0.85	1.38	1.33	0.58	0.34	0.24	2.17	

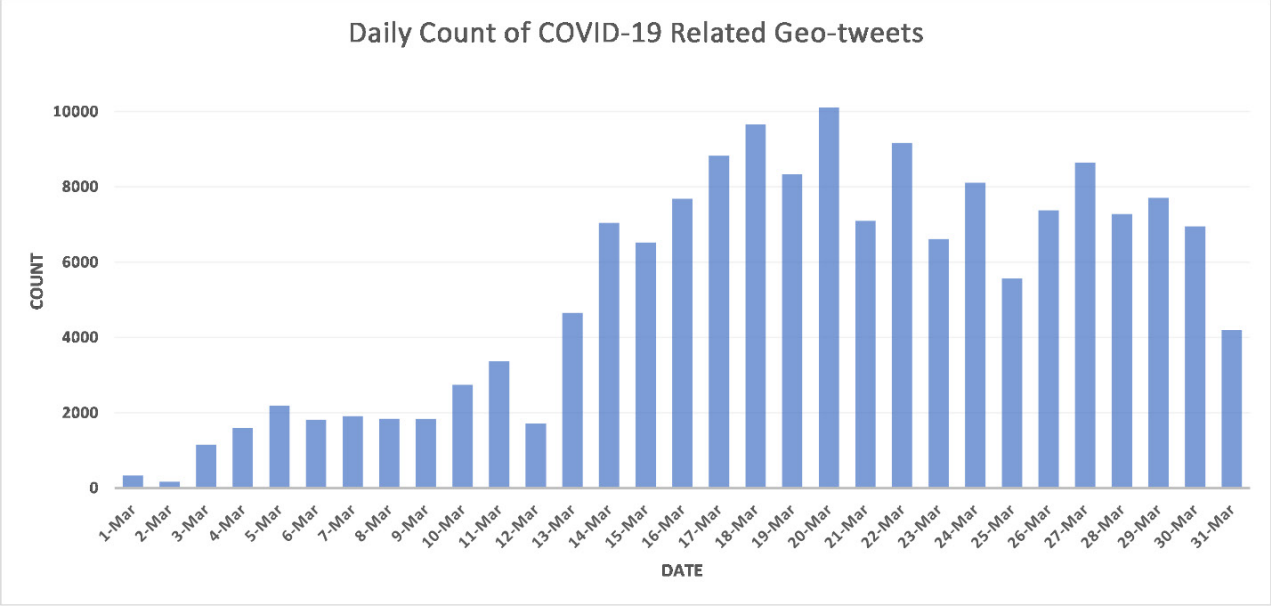

Figure 2. Daily count on global geo-tweets related to COVID-19 topics.

hospital name, address, city, county name, county FIPS code, state name, population, longitude, latitude, North American Industry Classification System (NAICS) code, NAICS description, and some other attributes.

3.1.5 Social Media Data

The social media data is collected from Twitter based on an open-sourced tool Twint, a Python library for advanced Twitter scraping and automated collection process (Grzybowski, Juralewicz, & Piasecki, 2019). Based on the name variations of the novel coronavirus, we collected tweets based on hashtags from January 1, 2020, including #coronavirus, #sarscov2, #2019ncov, and #covid19. There are 1,242,471 tweets collected as of March 31, 2020.

However, the Tweets collected earlier do not provide geolocation information. The Center for Geographic Analysis (CGA) at Harvard University has been harvesting geo-located tweets or "geo-tweets" to a Geo-Tweet Archive since late 2012. Geo-tweets are tweets that contain a pair of geographic coordinates from the originating device denoting the location at which the tweet was created. We extracted most recent geo-tweets in March including the

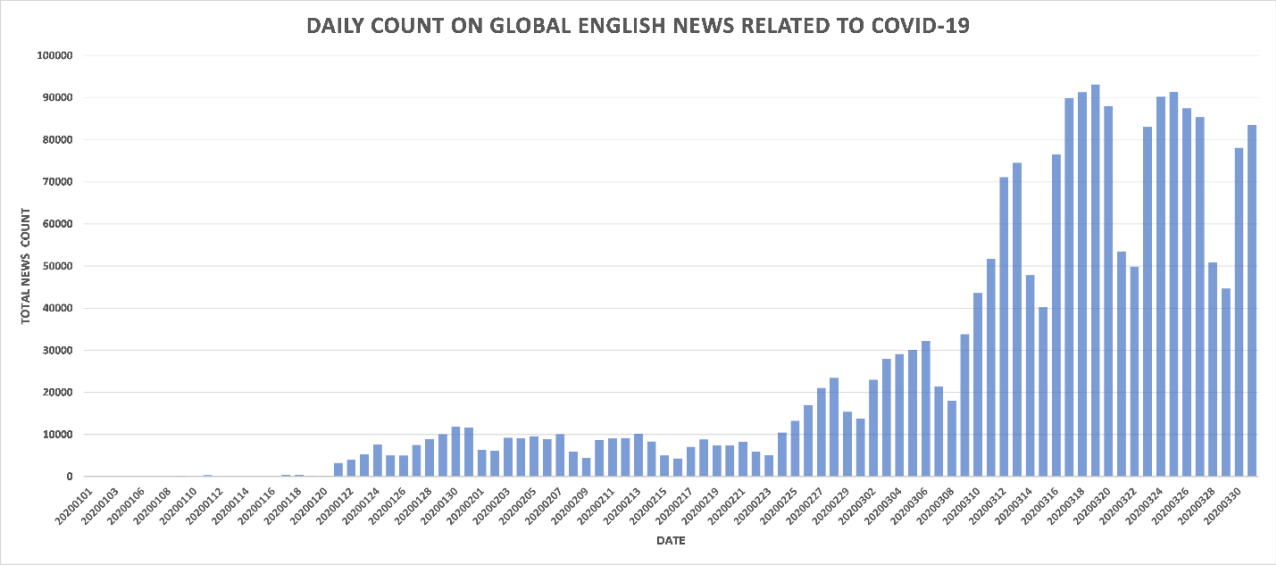

Figure 3. Daily counts of global news in English related to COVID-19 topics.

key words that are the same as the ones mentioned in Grzybowski et al. (2019). Figure 2 presents the daily count of geo-tweets related to COVID-19 topics during March.

3.1.6 Global News

The GDELT Project (https://www.gdeltproject.org/) monitors the world's broadcast, print, and web news from nearly every country in over 100 languages and identifies people, locations, organizations, themes, sources, emotions, counts, quotes, images, and events. It offers a free open platform for public. There are three data files updated every 15 min with download links, including events, mentions, and knowledge graph. We used a Python-based script for downloading the files from January 1, 2020 to current. To extract the news related to COVID-19 topics, a key words list including various COVID-19 terms is built to match the titles of each news. Key words include coronavirus, koronavirus, covid-19, 2019ncov, 2019-ncov, and sars-cov-2.

Figure 3 presents the daily counts of global news in English related to COVID-19 topics. Before Wuhan's lockdown on January 22, 2020, there were only few news on COVID-19. The number of news started to climb from the end of February. The number of published news in the weekend is much fewer than that in the weekdays.

3.1.7 Climate Data

The climate data is collected from the China National Environmental Monitoring Center (CNEMC), an online platform that offers real-time primary pollutants and AQI data for 367 cities in China. Based on the new ambient air quality standard (China, M., 2015), the basic air pollutants include PM2.5, PM10, SO2, NO2, O3, and CO (see Table 3 for details). The AQI level is calculated by the Chinese Standard (China, M.o.E.P.o., 2012) using the above six air pollutants, reported by those monitoring stations in each city. We calculated the daily average, minimum, maximum, and standard deviation of each city with the original data recorded by hours. The daily standard deviation expresses the fluctuation within each day. The minimum and maximum values show the two extreme conditions each day. The primary pollutants can be used as individual variables and AQI is a composite index. The data reflect changes of air quality by day. The current data contain records started from January 1, 2020 to current.

3.1.8 Policy and Regulation

The policy and regulations data file records Chinese and global countries' policies, infection case reports, medical progress, and official warnings, organized as a chronicle, selected for their significant impact on the COVID-19 outbreak. Tracking policies released by the government can help decision-makers and citizens understand the robustness of governmental responses in a consistent way,

Table 3
Metadata Table for Air Quality Data

Variable name	Description	Unit
CO	Content of carbon monoxide in the air	mg/m³
SO_2	Content of sulfur dioxide in the air	mg/m³
NO_2	Content of nitrogen dioxide in the air	mg/m³
O_3	Content of ozone in the air	mg/m³
PM2.5	Suspended particulates <2.5 mm	mg/m³
PM10	Suspended particulates <10 mm	mg/m³

aiding efforts to fight the pandemic (University of Oxford, 2020). For example, "On 23 January 2020, the central government of China imposed a lockdown in Wuhan and other cities in Hubei." This policy had significant impacts on the population flows and the spread of virus. The chronicle is compiled with the information from a variety of authoritative or professional presses, such as *China Daily*, BBC, and CNN News. Each record includes date, title, content, source, link, and category (policy, infection cases, medical progress, and official warning). For China's chronicle, the province or city name is added; for global chronicle, the continent and country names are added.

3.1.9 Scholarly Articles

The Web of Science (WoS) gives access to multiple databases that reference cross-disciplinary research that in turn allows for in-depth exploration of specialized subfields within an academic or scientific discipline (COVID & Team, 2020). We used the WoS core collection as the data source for literature related to COVID-19 topics. The query conditions for TOPIC are "2019-ncov" OR "covid-19" OR "new coronavirus" OR "novel coronavirus" OR "sars-cov-2." The document language is limited to English and the time span is the year of 2020. Using these search conditions, 1,251 records were obtained as of April 21, 2020. The records include authors, publication type, document title, key words, abstract, e-mail address, publisher, cited references, publication date, volume, issue, DOI, and so on. Table 4 presents the top 20 subjects classified by WoS and key words. The results show that researchers are more focused on the analysis of viruses and the comparison of COVID-19 with SARS and MERS.

Table 4
Top 20 Subject Categories and Key Words

ID	Subject	Count	Key word	Count
1	General and internal medicine	440	eCOVID-19	221
2	Infectious diseases	100	SARS-CoV-2	95
3	Public, environmental, and occupational health	67	coronavirus	69
4	Virology	53	2019-nCoV	60
5	Research and experimental medicine	51	Coronavirus	60
6	Radiology, nuclear medicine, and medical imaging	50	China	21
7	Science and technology—other topics	50	Pneumonia	20
8	Biochemistry and molecular biology	49	Wuhan	19
9	Microbiology	49	pandemic	17
10	Immunology	43	pneumonia	17
11	Oncology	38	SARS	16
12	Cell biology	37	novel coronavirus	15
13	Pediatrics	28	epidemiology	14
14	Pharmacology and pharmacy	28	Novel coronavirus	12
15	Life sciences and biomedicine—other topics	26	Epidemiology	11
16	Respiratory system	23	SARS-CoV	11
17	Anesthesiology	21	outbreak	11
18	Emergency medicine	21	epidemic	9
19	Psychiatry	19	public health	9
20	Health care sciences and services	18	2019 novel coronavirus	8

3.2 Data Integration

Data from different sources usually have different formats, which may not be comparable across sources and regions (e.g., the city list with infection cases reported may be different from the city list for Baidu population mobility data). We select some GIS base maps for the integration of data from different sources. Our data collections are

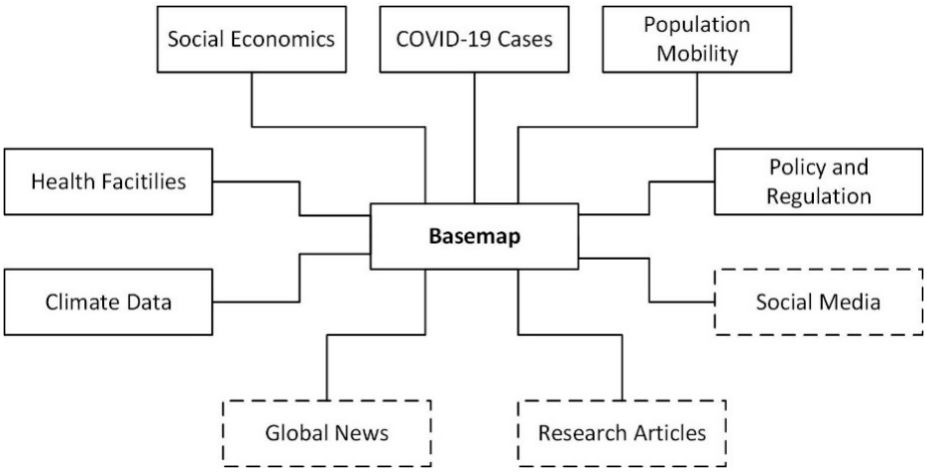

Figure 4. The integration of data from different sources with base maps.

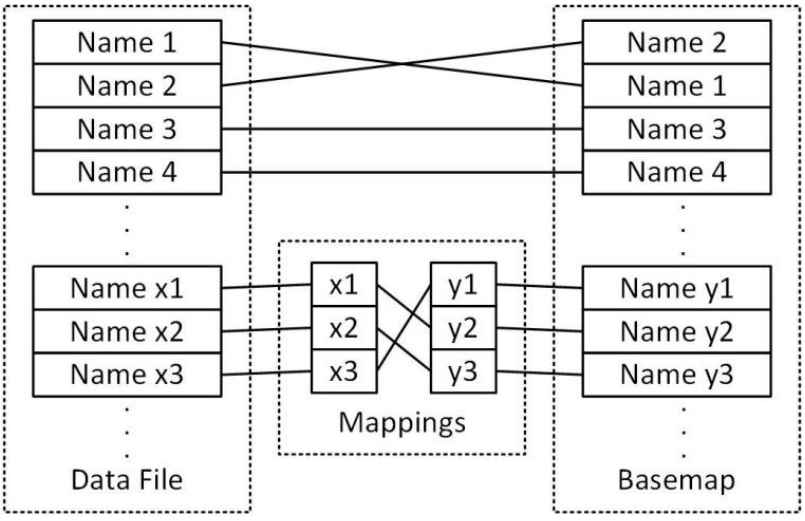

Figure 5. Place names mapping between data files and base maps.

primarily focused on China and the United States. The China and US base maps are provided by the China Data Institute, which allows COVID-19 data to match the administrative boundary maps with selected historical census data variables. Figure 4 shows the data files matching with base maps.

For the US data files, such as daily cases and socioeconomic data, the raw data include state and county's Federal Information Processing Standard (FIPS) code, thus the data can be matched with the US base maps by the FIPS code. For the China data files, such as China's daily cases, health facilities, climate, policies, and regulations, as well as global daily cases retrieved from JHU, the raw data do not provide province/city codes, thus denoting place name matching is the only way to match the records. Figure 5 presents the matching method. In the data files and base maps, some administrative unit names are spelled the same, thus they can be matched easily. However, when the place names are spelled with variations, we created a mapping list to facilitate the matching. Each row of the mapping list records the same place with different names from data files and base maps. For example, when mapping global cases with a global base maps, the country "Congo" in the global daily cases file needs to match with "Democratic Republic of the Congo" in the base map. Thus, "Congo" and "Democratic Republic of Congo" will be added as a record in the mapping list.

Table 5
List of COVID-19 Data Collections

ID	Dataset	Region	DOI link	Starting date	Update frequency	Source
1	China daily cases with base map	China	https://doi.org/10.7910/DVN/MR5IJN	January 14, 2020	Weekly	DXY.com
2	US daily cases with base map	USA	https://doi.org/10.7910/DVN/HIDLTK	January 22, 2020	Weekly	New York Times
3	Global daily cases with base map	Global	https://doi.org/10.7910/DVN/L20LOT	January 22, 2020	Weekly	John Hopkins University
4	Baidu mobility data	China	https://doi.org/10.7910/DVN/FAEZIO	January 01, 2020	Weekly	Baidu.com
5	Health facilities	China, USA	https://doi.org/10.7910/DVN/KRSGT3			AutoNavi/US Department of Homeland Security
6	Social economics	USA	https://doi.org/10.7910/DVN/B4WHRQ			US Census Bureau
7	Climate	China	https://doi.org/10.7910/DVN/XETLSS	January 01, 2020	Weekly	China Meteorological Administration
8	Policy and regulation	China, USA, and global	https://doi.org/10.7910/DVN/OAM2JK	January 01, 2020	Weekly	BBC, CNN, China Daily, Tencent, etc.
9	Scholarly articles	Global	https://doi.org/10.7910/DVN/MHL8JC	January 01, 2020	Weekly	Web of Science

3.3 Data Sharing on Harvard Dataverse

3.3.1 Harvard Dataverse

Harvard Dataverse is an online data repository in which users can share, preserve, cite, explore, and analyze research data. A Dataverse repository contains multiple virtual archives called dataverse. Each dataverse may contain several datasets and each dataset contains descriptive metadata and data files. Each dataset has its unique DOI number, which can be used to permanently identify an article or document as a hyperlink of web. In this project, each dataverse represents a different data category, such as COVID-19 daily cases, public health, population mobility, and so on. Each dataverse includes multiple datasets. For example, in the COVID-19 Daily Cases dataverse, the datasets may include data from different countries and regions, such as the United States, China, and other countries. Once a dataset is published, it will be assigned a unique DOI number for data sharing. More data files and documents can be added to a dataset anytime. Users can preview spatial and non-spatial data with most web browsers. For spatial data, it provides a link to Harvard WorldMap for visualizing the data on map. For non-spatial data, it provides a data explorer tool for viewing the tabular data.

3.3.2 Data Organization and Sharing

The datasets collected and integrated in this project are deposited to Harvard Dataverse (https://dataverse.harvard.edu/dataverse/2019ncov). Table 5 listed available datasets on the Dataverse, including name, DOI link, start date, and the source of the dataset. There is metadata (the description of data) for each dataset, which can help understand the collected data. Due to data sharing policy by Twitter, the tweets data can only be shared on the cloud server (http://chinadatalab.org/), a private cloud-based platform for data sharing, processing, and analysis. Additionally, newly released and updated resources are added frequently to the China Data Lab cloud for direct data analysis with available tools on the cloud.

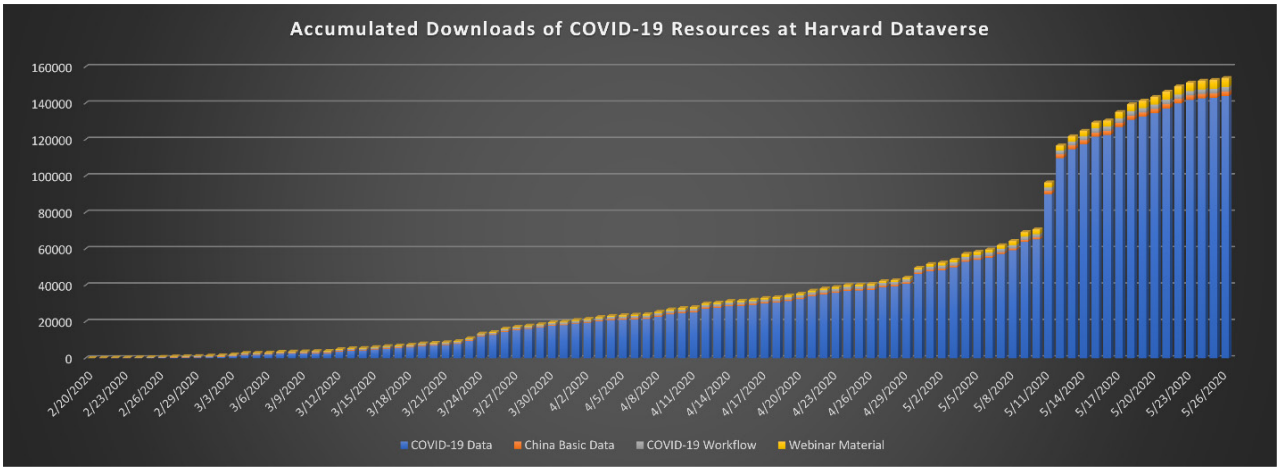

Figure 6. Accumulated downloads of "COVID-19 Resources" on Harvard Dataverse.

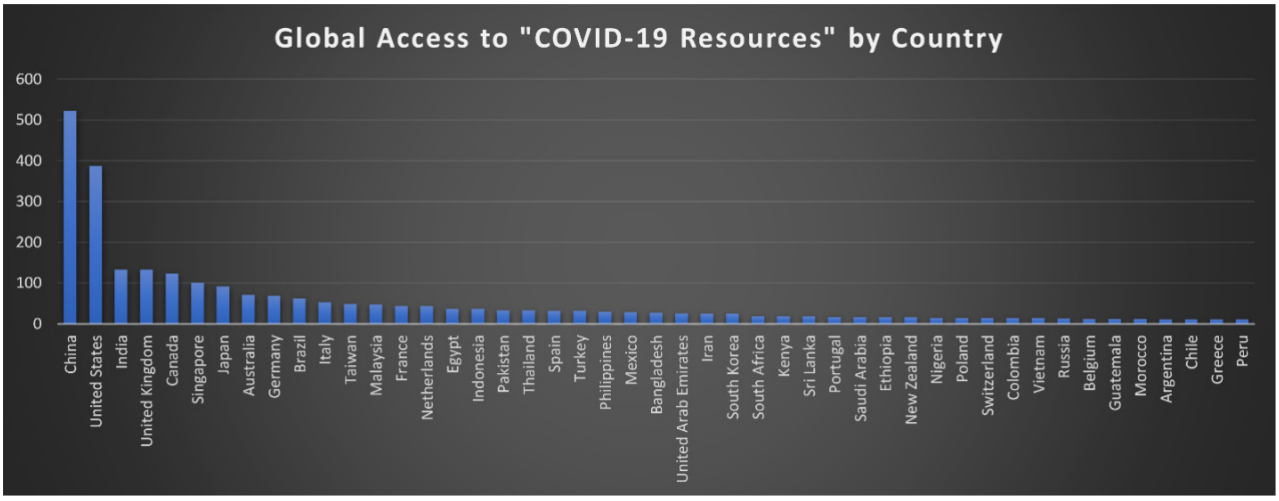

Figure 7. Global access to "COVID-19 Resources" by country by May 26, 2020.

3.3.3 Download Metrics

Harvard Dataverse provides usage statistics of each dataverse, including dataset name, actions (download, data explorer, and view data), file name, and file ID. For easy access, users are not required to log into Harvard Dataverse for downloading the data provided by this project. We published four dataverses: COVID-19 data, China basic data, COVID-19 workflows, and Webinars and training workshop materials. The China basic data includes census data, socioeconomic data, administrative boundary maps, and some other GIS map files. COVID-19 workflows provide some modules for preliminary data analysis built with KNIME, a free workflow tool developed in Germany, which allows users to easily replicate or reproduce the data analysis. Figure 6 is a bar chart for recent usage statistics for our data collections deposited to Harvard Dataverse.

Beside usage statistics for data download, Harvard Dataverse also records visitors' IP information using Google Analytics. Figure 7 summarizes the visits by countries. It shows that China, the United States, India, the United Kingdom, Canada, Singapore, Japan, Australia, Germany, and Brazil are top 10 countries for data visits. Figure 8 is the choropleth map showing global visits to our published datasets on Harvard Dataverse. As of May 26, 2020, over 120 countries have viewed our data resources pages.

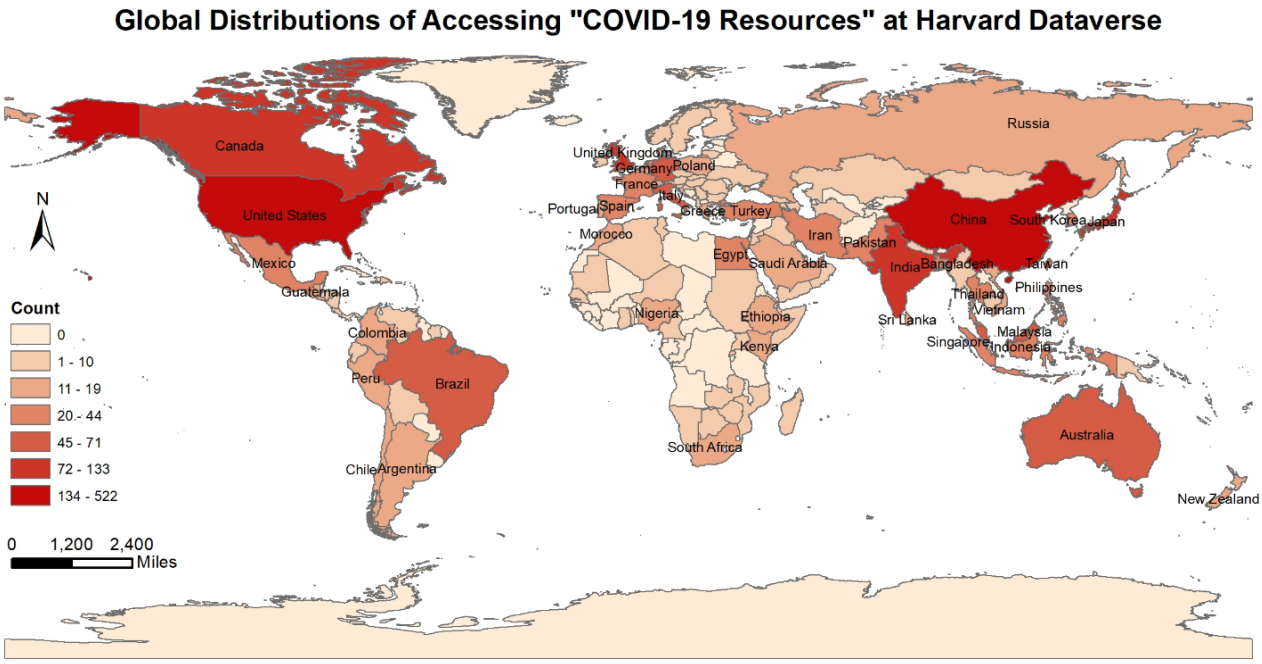

Figure 8. The global distributions of visitors to "COVID-19 Resources" on Harvard Dataverse.

4 Preliminary Analysis of COVID-19 Data

4.1 Global COVID-19 Data Analysis

As of May 23, 2020, there are 5,103,006 total confirmed cases and 333,401 death cases. The United States has been taking the lead in total infection population since the end of March 2020, which is much higher than Eastern Mediterranean, Europe, South-East Asia, Africa, Western Pacific, and other WHO regions (WHO, 2020c). The logarithmic scale chart can help visualize the exponential growth of COVID-19 cases. Figure 9 illustrates the infection cases variations in log scale of top 20 countries from January 22 to May 23, 2020. The growing trend of infection cases of different countries can be well depicted in log scale, especially when the case number passes 100. Most countries have slow start before the infected population reach 100 and then gradually became stable after 1 month. Some countries have much faster start from the beginning, including Brazil, Turkey, Chile, and Peru, which may reflect the differences in people's living habits and local policies. China is the first country that reported COVID-19 cases with a fast growth rate of confirmed cases within the first month. At the end of February, the new daily confirmed cases in China become stabilized. From the end of March, the United States, Italy, Spain, Germany, France, and Iran became the leading countries with confirmed cases and the growing curve is sharper than other countries. On April 11, 2020, the United States passed Italy to become the country with the most COVID-19 deaths. However, as a proportion of the total population in the United States, virus deaths remain at about one-sixth of those in hard-hit Italy or Spain (Grace, Karl, Veronica, & Mitchell, 2020). By May 23, 2020, 1,622,612 confirmed cases are reported in the United States, which is dramatically higher than any other countries in the second and third tiers, such as Brazil and Russia. While the confirmed cases have been stabilized in many other countries, they are still keeping growing the United States at a slower pace. With the implementation of reopening policies at different states in the United States, it will be a challenge for the government to control the spreading of COVID-19 virus in the short term.

4.2 Spatial Clustering Studies on COVID-19 Data

The spatial patterns of confirmed cases can help us identify the hot spots and understand how hot spots emerge and evolve. The Local Indicators of Spatial Association (LISA) (Anselin, 1995) has been widely used for detecting spatial

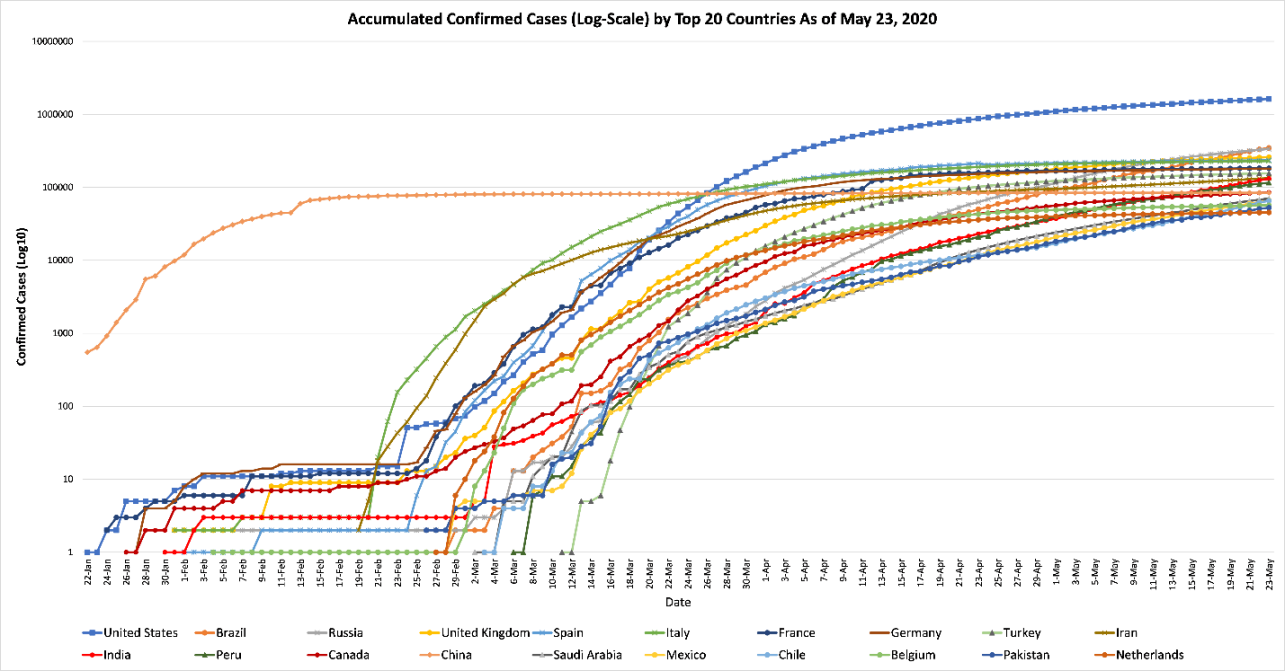

Figure 9. Accumulated confirmed cases (log scale) of COVID-19 by top 20 countries as of May 23, 2020.

autocorrelation and identifying spatial clusters. The LISA statistics can be defined as:

$$I_i = s_i \sum_j w_{ij} S_j \quad (1)$$

where s_i and s_j are the observed values of areas i and j. In this study, we applied log normalization before calculating LISA as the test with the original observations could be skewed by a few countries with extremely large number of confirmed cases. The w_{ij} represents the spatial adjacency relationship between two countries with normalized row vectors. We applied the queen-based spatial weight matrix that defines two areas as neighbors when both share a border or vertex. The LISA statistics can identify four different types of clusters, including clusters of high values (high–high), clusters of low values (low–low), low values surrounded by high values (low–high), and high values surrounded by low values (high–low). A positive LISA statistic suggests a tendency of spatial clustering. This work focuses on the high–high clustering that presents the hot spots of confirmed virus cases. The significance of the LISA statistics can be tested by using Monte Carlo simulations to generate pseudo p-values. We used GeoDa for the LISA test and set 999 as the number of permutations (Anselin, 2010).

Figures 10–12 show the spatial patterns of confirmed virus cases in mainland China, Italy, and the United States at similar time points. The starting time is chosen to be when the total number of confirmed cases reaches 100 for each country. The dates are January 18, 2020 in mainland China, February 25, 2020 in Italy, and March 2, 2020 in the United States. To explore the spatial distribution patterns of COVID-19 over time, we calculate and visualize results from LISA tests every 10 days. As shown in Figure 10, the initial hot spot in mainland China is concentrated in Wuhan. After 10 days, the hot spots start to emerge in Beijing, Shanghai, and Guangdong areas. The hot spot in Wuhan spread rapidly to nearby cities and provinces. Since then, these hot spots stay mostly unchanged, which might be related to the strict lockdown policies. The Moran coefficient starts with -0.004 on January 18, 2020, and then rise to 0.553 on January 28, 2020, indicating a strong and positive spatial autocorrelation. On February 17, 2020, the coefficient grows to 0.625 suggesting an increasingly strong spatial autocorrelation. The cases in Italy exhibit similar patterns as shown in Figure 11. The single largest hot spot remains in north Italy and is gradually expanding to nearby regions. Figure 12 presents the spatial trends in the United States. Small hot spots first appear in the states Washington and California. After 10 days, several other hot spots start to emerge in Michigan, Florida, and New York. These hot spots span rapidly and grow into large regional clusters. The hot spots continue to expand after 30 days.

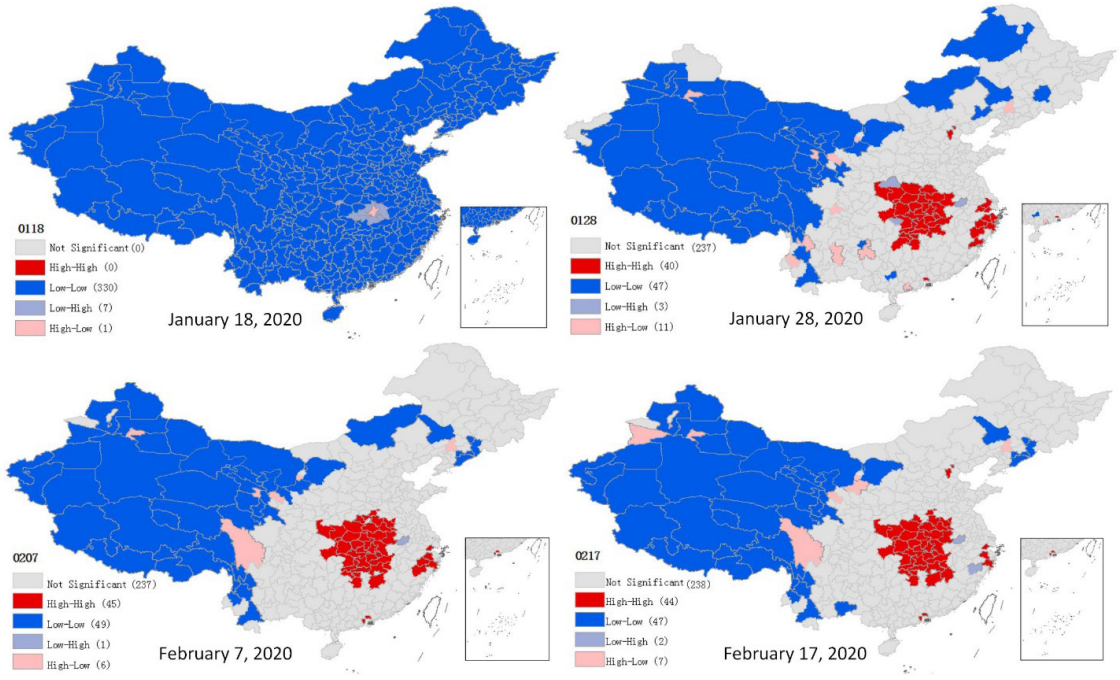

Figure 10. Spatiotemporal COVID-19 cases correlations in mainland China.

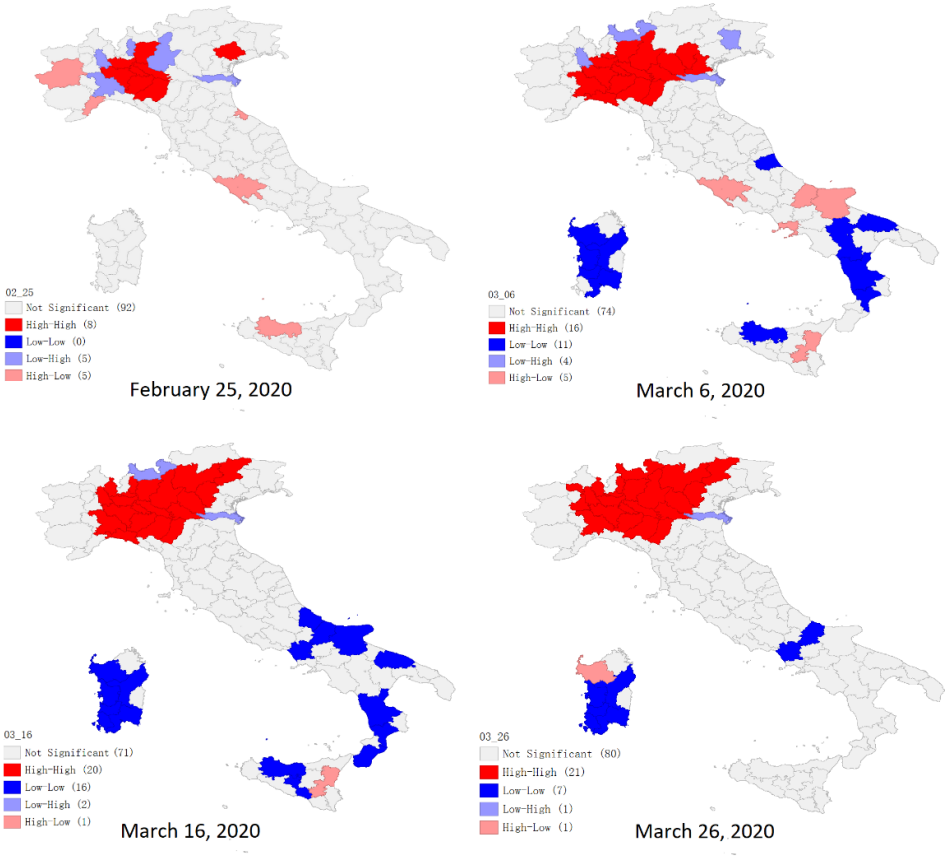

Figure 11. Spatiotemporal COVID-19 cases correlations in Italy.

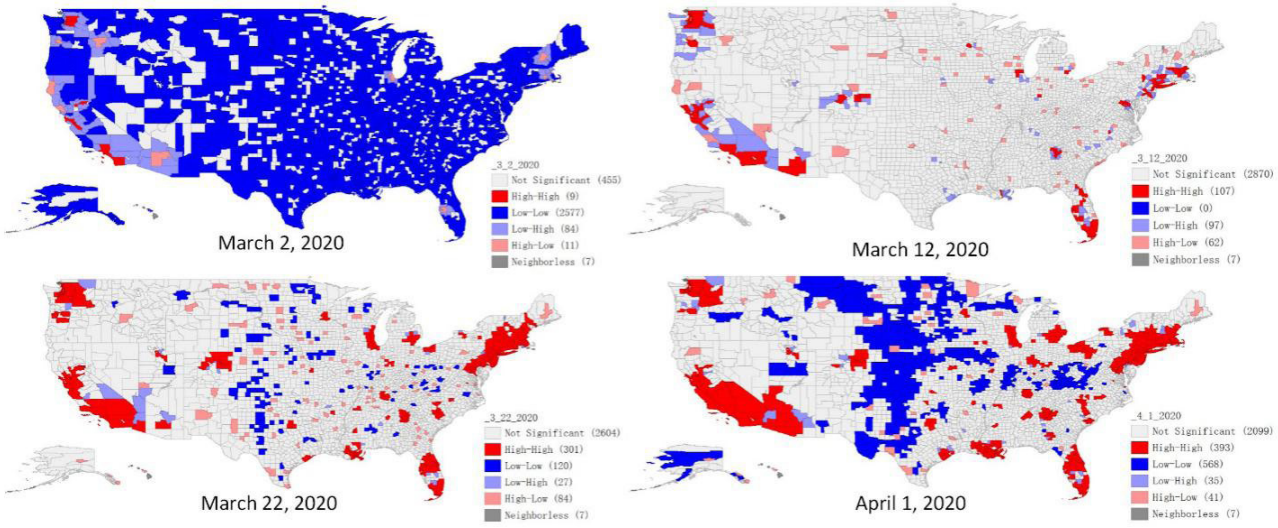

Figure 12. Spatiotemporal COVID-19 cases correlations in the United States.

5 Summary and Discussions

COVID-19 is a global pandemic of the century. Understanding its transmissibility, risk of geographic spreading, routes of transmission, risk factors for infections, and economic impacts of COVID-19 are among the critical issues for current and future control of the outbreaks (Xu et al., 2020). Data sharing allows creative innovation from archival datasets, generation of new knowledge, promotion of new discoveries, formulation of new hypotheses, creation of new meanings by connecting existing datasets, and verification of existing results (Aleixandre-Benavent, Vidal-Infer, Alonso-Arroyo, Peset, & Ferrer Sapena, 2020). This study introduces the process of collecting data from different sources, including daily COVID-19 confirmed cases, global news, social media data, population mobility, climate, health facilities, socioeconomic data, events chronicle, and scholarly articles. To keep the data files for permanent collections, we deposited the data to the Harvard Dataverse, an open online data management and sharing platform with a permanent DOI number for each dataset. Within the first 2 months, the data files have been downloaded over 160,000 times by users from more than 120 countries.

We also conducted some preliminary data analysis based on the COVID-19 datasets. The results show that China had the leading confirmed cases before March 25, 2020, while Italy started to take lead in the confirmed cases after that. European countries have more confirmed cases than other continents in the middle of March. The United States has become the leading country of the COVID-19 outbreak since the beginning of April. In addition, we applied the spatial analysis with LISA to compare the spatiotemporal trends based on the same outbreak stage in mainland China, Italy, and the United States. The results showed that a higher number of cases are centralized in Hubei Province of China and the northern region of Italy. The hot spots of confirmed cases are diversified in different locations of the United States, including California, Washington, Florida, New York, New Jersey, Massachusetts, Michigan, and Illinois. It is important to understand the driving factors of those differences, such as policies, population mobility, health facilities, locations, climate, and so on. The open data can help attract scholars from different fields, facilitate those complex studies of COVID-19 outbreaks, and help us understand future risks.

There are several limitations on our datasets:

①. *Missing values*. For example, missing values account for 3% of the final processed air quality data (there is no climate data reported for 10 out of 377 cities in China from the CNEMC data source).

②. *Base maps*. To match the census data, the China base maps used in this project was based on the 2010 census year. There have been some changes in the administrative divisions, especially in prefecture cities and counties. The census data are collected every 10 years. We may add new base maps of 2020 for our data collections once the 2020 census data is released.

Our team plans to continue updating the datasets on the Harvard Dataverse. We also plan to enrich the datasets by including more data resources and engaging more researchers who would like to share their datasets. For example, the global flight data and fast train data in China

would be a valuable addition. These data are helpful to understand the relationship between population mobility and disease outbreaks. We also plan to implement more data quality control mechanisms using various rules. Finally, we would like to collaborate with global scholars in data sharing, quality improvement, data analysis, and data applications, which may help facilitate many potential research initiatives, including spatiotemporal patterns of disease transmissions, global chain impacts, optimized capacities of health care facilities, and potential risks of future epidemic outbreaks.

Acknowledgments: This project is partially supported by NSF funding (No.: 1841403) and (No.: 2027540), Key Program of National Natural Science Foundation (No.:41830645), and CAE Advisory Project (No.: 2020-ZD-16). The Harvard Dataverse team provided instrumental support for COVID-19 resources management.

References

Aleixandre-Benavent, R., Vidal-Infer, A., Alonso-Arroyo, A., Peset, F., & Ferrer Sapena, A. (2020). Research data sharing in Spain: Exploring determinants, practices, and perceptions. *Data, 5*(2), 29.

Aminer. (2020). *COVID-19 open datasets*. Retrieved from https://covid-19.aminer.cn/

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis, 27*(2), 93-115.

Anselin, L., Syabri, I., & Kho, Y. (2010). GeoDa: An introduction to spatial data analysis. In M. M. Fischer & A. Getis(Eds.), *Handbook of applied spatial analysis* (pp. 73-89). Heidelberg, Germany: Springer.

Bank, T.W. (2020). *Understanding the coronavirus (COVID-19) pandemic through data*. Retrieved from http://datatopics.worldbank.org/universal-health-coverage/covid19/.

BigQuery Public Datasets Program. (2020). *COVID-19 public datasets*. Retrieved from https://console.cloud.google.com/marketplace/details/bigquery-public-datasets/covid19-dataset-list?preview=bigquery-public-datasets.

Bogoch, I. I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M. U., & Khan, K. (2020). Pneumonia of unknown etiology in Wuhan, China: Potential for international spread via commercial air travel. *Journal of Travel Medicine, 27*(2), 1-3.

Bureau, U.S.C. (2020). *Census data*. Retrieved from https://data.census.gov/cedsci/.

Chen, E., Lerman, K., & Ferrara, E. (2020). *Covid-19: The first public coronavirus twitter dataset. ArXiv Preprint*. arXiv: 2003.07372. Retrieved from https://arxiv.org/pdf/2003.07372.pdf

China, M., Ministry of Environmental Protection of the Peoples' Republic of China. (2015). *2015 report on the state of the environment in China*. Retrieved from http://english.mee.gov.cn/Resources/Reports/soe/Report/201706/P020170614504782926467.pdf

China, M.o.E.P.o. (2020). *Technical regulation on ambient air quality index (HJ633-2012)*. Retrieved from http://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/jcffbz/201203/t20120302_224166. shtml

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., & Viboud, C. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science, 368*(6489), 395-400.

Covid, C. D. C., & Team, R. (2020). Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep, 69*(12), 343-346.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases, 20*(5), 533-534.

Foundation, K.F. (2020). *State data and policy actions to address coronavirus*. Retrieved from https://www.kff.org/report-section/state-data-and-policy-actions-to-address-coronavirus-sources/.

Fujita, H. (2013). Geo-tagged Twitter collection and visualization system. *Cartography and Geographic Information Science, 40*(3), 183-191.

Grace H., Karl G., Veronica B., & Mitchell T. (2020) Three months in: A timeline of how COVID-19 has unfolded in the US. *USA TODAY*. Retrieved from https://www.usatoday.com/in-depth/news/nation/2020/04/21/coronavirus-updates-how-covid-19-unfolded-u-s-timeline/2990956001/

Grzybowski, P., Juralewicz, E., & Piasecki, M. (2019, September). Sparse coding in authorship attribution for Polish tweets. *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP 2019), 409-417. Shoumen, Bulgaria: INCOMA Ltd. doi: 10.26615/978-954-452-056-4_048

Henrik, P., Byron, M., & Sergio, H. (2020, June 18). Tracking coronavirus' global spread. *CNN Health*. Retrieved from https://www.cnn.com/interactive/2020/health/coronavirus-maps-and-cases/.

Hu, T., She, B., Duan, L., Yue, H., & Clunis, J. (2019). A systematic spatial and temporal sentiment analysis on Geo-Tweets. *IEEE Access, 8*, 8658-8667.

Lai, S., Bogoch, I. I., Ruktanonchai, N. W., Watts, A., Lu, X., Yang, W., & Tatem, A. J. (2020). Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: A travel network-based modelling study. *MedRxiv*. doi:10.1101/2020.02.04.20020479

Lin, I. (2020). *DXY-COVID-19-Data*. Retrieved from https://github.com/BlankerL/DXY-COVID-19-Data

Liu, Q., Sha, D., Liu, W., Houser, P., Zhang, L., Hou, R., & Yang, C. (2020). Spatiotemporal patterns of COVID-19 impact on human activities and environment in mainland China using nighttime light and air quality data. *Remote Sensing, 12*(10), 1576.

Luo, W., Majumder, M., Liu, D., Poirier, C., Mandl, K., Lipsitch, M., & Santillana, M. (2020). The role of absolute humidity on transmission rates of the COVID-19 outbreak. *MedRxiv*. doi:10.1101/2020.02.12.20022467

Merlière, A.A.A.B.C.M.G.J.D.M.F.P.R.T.P.-N.T. (2020). *OpenCOVID19 France*. Retrieved from https://github.com/opencovid19-fr.

Remuzzi, A., & Remuzzi, G. (2020). COVID-19 and Italy: What next? *The Lancet, 395*(10231), 11-17.

Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. (2020). *Temperature and latitude*

analysis to predict potential spread and seasonality for COVID-19. Retrieved from SSRN 3550308.

Scholar, S. (2020). *COVID-19 open research dataset (COVID-19)*. Retrieved from https://pages.semanticscholar.org/coronavirus-research.

Tian, H., Li, Y., Liu, Y., Kraemer, M. U., Chen, B., Cai, J., & Cui, Y. (2020). Early evaluation of Wuhan city travel restrictions in response to the 2019 novel coronavirus outbreak. *Medrxiv*. Retrieved from https://www.medrxiv.org/content/medrxiv/early/2020/02/02/2020.01.30.20019844.full.pdf

Times, T.N.Y. (2020). *Coronavirus (Covid-19) data in the United States*. Retrieved from https://github.com/nytimes/covid-19-data.

University of Oxford. (2020). *COVID-19 government response tracker*. Retrieved from https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker.

Wang, J., Tang, K., Feng, K., & Lv, W. (2020). *High temperature and high humidity reduce the transmission of COVID-19*. Retrieved from http://dx.doi.org/10.2139/ssrn.3551767.

Warren, M. S., & Skillman, S. W. (2020). Mobility changes in response to COVID-19. *ArXiv Preprint*. ArXiv: 2003.14228. Retrieved from https://arxiv.org/pdf/2003.14228.pdf

Wells, C. R., Sah, P., Moghadas, S. M., Pandey, A., Shoukat, A., Wang, Y., & Galvani, A. P. (2020). Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences, 117*(13), 7504-7509.

W.H.O. (2020a). *Coronavirus disease (COVID-2019) situation reports*. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

W.H.O. (2020b). *WHO announces COVID-19 outbreak a pandemic*. Retrieved from http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic.

W.H.O. (2020c). *Global research on coronavirus disease (COVID-19)*. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov

W.H.O. (2020d). *Coronavirus disease (COVID-2019) situation reports*. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200523-covid-19-sitrep-124.pdf?sfvrsn=9626d639_2.

Worldometer. (2020). *COVID-19 coronavirus pandemic*. Retrieved from https://www.worldometers.info/coronavirus/#countries.

Wu, T., Ge, X., Yu, G., & Hu, E. (2020). Open-source analytics tools for studying the COVID-19 coronavirus outbreak. *MedRxiv*. doi:10.1101/2020.02.25.20027433

Xu, B., Kraemer, M. U., Gutierrez, B., Mekaru, S., Sewalk, K., Loskill, A., & Li, S. (2020). Open access epidemiological data from the COVID-19 outbreak. *The Lancet Infectious Diseases, 20*(5), 534.