China Data Institute | Center for Geographic Analysis, Harvard University | 社会地理计算联合研究中心 Geocomputation Center for Social Science | Spatiotemporal Innovation Center

An Overview of Research Data for COVID-19 Modeling

Wendy Guan[1], Tao Hu[1], Xinyan Zhu[2], Yuanzhen Shao[2], Xiaokang Fu[2], Chaowei Yang[3], Dexun Sha[3] , Qian Liu[3], Yanfang Su[4], Xiaocheng Dai[4] and Shuming Bao[5]

[1]Harvard University

[2]Wuhan University

[3]George Mason University

[4]University of Washington

[5]China Data Institute and Future Data Lab

About CDL

Establishment

China Data Lab

A cloud-based geospatial data analysis platform for geospatial data gathering, management, analysis, visualization, and sharing.

Sponsored by the Spatiotemporal Innovation Center of the NSF Industry-University Cooperative Research Centers (I/UCRC) Program

CGA

The Center for Geographical Analysis (CGA) at Harvard University. Its core mission is to support research and teaching in all disciplines across Harvard University with emerging geospatial technologies.

CDI

The China Data Institute, a Michigan based not-for-profit organization. It aims to promote the use and sharing of China data; support quantitative research on China in social science, digital humanity and other research subjects.

GCSS

The GeoComputation Center for Social Science at Wuhan University. It promotes the scientific research on the theory and method of spatial data in scientific research, personnel training, international cooperation and social practice.

Partners for Data, Tools and Case Studies

esri China BEIJING THE SCIENCE OF WHERE · CSMAR · ACMR 北京华通人商用信息有限公司 · GAUSS · Knowledge@Share 智享

Challenges and Objectives

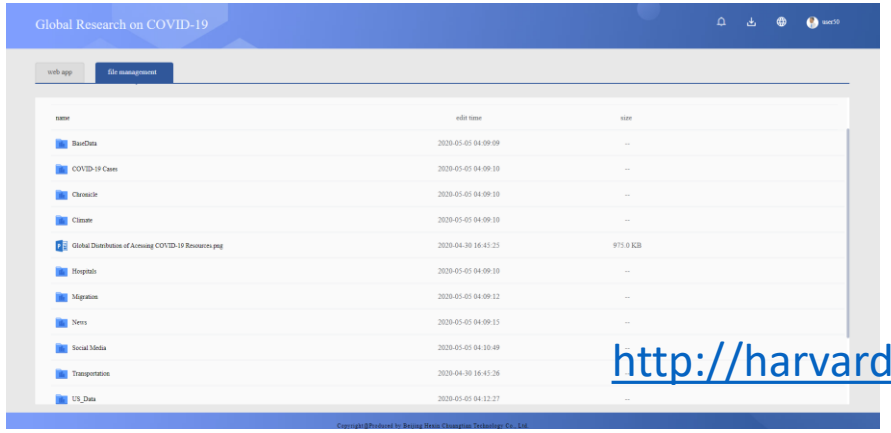
- **Data Sharing**
  - Licensed data
  - Restricted data
  - Sensitive data
  - Large size data
- **Tool Sharing**
  - Licensed Tool
  - Maintenance and updates
  - Operating environment
- **Workflow/Research Sharing**
  - Discoverable
  - Replicable
  - Reusable
  - Modifiable and expandable
- **Dissemination**
  - Publication
  - Teaching
  - Decision support

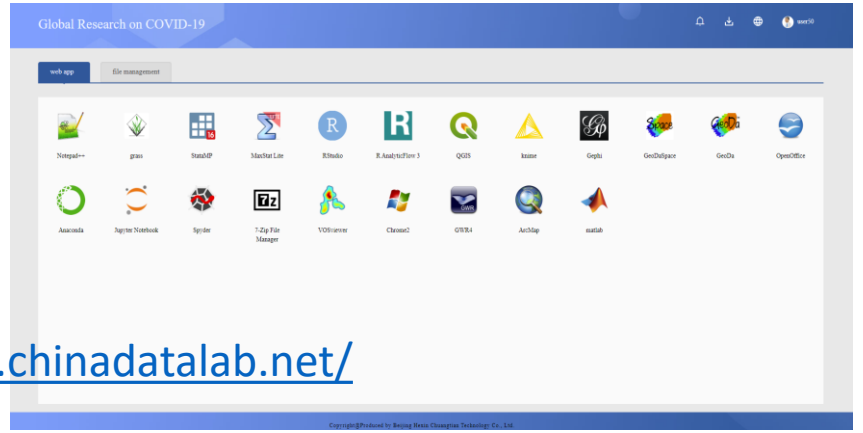
China Data Lab Platform

http://chinadatalab.net

- A Data Library
- An Analytical Lab
- A Collaboration Hub
- A Training Center

Alteryx GAUSS GeoDa GWR4 ArcMap Jupyter RAnalyticFlow KNIMEAnalytics





http://harvard.chinadatalab.net/

Advisory Committee

Jason Ur, **Committee Chair**
Professor of Archaeology
Director of the Center for Geographic Analysis Harvard University

Peter K. Bol
Charles H Carswell Professor
Dept of East Asian Languages and Civilizations
Harvard University

Luc Anselin
Professor of Sociology
Director, Center for Spatial Data Science
University of Chicago

Daniel Sui
Distinguished Professor of Geography
Vice Chancellor for Research and Innovation University of Arkansas

Peng Gong
Professor, Department of Earth System Science
Dean, School of Science
Tsinghua University

Yasheng Huang
Epoch Foundation Professor of International Management
Professor of Global Economics and Management, Massachusetts Institute of Technology

Gary King
Weatherhead University Professor
Director of the Institute for Quantitative Social Science
Harvard University

Peter X Zhou
Director and Assistant University Librarian
C.V. Starr East Asian Library
University of California, Berkeley

Yu Xie
Professor of Sociology
Institute of International and Regional Studies
Princeton University

Pinde Fu
Platform Engineering Team Lead, ESRI
Adjunct Faculty at University of Redlands and Harvard Extension School

Resources for COVID-19 Research

HARVARD UNIVERSITY HARVARD.EDU

China Data Lab @Center for Geographic Analysis, IQSS

HOME People Resources Partners Events About

FILTER NEWS BY MONTH

May 2020 (1)
March 2020 (1)
December 2019 (3)
November 2019 (1)
October 2019 (1)

1 of 2 »

New Releases and Updates for COVID-19 Resources

May 20, 2020

"Resources for Novel Coronavirus and Global Research" is a new initiative on global research on COVID-19 sponsored by the China Data Lab project.

Below are new releases and updates of data collection for COVID-19 study on Resources for COVID-19 at Harvard Dataverse data archive.

New Releases through 2020-05-20

- US Metropolitan Daily Cases with Basemap: This dataset includes US Metropolitan level daily confirmed and death cases from **Jaunary 22, 2020** to **May 18, 2020**
  https://doi.org/10.7910/DVN/5B8YM8
- Climate: The climate dataset includes time-series' air quality data in China cities: AQI, CO, NO2, O3, PM10, PM2.5, and SO2. It is started from **January 1, 2020** to **May 16, 2020** and each record has daily max, min, and Standard Deviation value.
  https://doi.org/10.7910/DVN/XETLSS
- South Africa COVID-19 Case Data with Basemap (STC): This dataset includes South Africa admin1 level daily confirmed, recovered, death cases, and summary reports from **03-05-2020** to **05-17-2020**.
  https://doi.org/10.7910/DVN/AKYTCM
- Italy COVID-19 Case Data with Basemap (STC): This dataset includes Italy admin1 level daily confirmed, recovered, death cases, and summary reports from **02-24-2020** to **05-17-2020**.
  https://doi.org/10.7910/DVN/4Z8ZKI
- Russia COVID-19 Case Data with Basemap (STC): This dataset includes Russia admin1 level daily confirmed, recovered, death cases, and summary reports from **03-22-2020** to **05-17-2020**.
  https://doi.org/10.7910/DVN/ZR5KW9

HARVARD UNIVERSITY HARVARD.EDU

China Data Lab @Center for Geographic Analysis, IQSS

HOME People Resources Partners Events About

RESOURCES

- Enter the Lab
- Resources for COVID-19
- Publications
- Presentations
- Data Case Studies

Resources for COVID-19

COVID-19 HARVARD Dataverse

- Resources for COVID-19 at Dataverse
- User Guide for "Resources for COVID-19"

- Cloud for Global Research on COVID-19
- User Guide for "Cloud Platform for COVID-19"

CALL FOR PAPERS COVID-19

- **China Economic Review:** Special Issue on *Virtual Pandemic Disease*
- **Journal of Data Science:** Special Issue on *Data Science in Action in Response to the Outbreak of COVID-19*
- **Transportation Research Interdisciplinary Perspectives:** Special Issue on *Transportation Impacts and Responses to the Coronavirus (COVID-19)*
- **Technological Forecasting and Social Change:** Special Issue on *Social-Economic Impacts of Epidemic Diseases*
- **Geospatial Approaches for Understanding the Social, Economic and Environmental Impacts of COVID-19**: Special Issue on *IJGI*

Free User Account on Cloud Platform for Global Research on COVID-19

- Guideline for Applications
- **Apply online:**
  - The Online Application in Chinese
  - The Online Application in English
- **Download the application form:**
  - The application form in Chinese
  - The application form in English

COVID-19 Data Challenges and Objectives

❑Challenges

- Fragmented information
- Incomparable and lack of standard
- Separate from maps
- Temporary data connection
- Repeated work with low efficiency

❑Objectives

- Integration of data from different sources
- Standardization of data formats
- Integration of data and maps
- Permanent collections
- Improvement with efficient technology
- Collaboration on COVID-19 studies

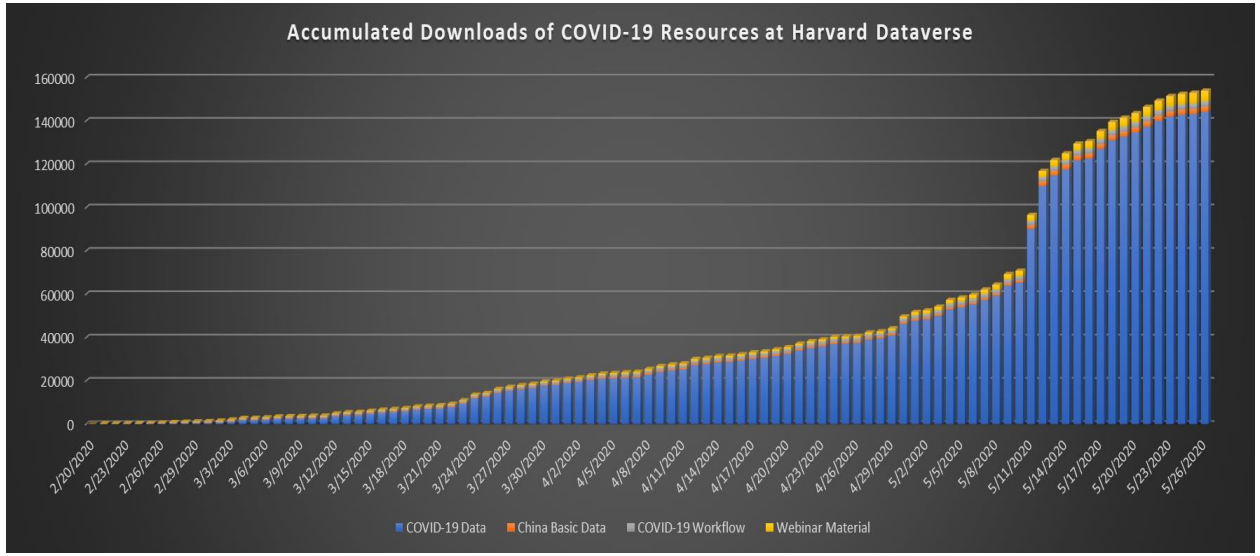
Data Collection and Integration

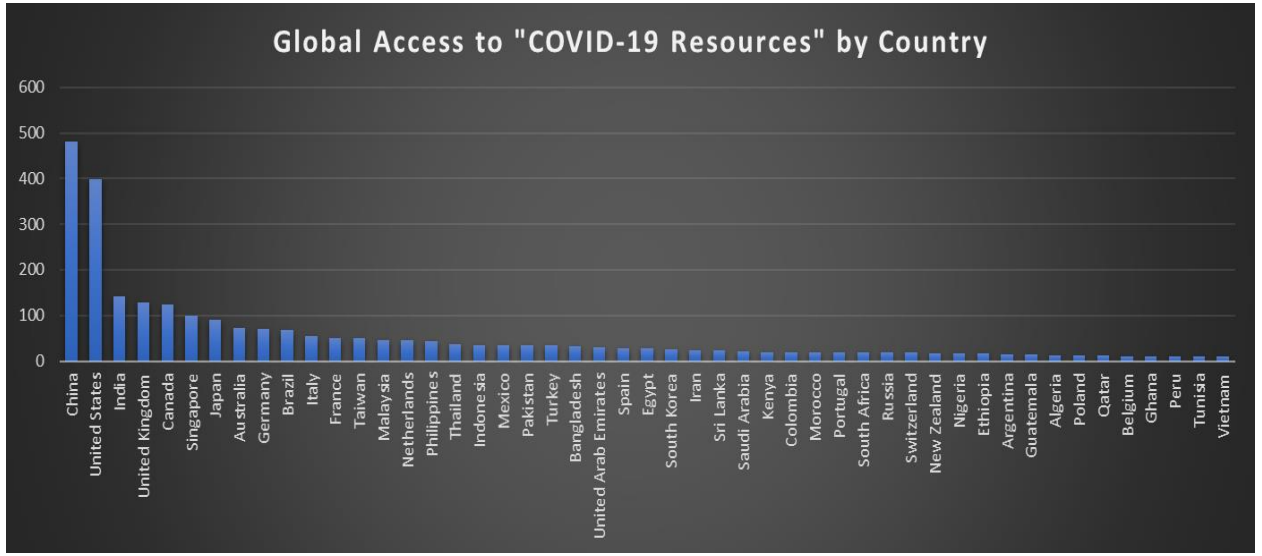
- ❑ Data standardization and association
- ❑ Rich spatial variables for geographic research

ID	Data Sets
1	Coronavirus cases
2	Population mobility
3	Health facilities
4	Traces
5	Flights
6	High-speed train
7	Global News
8	Social media
9	Policies
10	Meteorological data
11	Air Quality
12	Socioeconomic Data

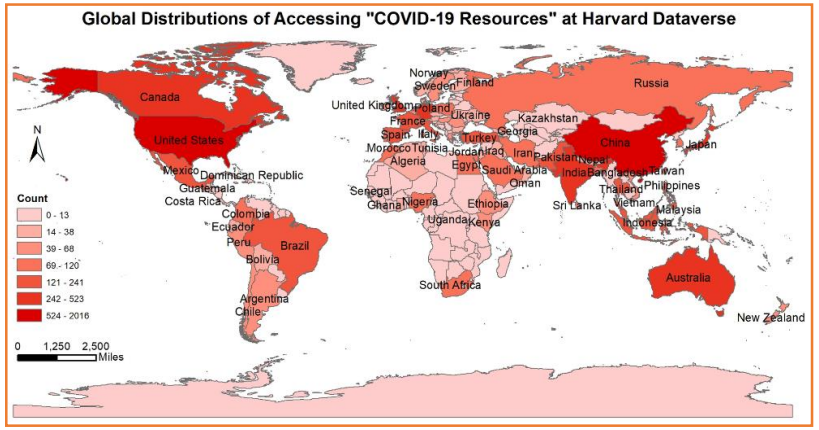
COVID-19 Resources Access

Global users from over ***150*** countries & regions with over ***230,000*** data downloads



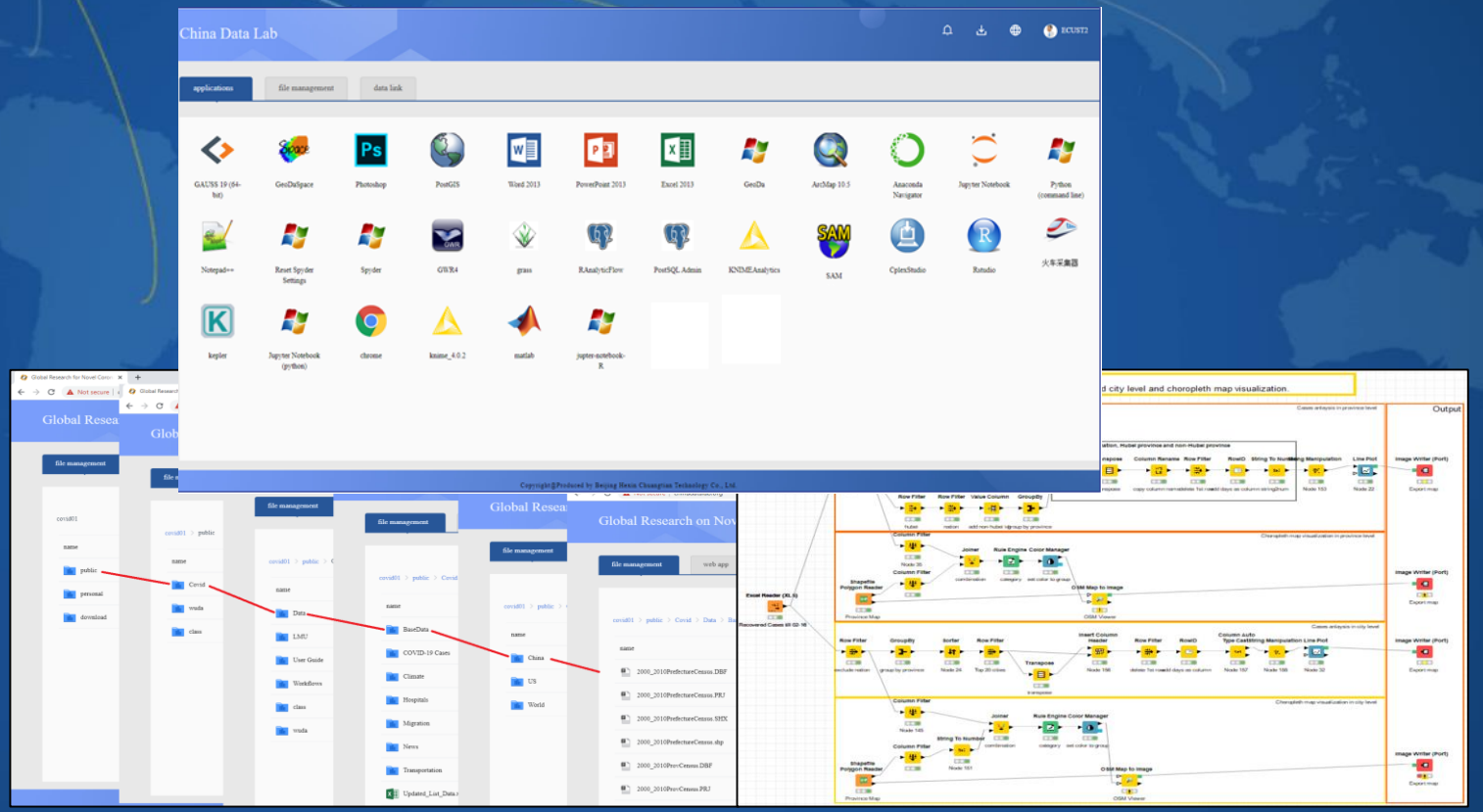


- Over ***120*** researchers from ***9*** countries applied for using the CDL platform for COVID-19 research, including *Oxford University*, *University of Illinois at Chicago*, *University of Maryland at College Park*, and so on.
- Research topics cover health inequality, economic loss, public opinions, correlation analysis between climate and virus spreading during COVID-19 epidemic, and so on.



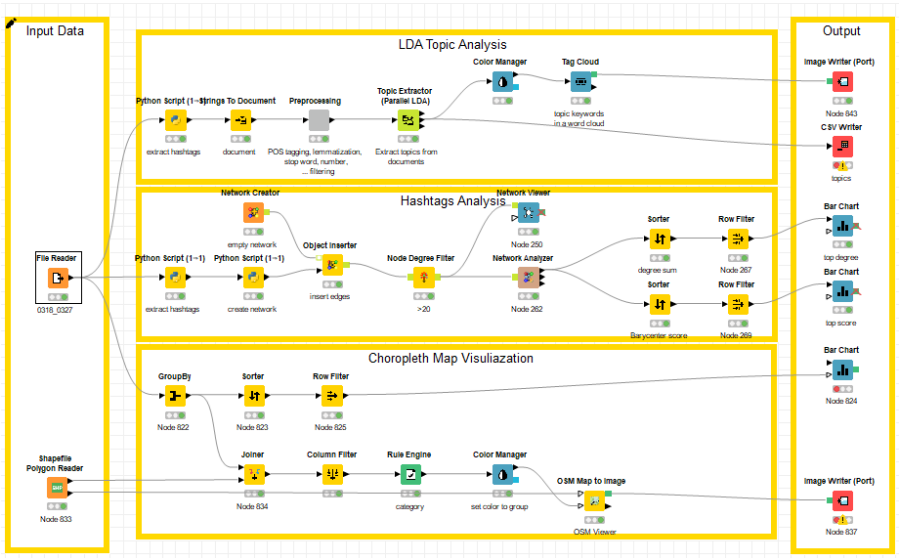
Methodology and Technology

Global Research on Novel Coronavirus

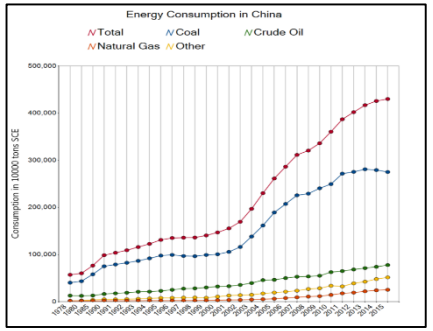


Research & Development

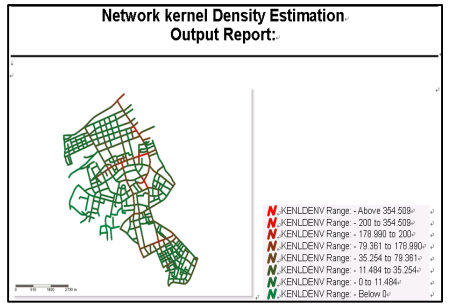
- **Workflow-based Data Analysis**
  - Low barrier to study
  - Easy to operate
  - Modularity
  - Easy to reproduce
  - Easy to extend



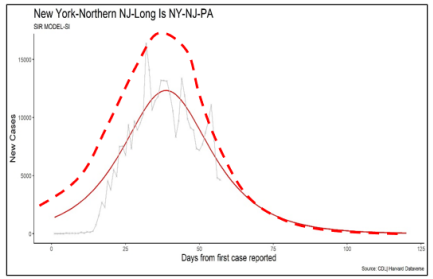
Workflow



Environment

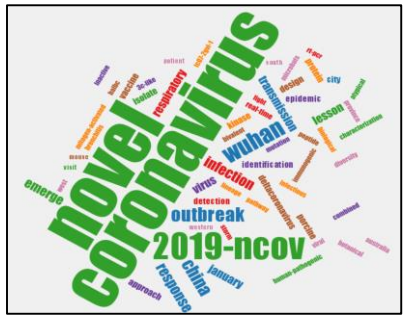


Transportation

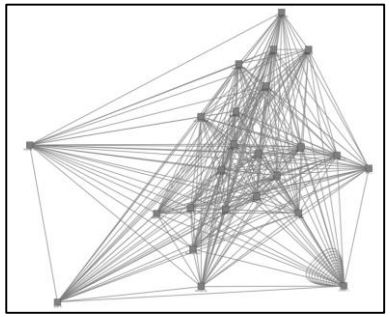


Public Health

Disease Transmission



Social-media

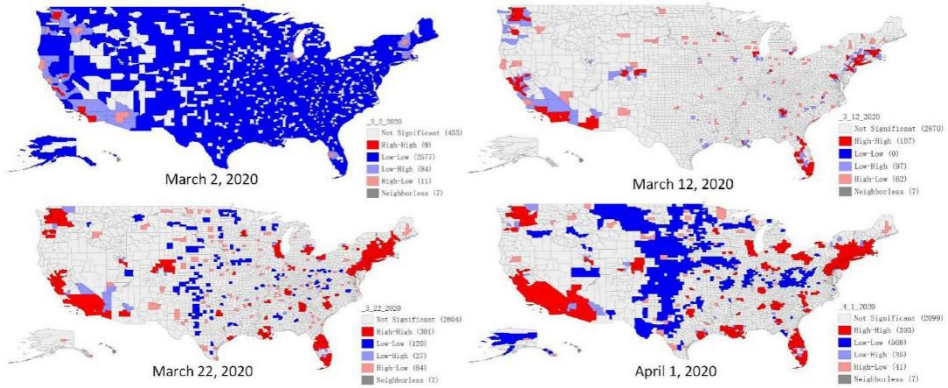


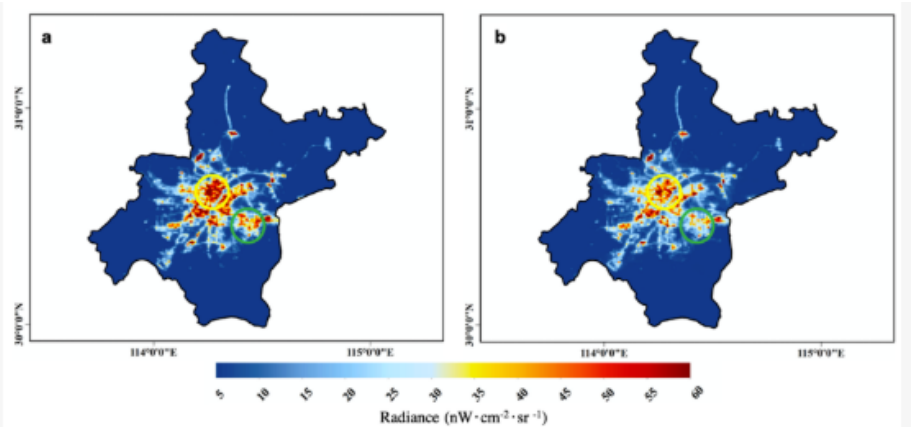
Publication

Grant and Publication

- **NSF RAPID Grant (Harvard, GMU and CDI)**
- **Published and Preprint Papers**

ID	Title	Journal/Preprint
1	Build an Open Repository for COVID-19 Research	Data and Information Management
2	Taking the pulse of COVID-19: A spatiotemporal perspective	International Journal of Digital Earth (Under Review)
3	Spatiotemporal Patterns of COVID-19 Impact on Human Activities and Environment in Mainland China Using Nighttime Light and Air Quality Data	Remote Sensing
4	Spatiotemporal Analysis of Medical Resource Deficiencies in the U.S. under COVID-19 Pandemic	Plosone (Under Review)
5	Spreading of COVID-19: Density matters	Plosone
6	Racial Segregation, testing sites accessibility, and COVID-19 incidence rate in cities/towns of Massachusetts: a spatial regression approach	In Draft





- The datasets of “Resources for COVID-19 Research” are cited by many research studies and organizations, such as Domino Data Lab; UCGIS; Emory University Libraries, The World Bank/IMF Library, George Washington University Library, NTU Library et al.

Data Overview for COVID-19 Modeling

- Categories of Models and Data
  - COVID-19 Models
  - Applications
  - Data Categories for Modeling
- Data Sources
  - Data Sources for Quantitative Index
  - Data Sources for Derivable Index
- COVID-19 Models on CDL Platform (in progress)
  - Benchmark Datasets
  - Model Replication with Workflows
- Summary

Categories of COVID-19 Models

By Objectives:

- Prediction of Trends
- Health Facility Evaluation
- Policy Evaluation
- Impact Evaluation

By Methodology:

- Epidemiology Model
- Agent-based Model
- Machine Learning Model
- Regression Model

COVID-19 Predictive Models Reported to CDC

ID	Group	Model Names	Methods
1	Auquan Data Science	Auquan	SEIR model
2	Columbia University	Columbia	Metapopulation SEIR model
3	COVID Act Now	CAN	SEIR model
4	COVID-19 Simulator Consortium	Covid19Sim	SEIR model
5	Georgia Institute of Technology, Center for Health and Humanitarian Systems	GT_CHHS	Agent-based model
6	Georgia Institute of Technology, College of Computing	GT-DeepCOVID	Deep learning.
7	Imperial College, London	Imperial	Ensembles of mechanistic transmission models
8	Institute of Health Metrics and Evaluation	IHME	Combing a mechanistic disease transmission model
9	Iowa State University	ISU	Nonparametric spatiotemporal model
10	Johns Hopkins University, Infectious Disease Dynamics Lab	JHU	Stochastic metapopulation SEIR model
11	London School of Hygiene and Tropical Medicine	LSHTM	Renewal equation
12	Los Alamos National Laboratory	LANL	Statistical dynamical growth model
13	Massachusetts Institute of Technology, COVID-19 Policy Alliance	MIT-CovAlliance	SIR model
14	Massachusetts Institute of Technology, Operations Research Center	MIT-ORC	SEIR model
15	Northeastern University, Laboratory for the Modeling of Biological and Socio-technical Systems	MOBS	Metapopulation, age-structured SLIR model
16	Notre Dame University	NotreDame-Mobility NotreDame-FRED	NotreDame-Mobility: SEIR model NotreDame-FRED: Agent-based model
17	Oliver Wyman	Oliver Wyman	Time-dependent SIR model
18	Predictive Science Inc.	PSI	Stochastic SEIRX model
19	Snyder Wilson Consulting	SWC	Bayesian SEIR model
20	US Army Engineer Research and Development Centertxt icon	ERDC	SEIR model
21	University of Arizona	UA	SIR mechanistic model
22	University of California, Los Angeles	UCLA	Modified SEIR model
23	University of Geneva / Swiss Data Science Center	Geneva	Exponential and linear statistical models
24	University of Massachusetts, Amherst	UMass-MB, Ensemble	UMass-MB: Bayesian SEIRD model
25	University of Southern California	USC	SIR Model
26	University of Texas, Austin	UT	Nonlinear Bayesian hierarchical regression
27	Youyang Gu	YYG	SEIS mechanistic model

Data Inputs for Models

IHME

University of Washington

- Daily deaths
- Case data
- Testing data
- Mobility data
- Pneumonia seasonality
- Mask use
- Population density
- Air pollution
- Low altitude

http://www.healthdata.org/

COVIDScenarioPipeline

John Hopkins University

- JHU-CSSE COVID-19 Dashboard
- USAFacts
- Airport Coordinates
- Mobility matrix
- non-pharmaceutical interventions
- Hospital and ICU admissions
- COVID Severity: age distribution

https://github.com/HopkinsIDD/COVIDScenarioPipeline

GT_CHHS

Georgia Institute of Technology

- Case data
- Demographics
- Socioeconomics
- Policy: school closure, shelter in place, voluntary quarantine
- Hospital beds
- ICU beds
- Ventilators
- Disease risk factors

https://github.com/pkeskinocak/COVID19GA

GT-DeepCOVID

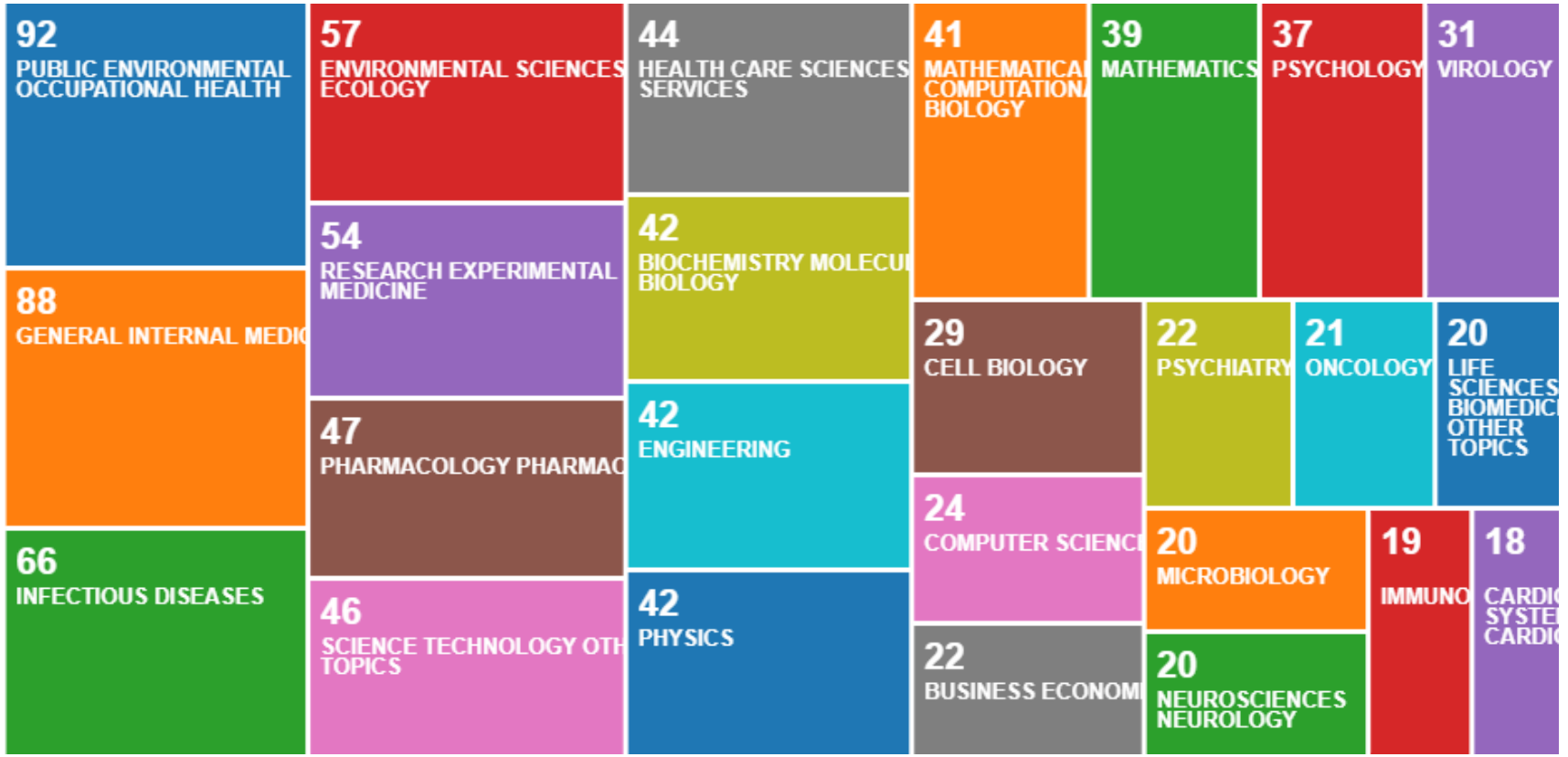
Georgia Institute of Technology

- JHU Case data
- Demographics
- Socioeconomics
- Mobility
- Symptomatic Data
- Point-of-care claims data

https://www.cc.gatech.edu/~badityap/covid.html

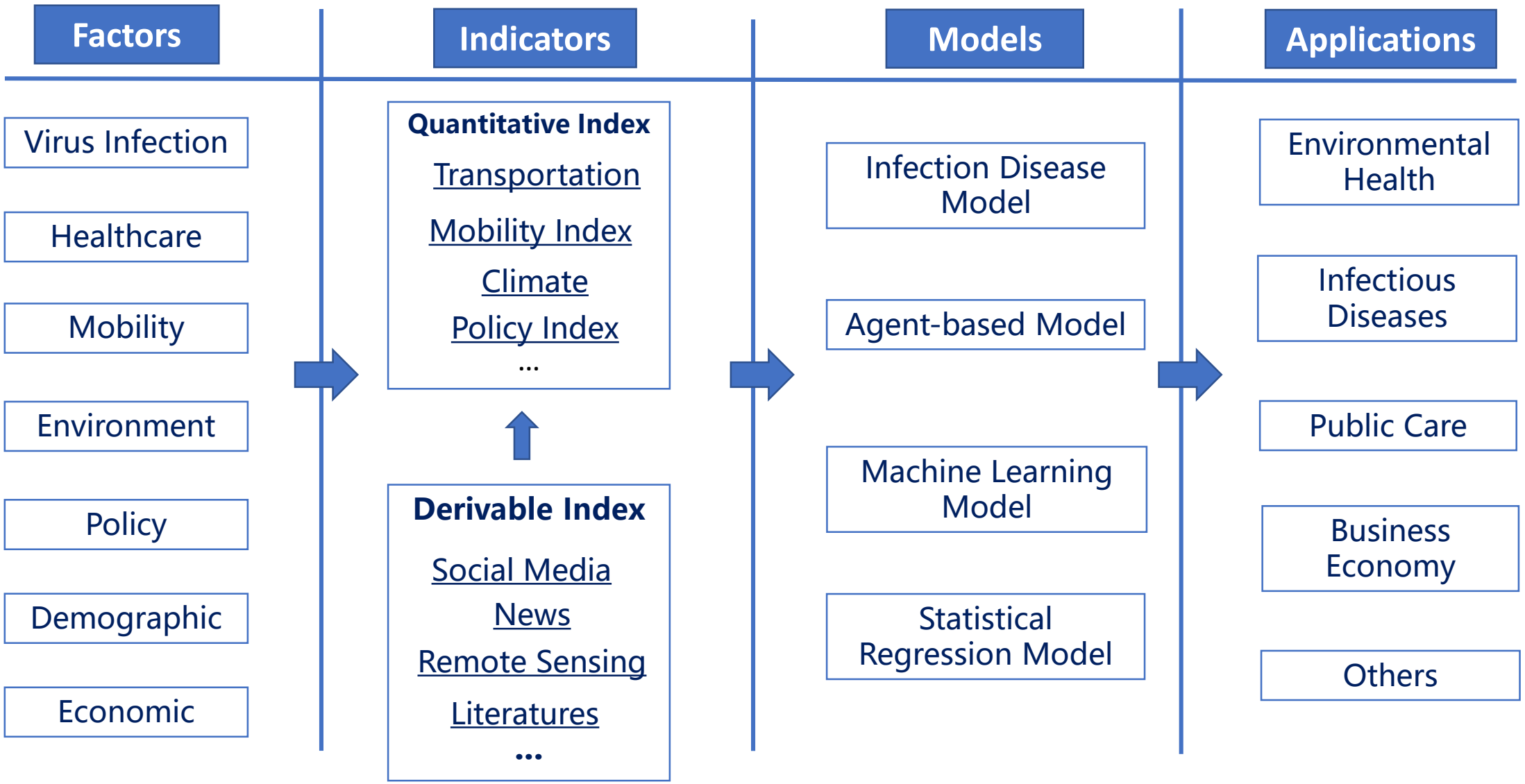
Model Applications from Text Analysis

- Environmental Health
- Infectious Diseases
- Public Care
- Life Science
- Business Economy
- Others



Source: www.webofknowledge.com (keyword: covid-19 + model)

The Flowchart of COVID-19 Data Analysis



Data Sources for Quantitative Index

- COVID-19 Cases
- Health Data
- Demographics and Economics Data
- Mobility Data
- Policy Data
- Environmental Data

COVID-19 Cases Data

■ Worldwide COVID-19 Cases Data

Source: https://ourworldindata.org/covid-sources-comparison

❑ WHO (World Health Organization)

- **Link:** *WHO Situation Reports*
- **Reporting time and frequency:** The daily Situation Reports are published every day;
- **Metrics covered:** Total confirmed cases/deaths; daily new cases/deaths
- **Underlying source of data:** Direct reports from national governments

Pros: higher reliability;
Cons: single data source

❑ ECDC (Europe CDC)

- **Link:** *COVID-19 Dashboard; Situation reports; The daily data tables*
- **Reporting time and frequency:** every day;
- **Metrics covered:** daily new cases; daily new deaths;
- **Underlying source of data:** Direct reports from national governments

Pros: multiple data download format;
Cons: only cases and deaths

❑ Johns Hopkins University

- **Link:** *COVID-19 Dashboard*
- **Reporting time and frequency:** every day;
- **Metrics covered:** Total confirmed cases/deaths; daily new cases/deaths; recovered;
- **Underlying source of data:** *China CDC (CCDC), Hong Kong Department of Health, Macau Government, Taiwan CDC, ECDC, WHO, US CDC*

Pros: integrate and verify multiple data sources ;
Cons: higher numbers (presumptive positive cases)

COVID-19 Cases Data

Source: https://source.opennews.org/articles/comparison-four-major-covid-19-data-sources

The US COVID-19 Cases Data

- Johns Hopkins University (JHU)
- USAFacts
- The New York Times (NYT)
- COVID Tracking Project

	U.S. states	U.S. counties	Global cases	Total tests	ICU / Ventilator use	Hospitalizations
Johns Hopkins University *	✓	✓	✓	✓		✓
Covid Tracking Project	✓			✓	✓	✓
USAFacts	✓	✓				
The New York Times	✓	✓				

No.	Difference Aspects	Descriptions
1	Reporting Time	JHU and NYT update later than the others (higher cases)
2	Death Cases	COVID Tracking Project is not yet including ***presumed deaths*** if they are reported separately
3	Geographic Area	**USAFacts** reports New York City data broken down into the ***five*** boroughs that make up the city, while other sources report one number for the entire city.
4	Data Quality	**COVID Tracking Project** awards each state a letter grade between A to D based on the state's reliability and completeness.

Hospital Data

- **Data Source**: Definitive Healthcare - Hospital data
- **Download**: https://coronavirus-resources.esri.com/datasets/definitivehc::definitive-healthcare-usa-hospital-beds/
- **Format**: Point-based Shapefile (POI)
- **Category**: Health Data

Index:

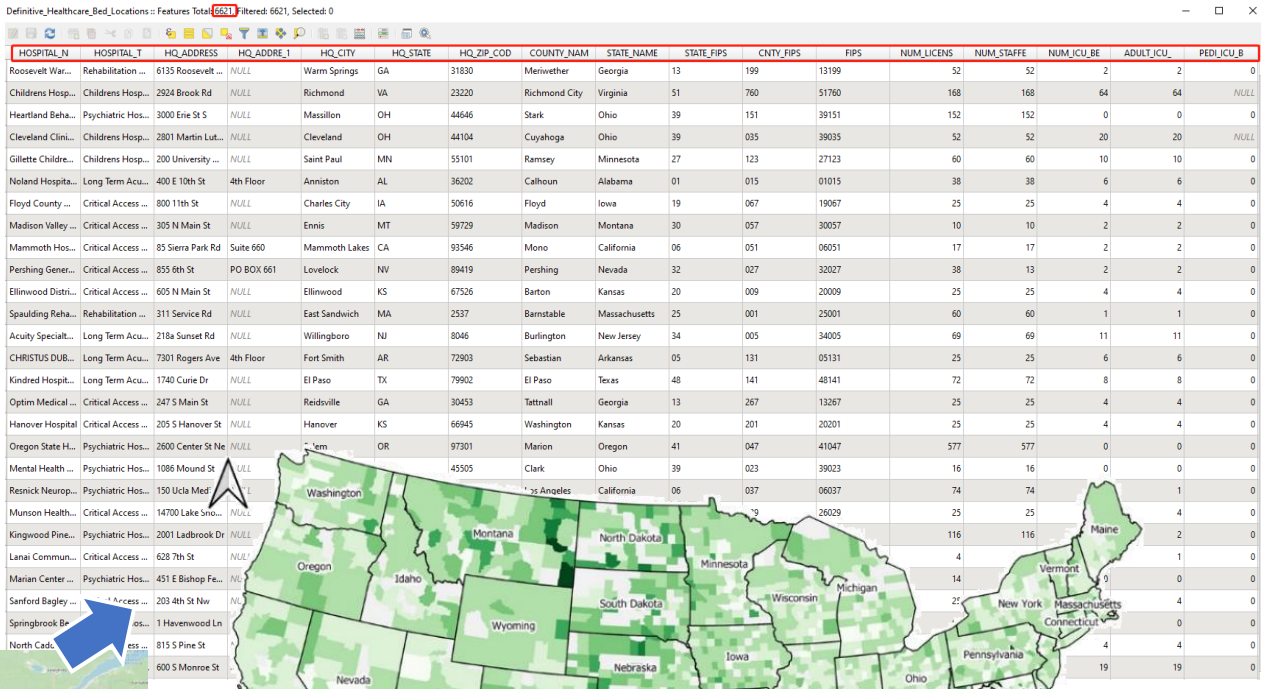
- **Number of Licensed Beds**
- **Number of Staffed Beds**
- **Number of ICU Beds - ICU (Intensive Care Unit) Beds**
- **Adult ICU Beds (Legacy)**
- **Pediatric ICU Beds**
- **Average Ventilator Usage**
- **Bed Utilization Rate**
- **Potential Increase in Bed Capacity**
- **Hospital Definition**

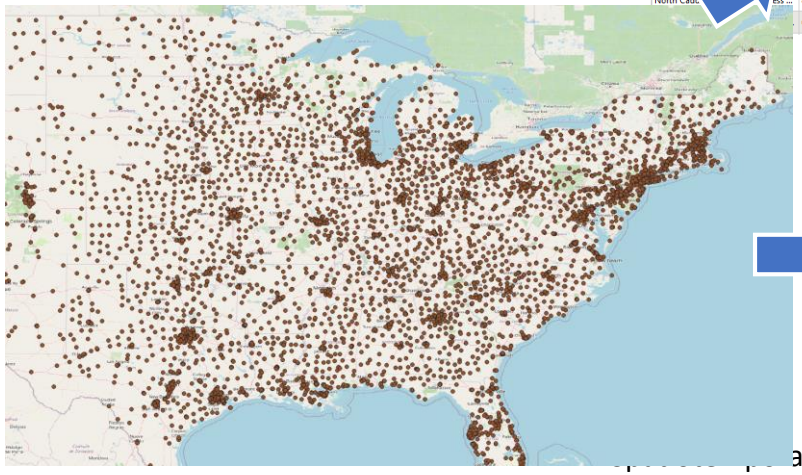
Hospital Type

- Short Term Acute Care Hospital (STAC)
- Critical Access Hospital (CAH)
- Religious Non-Medical Health Care Institutions
- Long Term Acute Care Hospitals
- Rehabilitation Hospitals
- Children's Hospitals
- Psychiatric Hospitals
- Veteran's Affairs (VA) Hospital
- Department of Defense (DoD) Hospital

Cons:

- Static Data





al Analysis of Medical Resource Deficiencies in the US under COVID-19 Pandemic (Sha, 2020)

Health Data

- **Data Source**: Centers for Medicare Medicaid Services - Critical care staff (CCS)
- **Download**: National Provider Identifier Registry (NPI) database
- **Format**: table (csv)
- **Category**: Health Data

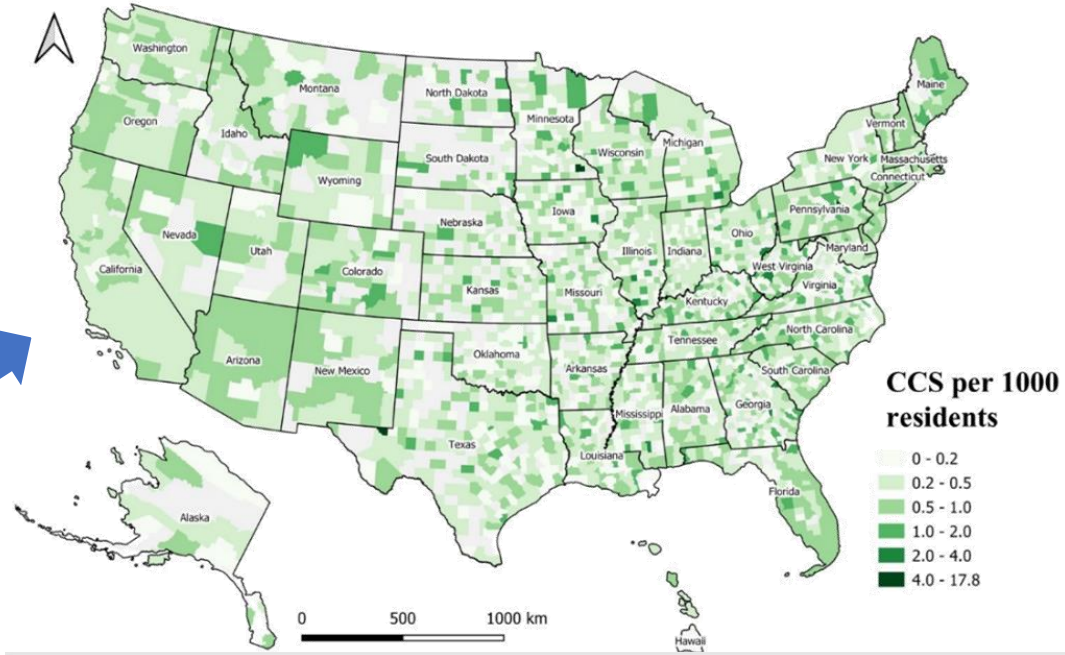
Index:

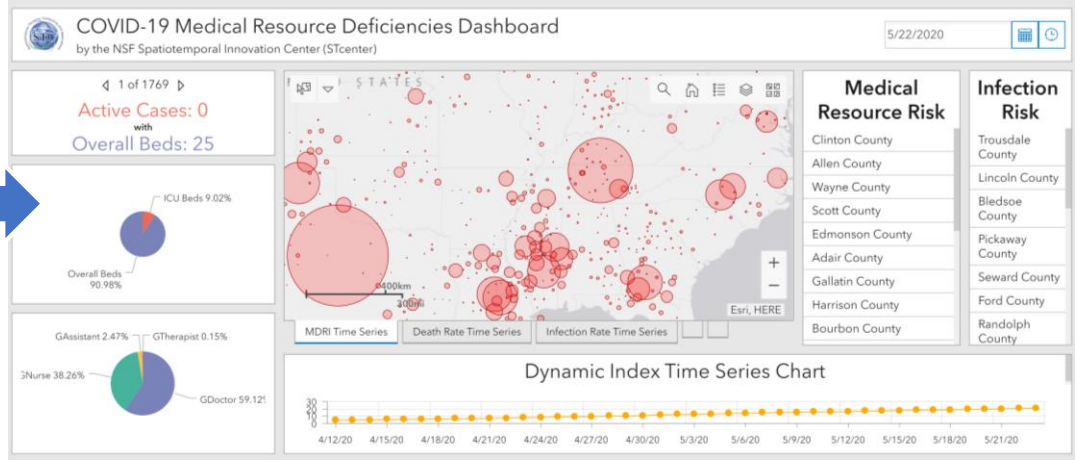
- NPI
- Enumeration Date
- NPI Type
- Status
- Authorized Official Information
- Mailing Address
- Primary Practice Address
- Health Information Exchange
- Other Identifiers
- Taxonomy Code

Cons:

- Static data
- Without geocoding

Taxonomy	Group
Emergency Medicine	Doctor
Anesthesiology: Critical Care Medicine	Doctor
Hospitalist	Doctor
Internal Medicine: Infectious Disease	Doctor
Internal Medicine: Critical Care Medicine	Doctor
Internal Medicine: Pulmonary Disease	Doctor
Surgical Critical Care	Doctor
Nurse Anesthetist, Certified registered	Nurse
Nurse Practitioner: Critical Care Medicine	Nurse
Respiratory Therapist, Certified: Critical Care	Therapist
Respiratory , Registered: Critical Care	Therapist
Anesthesiologist Assistant	Assistant





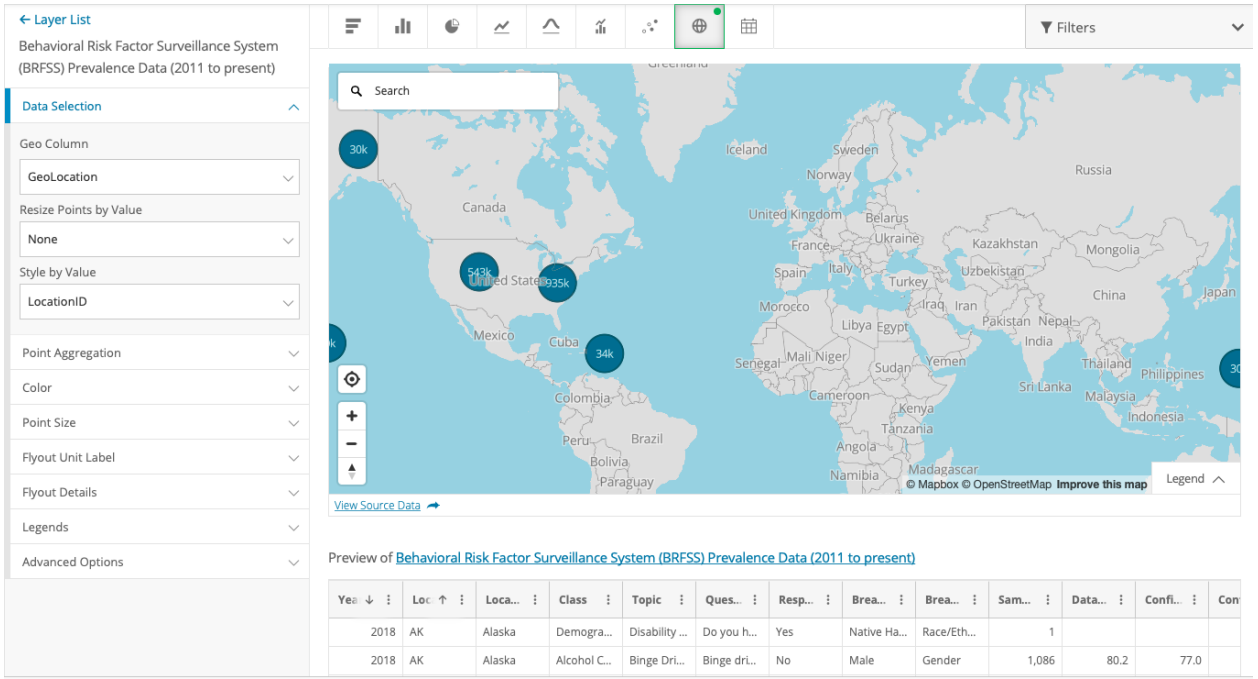
Spatiotemporal Analysis of Medical Resource Deficiencies in the US under COVID-19 Pandemic (Sha, 2020)

Health Data

- **Data Source**: Chronic Disease Risk Factors from CDC
- **Download**: https://chronicdata.cdc.gov/browse?category=Behavioral+Risk+Factors
- **Format**: table (csv)
- **Category**: health data

Index:

- **Alcohol consumption**
- **Cholesterol awareness**
- **Chronic health indicators**
- **Colorectal cancer screening**
- **E-cigarette use**
- **Demographics**
- **Fruits and vegetables**
- **Health care access/coverage**
- **Health status**
- **HIV-AIDS**
- **Hypertension awareness**
- **Immunization**
- **Injury**
- **Oral health**
- **Overweight and obesity (BMI)**
- **Physical activity**
- **Prostate cancer**
- **Tobacco use**
- **Women's health**



Cons:

- Sampling data

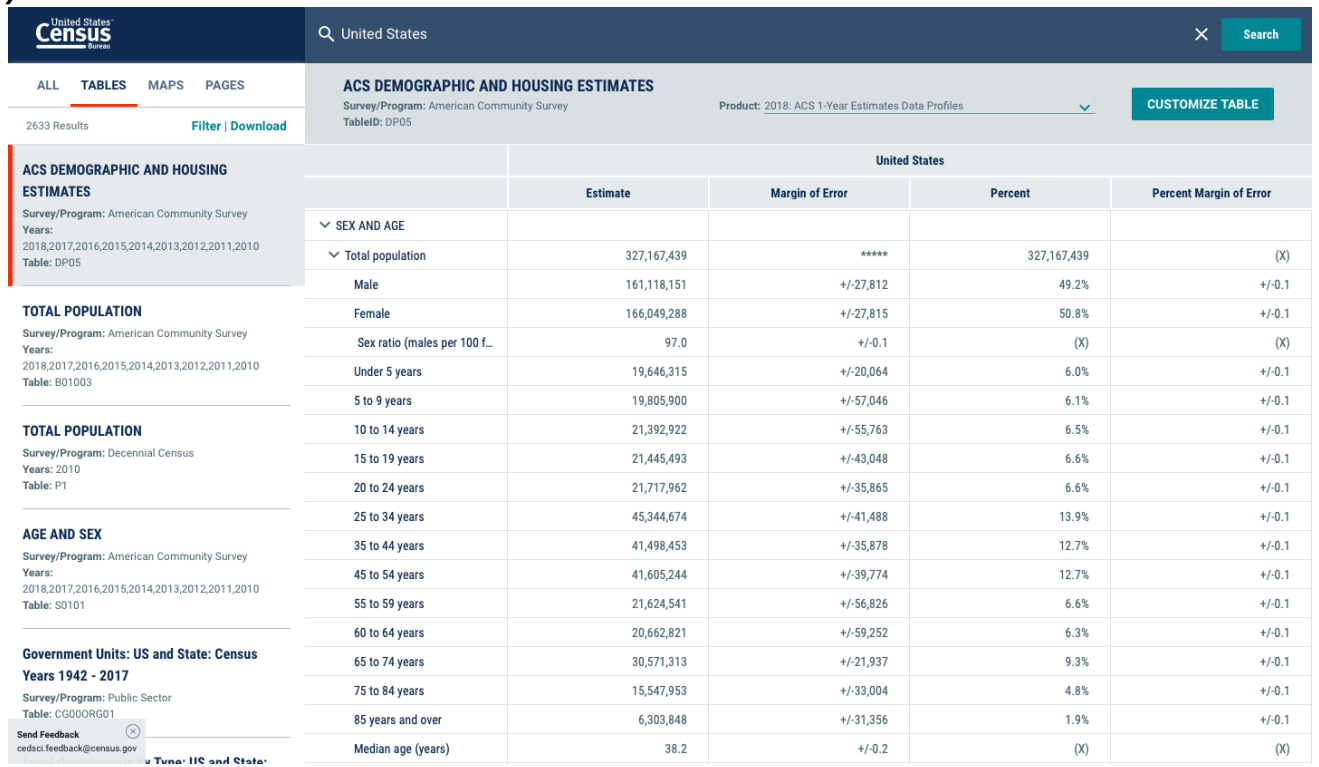
Provided by Regina Liu, Mercer University

Socioeconomic and Demographic Data

- **Data Source**: American Census Bureau (ACS)
- **Download**: https://data.census.gov/cedsci/
- **Format**: table (csv)
- **Category**: demographics

Index:

- **Age and sex**
- **Business and economy**
- **Education**
- **Emergency preparedness**
- **Employment**
- **Families and living arrangements**
- **Geographic mobility/migration**
- **Geography**
- **Health**
- **Hispanic origin**
- **Housing**
- **Income and poverty**
- **International trade**
- **Public sector**
- **Race**
- **Research**
- **Voting and Registration**
- **…**



Cons:

- The most recent available year is 2018

Provided by Regina Liu, Mercer University

Mobility Data

Mobile Phone Signaling Data

- **China**: Mobile，Telecom，Unicom
- **US:** Verizon Wireless, AT&T, T-Mobile, Sprint

pros： **accurate data**
cons： **private and confidential data**

Airline Flight Data

- **Opensky**: Global Flight Data
  https://opensky-network.org/
- **Variflight**: China Flight Data
  https://flightadsb.variflight.com/

pros： **mobilities inter/intro cities and countries**
cons： **without passenger numbers; sampling data**

Public Transit Data (China)

- Central city traffic data http://xxgk.mot.gov.cn
- Railway Data https://www.12306.cn

pros： **mobilities inter/intro-central cities**
cons： **hard to collect data; without passenger numbers**

Public Transit Data (US)

- Federal Transit Admin: Monthly Ridership Time Series
  https://www.transit.dot.gov/ntd/ntd-data
- Weekly Subway Check-in/out: NYC
  http://web.mta.info/developers/turnstile.html

pros： **mobilities in urban areas**
cons： **delayed data； local mobility data**

Mobility Data

Data Source: https://opensky-network.org/

Social Activity Data

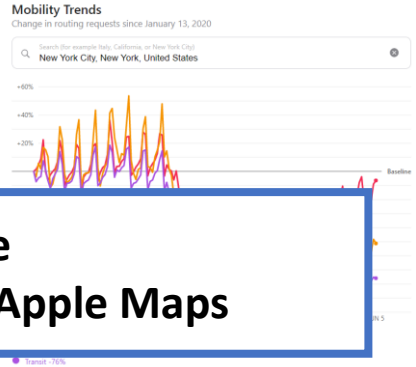
Google Community Mobility Report

- **Data:** https://www.google.com/covid19/mobility/
- **Format:** tabular/pdf
- **Category:** Mobility
- **Admin Level**: 0,1,2
- **Index**: POI access frequency

pros: global wide; one single file
cons: not comparable among countries

Apple Mobility Trends Report

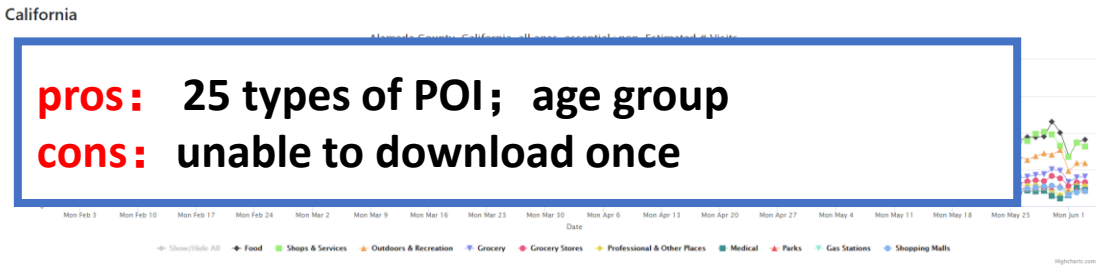
- **Data:** https://www.apple.com/covid19/mobility
- **Format:** tabular
- **Category:** Mobility
- **Admin Level**: 0,1,2, city
- **Index**: walking, biking, driving, and tı



pros: global wide; one single file
cons: requests for directions in Apple Maps

Foursquare Mobility Reports

- **Data:** https://visitdata.org
- **Format:** tabular
- **Category:** Mobility
- **Admin Level**: 1,2
- **Index**: POI access frequency



pros: 25 types of POI; age group
cons: unable to download once

SafeGraph COVID-19 Data Consortium

- **Data:** https://www.safegraph.com/covid-19-data-consortium
- **Format:** tabular
- **Category:** Mobility
- **Admin Level**: 1,2
- **Index**: POI access frequency

pros: varieties of data categories
cons: publicly accessible (application and Amazon S3)

Mobility Index

Baidu Mobility Index 迁徙指数

- **Data:** https://qianxi.baidu.com
- **Format:** json
- **Category:** Mobility Index
- **Admin Level:** province, city

Index
- In/out ratio
- in/out intensity
- Intro intensity

pros: inter/intro-city mobility index
cons: not publicly accessible after May 7th, 2020

Descartes Labs Social Distancing Index

- **Data:** https://github.com/descarteslabs/DL-COVID-19
- **Format:** tabular/json
- **Category:** Mobility Index
- **Admin Level**: state, count

Index
- m50
- m50_index

pros: accurate positioning data
cons: inter-city index not covered

Mobility Metrics and Social Distancing Index

- **Source: University of Maryland**
- **Data:** https://data.covid.umd.edu/methods/index.html
- **Format:** tabular/json
- **Category:** Mobility Index
- **Admin Level**: state, county

Index
- Social Distancing Index
- % staying home Trips/person
- % out-of-county trips
- % out-of-state trips Miles/person
- Work trips/person
- Non-work trips/person
- Transit mode share

pros: integrating multiple mobility sources
cons: inter-city index not covered

Camber Systems Social Distancing Reporter

- **Source: Facebook,** Camber, and Universities
- **Data:** https://covid19.cambersystems.com/
- **Format:** tabular/json
- **Category:** Mobility Index
- **Admin Level**: state, county

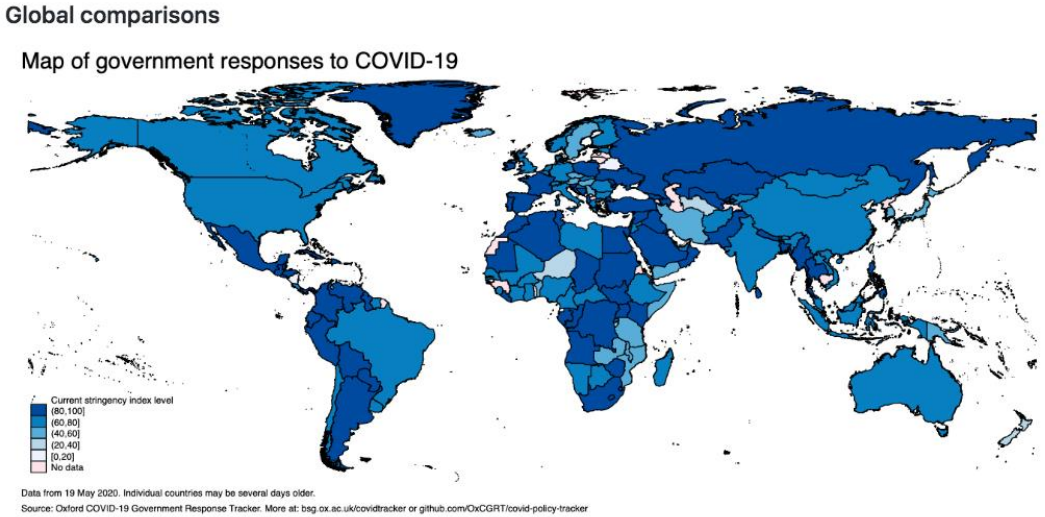
Index
- Radius of gyration
- Entropy

pros: accurate positioning data;
cons: application is needed for data accessibility

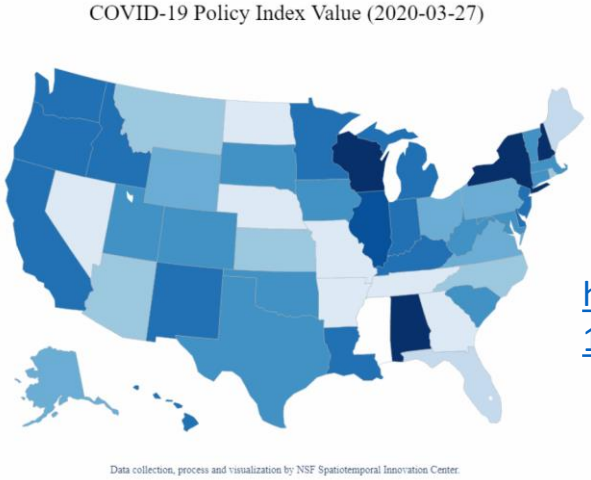
Policy Index

- **Source: Oxford COVID-19 Government Response tracker (OxCGRT)**
- **Data：** https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
- **Format：** tabular
- **Category：** Policy Data
- **Admin Level**: country

ID	Name	Type	Targeted/ General?
Containment and closure			
C1	School closing	Ordinal	Gegraphic
C2	Workplace closing	Ordinal	Gegraphic
C3	Cancel public events	Ordinal	Gegraphic
C4	Restrictions on gathering size	Ordinal	Gegraphic
C5	Close public transport	Ordinal	Gegraphic
C6	Stay at home requirements	Ordinal	Gegraphic
C7	Restrictions on internal movement	Ordinal	Gegraphic
C8	Restrictions on international travel	Ordinal	No
Economic response			
E1	income support	Ordinal	Sectoral
E2	debt/contract relief for households	Ordinal	No
E3	fiscal measures		No
E4	giving international support		No
Heatlh systems			
H1	Public information campaign	Ordinal	Gegraphic
H2	testing policy	Ordinal	No
H3	contact tracing	Ordinal	No
H4	emergency investment in healthcare		No
H5	investment in Covid-19 vaccines		No
Miscellanious			
M1	Other responses	Text	No



Cons： **Country-level policies only**



GMU STCenter

https://github.com/stccenter/COVID-19-Data/tree/master/Policy/US_Policy

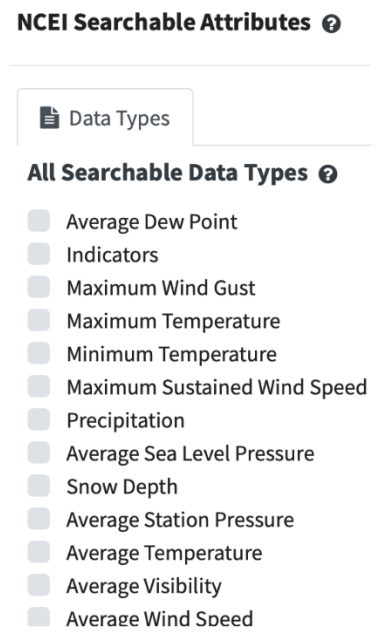
Environment Data

Meteorological Data

- **Data Source**: NOAA’ s Global Surface Summary of the Day
- **Access Point**: https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-day
- **Data Type**: **csv**
- **Region:** Global
- **Data Frequency:** Daily
- **Update Frequency:** usually 1-2 days after the date-time
- of the observations used in the daily summaries

Cons:

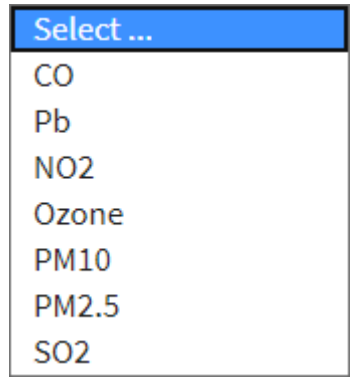
- Not every city has a station
- Not every station has data collected everyday
- Not every station has all types of data collected every day



Air Quality Data

- **Data Source**: United States Environmental Protection Agency
- **Access Point**: https://www.epa.gov/outdoor-air-quality-data/download-daily-data
- **Data Type**: **csv**
- **Region:** US
- **Data Frequency:** Daily
- **Update Frequency:** usually 1-2 days after the date-time
- of the observations used in the daily summaries

Index

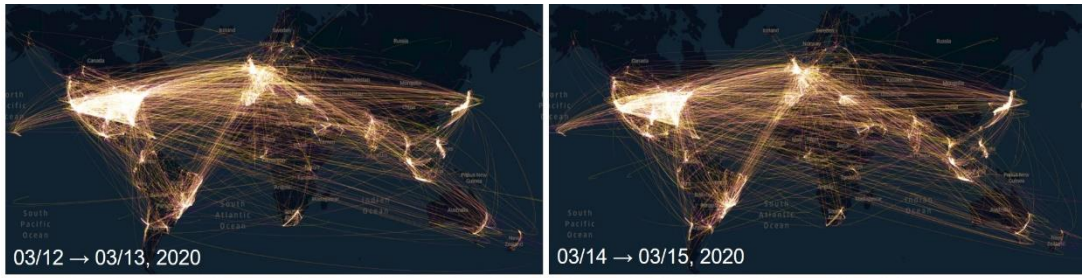


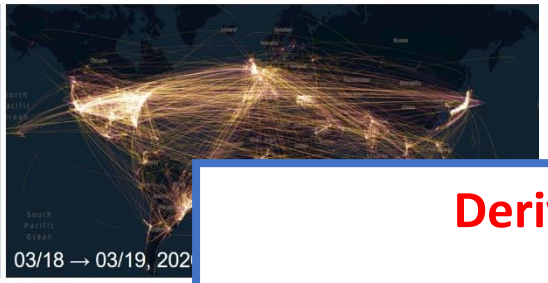
Data Source for Derivable Index

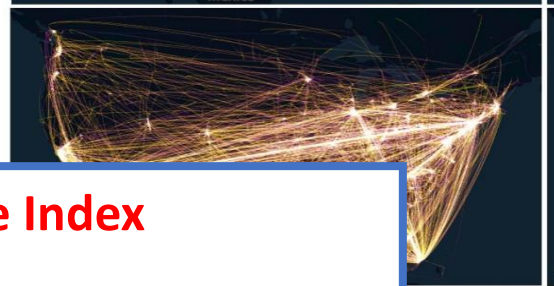
- Social Media
- Trends Data
- Global News
- Remote Sensing
- Night-light Image

Social Media

http://gis.cas.sc.edu/gibd/how-our-collective-efforts-of-fighting-the-virus-are-reflected-on-maps/







Derivable Index

- **Inter-City/Country Mobility Index**
- **Intra-City/Country Mobility Index**

Social Media

COVID-19-TweetIDs

- **Source:** University of Southern California
- **Data：** https://github.com/echen102/COVID-19-TweetIDs
- **Format：** txt
- **Category：** social media
- **Admin Level**: global
- **Index**: TweetID

Cons：
- Non-Geotagged Tweet
- No country information
- Gaps

Tracked since	Keyword
1/21/2020	Coronavirus; Corona; CDC; Ncov; Wuhan; Outbreak; China
1/22/2020	Koronavirus; Wuhancoronavirus; Wuhanlockdown; N95; Kungflu; Epidemic; Sinophobia
2/16/2020	Covid-19
3/2/2020	Corona virus
3/6/2020	Covid19; Sars-cov-2
3/8/2020	COVID–19
3/12/2020	COVD; Pandemic
3/13/2020	Coronapocalypse; CancelEverything; Coronials; SocialDistancing
3/14/2020	Panic buying; DuringMy14DayQuarantine; Panic shopping; InMyQuarantineSurvivalKit
3/16/2020	chinese virus; stayhomechallenge; DontBeASpreader; lockdown
3/18/2020	shelteringinplace; staysafestayhome; trumppandemic; flatten the curve
3/19/2020	PPEshortage; saferathome; stayathome
3/21/2020	GetMePPE
3/26/2020	covidiot
3/28/2020	epitwitter
3/31/2020	Pandemie

Geo-tagged Tweet

- **Source:** JNU, New Delhi
- **Data：** https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset
- **Format：** txt
- **Category：** social media
- **Admin Level**: global
- **Index**: TweetID

Cons：
- Start from March 20
- Gaps

- **Source:** CGA, Harvard University
- **Data：** https://doi.org/10.7910/DVN/2TOFZS/TIOZC8
- **Format：** txt
- **Category：** social media
- **Admin Level**: global
- **Index**: TweetID

Cons：
- '%corona%' or '%covid%'
- March and April

Trends Data

- **Source: Google**
- **Data:** https://trends.google.com/trends/?geo=US
- **Format:** csv
- **Category:** Media
- **Admin Level**: country and state

Pros:

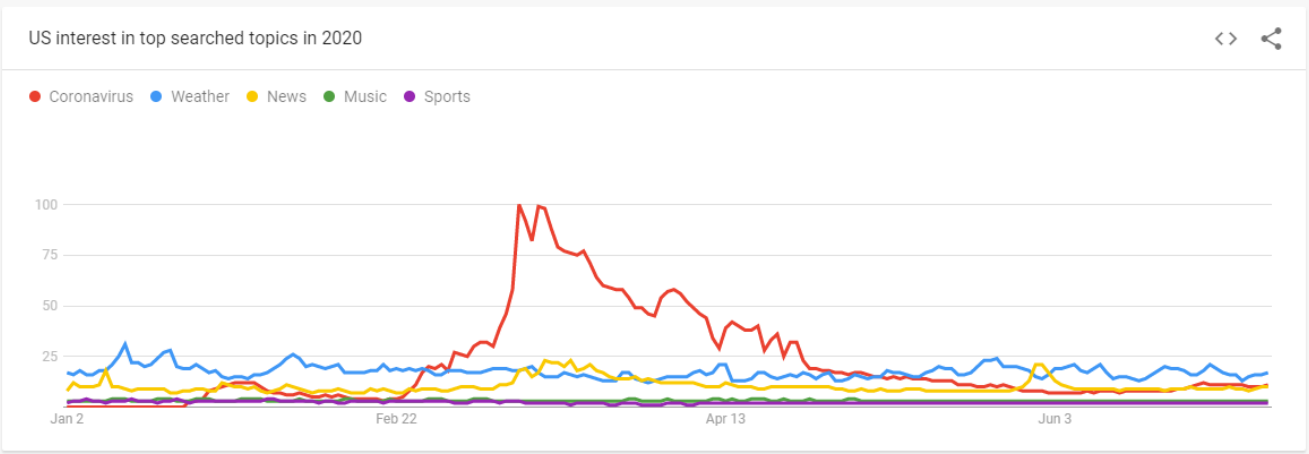
- Reflect public interest over time
- Closely associate with COVID-19 outbreak
- Verify the reliability

Cons:

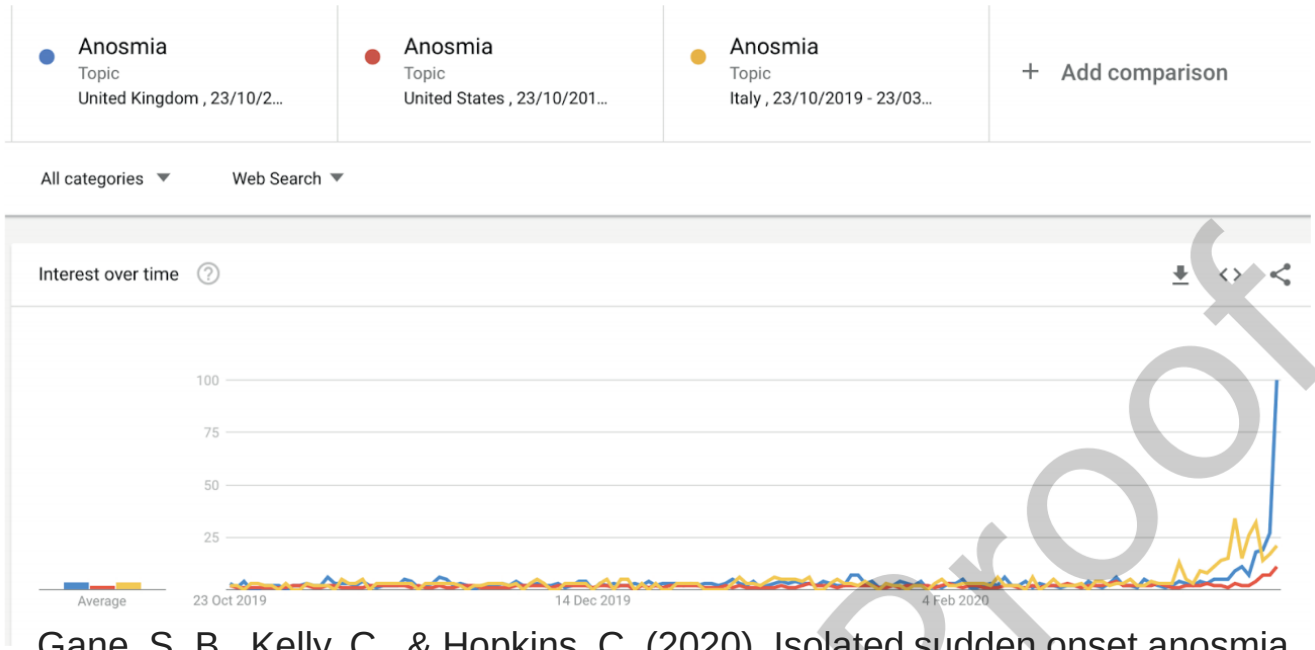
- Standardized data
- Do not necessarily represent users' opinions

Derivable Index

- **trends index**



https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en



Gane, S. B., Kelly, C., & Hopkins, C. (2020). Isolated sudden onset anosmia in COVID-19 infection. A novel syndrome. *Rhinology*, *10*.

Proof

Global News

- **Source: Google**
- **Data:** *https://www.gdeltproject.org*
- **Format:** json
- **Category:** Media
- **Admin Level**: global

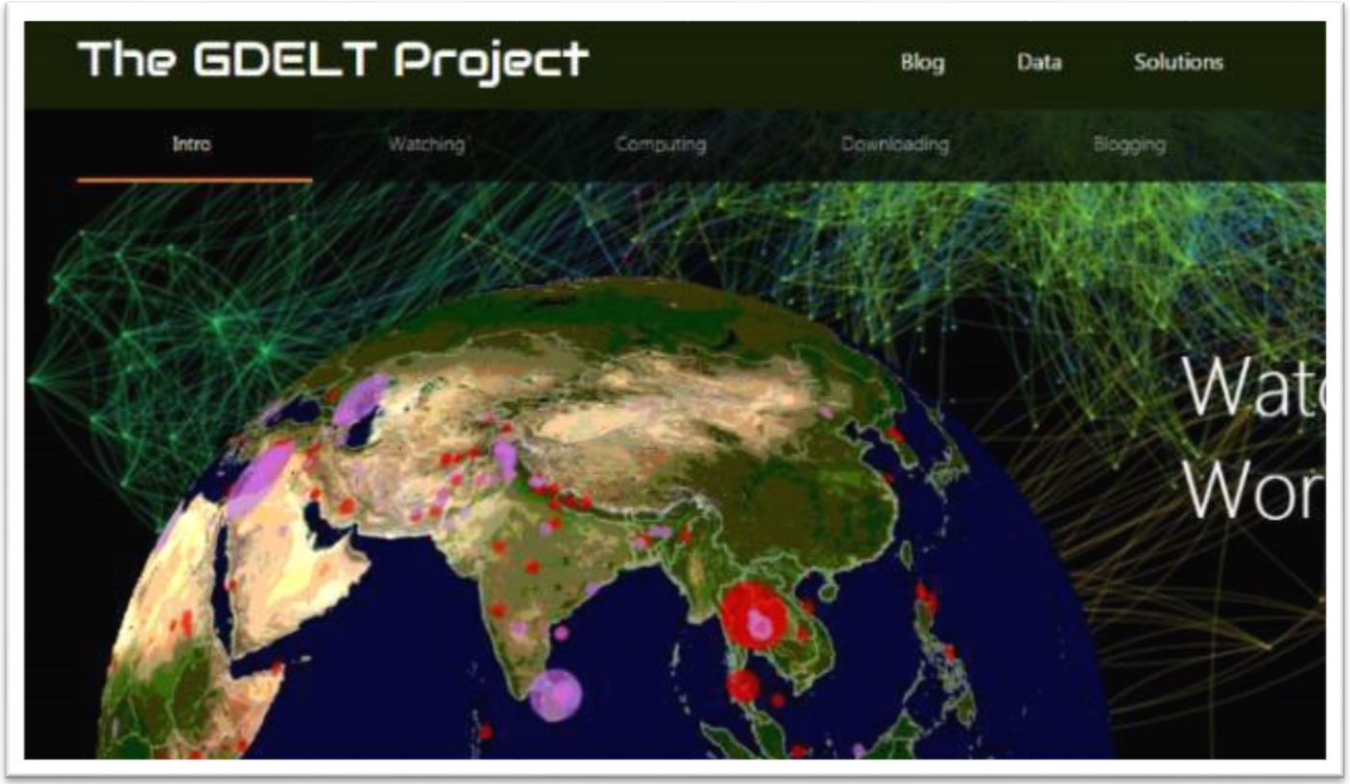
Index:

- Date
- Website
- Actor
- Event
- Tone
- Title
- Person
- Image
- Theme
- Country
- Geo
- …

Cons:

- News content is missing

Derivable Index
- Policy index
- Emotion index
- Association index



Source: https://www.gdeltproject.org

The GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over **100** languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world

Nighttime Light Data

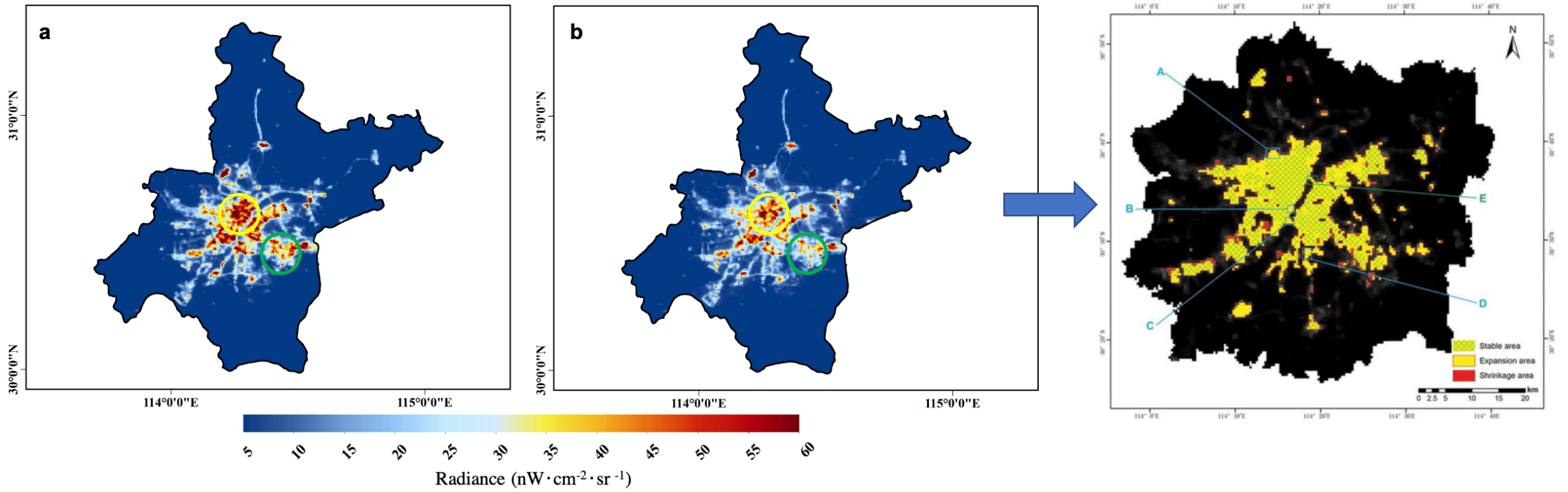
- **Source: VIIRS DNB**
- **Data:** https://ladsweb.modaps.eosdis.nasa.gov/search/order/1/VNP46A1--5000
- **Format: HDF5**
- **Category:** Remote Sensing
- **Admin Level**: global

Cons:

- Background noise: cloud cover

Derivable Index

- **Urban development changes**
- **Human activity changes**
- **Economic changes**



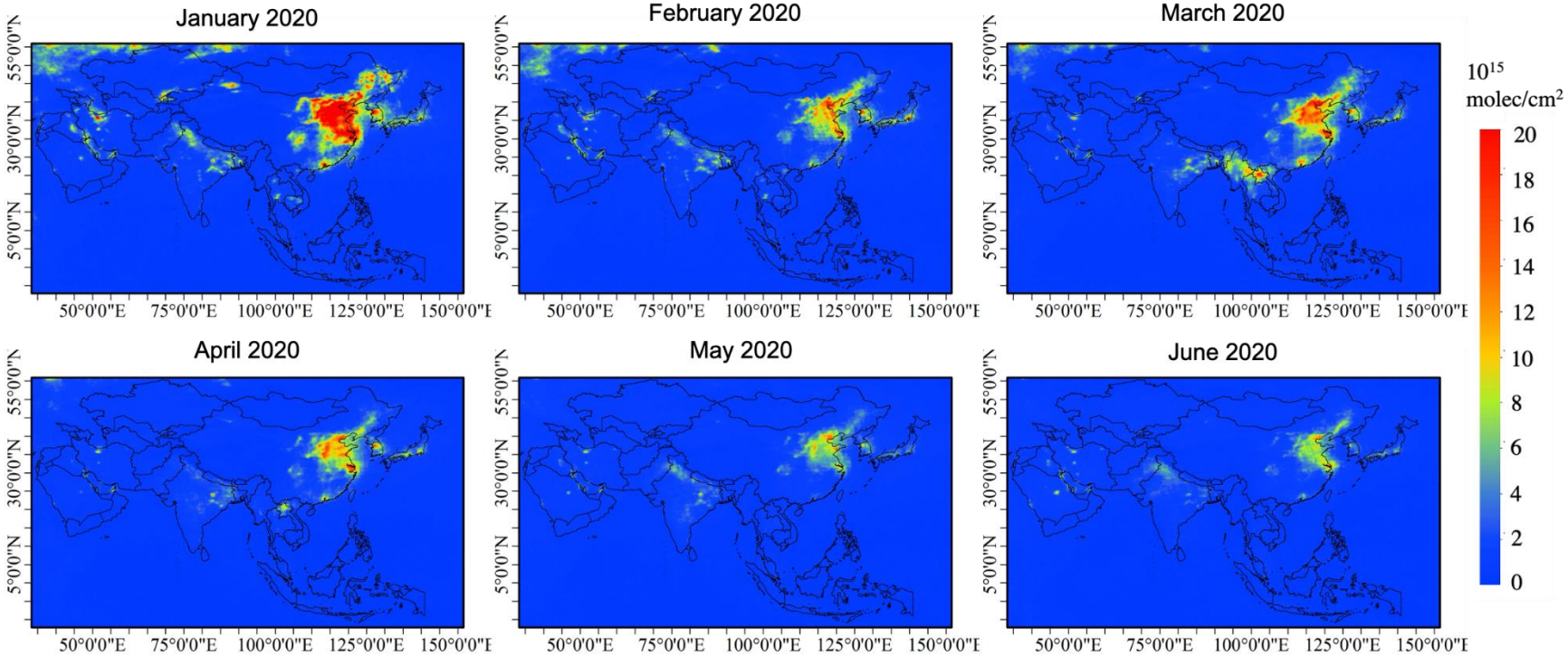
Liu, Q., et al. (2020). Spatiotemporal Patterns of COVID-19 Impact on Human Activities and Environment in Mainland China Using Nighttime Light and Air Quality Data. *Remote Sensing*, *12*(10)

Remote Sensing

- **Source:** Air Pollution Emission from **Aura OMI**
- **Data:** https://disc.gsfc.nasa.gov/datasets/OMTO3G_003/summary?keywords=omi
- **Format: HDF5**
- **Category:** Remote Sensing
- **Admin Level**: global

Cons:

- Non-ground station
- Low resolution (25km)



Derivable Index

- **Air Quality Index**
- **Land use category**
- **Ground impervious surfaces**

Liu, Q., et al. (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)

Literature Data

- **Source:** Allen Institute for AI (semantics scholar)
- **Data：** https://www.semanticscholar.org/cord19
- **Format：** json
- **Category：** Literature
- **Update Time**: Daily

Platform: PubMed, bioRxiv, medRxiv, arXiv.org, Microsoft Academic and the WHO COVID-19 database of publication

Source: WHO (World Health Organization)

Data： https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov

Format： json

Category： Literature

Update Time: Daily

Index：

- **Title**
- **Authors**
- **Source**
- **Database**
- **Type**
- **Language**
- **Publication Date**
- **Description**
- **Publication**
- **Fulltext link**
- **Abstract**
- **Volume**
- **Issue**
- **DOI**

Date	Changelog	Release file	md5	sha1
2020-07-07	changelog	cord-19_2020-07-07.tar.gz (3.5GB)	78b921fc	a2419b5d
2020-07-06	changelog	cord-19_2020-07-06.tar.gz (3.6GB)	ce134cfd	c27f9384
2020-07-05	changelog	cord-19_2020-07-05.tar.gz (3.5GB)	0d62cfad	f8d5607c
2020-07-04	changelog	cord-19_2020-07-04.tar.gz (3.5GB)	eb99050e	5a51a46f
2020-07-03	changelog	cord-19_2020-07-03.tar.gz (3.5GB)	b1e881b9	2034c097
2020-07-01	changelog	cord-19_2020-07-01.tar.gz (3.4GB)	0a33c4f2	e69b6934
2020-06-30	changelog	cord-19_2020-06-30.tar.gz (3.5GB)	e3869f74	4231b717
2020-06-29	changelog	cord-19_2020-06-29.tar.gz (3.4GB)	d73e1529	415d81e1
2020-06-28	changelog	cord-19_2020-06-28.tar.gz (3.4GB)	4d73b76a	1d964db9
2020-06-27	changelog	cord-19_2020-06-27.tar.gz (3.4GB)	4e21c52c	856bbb7e
2020-06-26	changelog	cord-19_2020-06-26.tar.gz (3.4GB)	a36a8e1b	6859bacd
2020-06-25	changelog	cord-19_2020-06-25.tar.gz (3.4GB)	ee83e780	d7e182c9
2020-06-24	changelog	cord-19_2020-06-24.tar.gz (3.3GB)	198d722f	c7697796
2020-06-23	changelog	cord-19_2020-06-23.tar.gz (3.4GB)	2e138620	4d427bb9
2020-06-22	changelog	cord-19_2020-06-22.tar.gz (3.3GB)	db7f0c3f	df755c6f
2020-06-21	changelog	cord-19_2020-06-21.tar.gz (3.3GB)	9a971c0e	d6fc3663
2020-06-20	changelog	cord-19_2020-06-20.tar.gz (3.3GB)	7aa91949	fac60eff
2020-06-19	changelog	cord-19_2020-06-19.tar.gz (3.3GB)	47b61215	fdd0490e
2020-06-18	changelog	cord-19_2020-06-18.tar.gz (3.3GB)	a922df2e	99c461a1
2020-06-17	changelog	cord-19_2020-06-17.tar.gz (3.3GB)	2ec0271f	19607fab
2020-06-16	changelog	cord-19_2020-06-16.tar.gz (3.2GB)	d85efa8b	cd48648d
2020-06-15	changelog	cord-19_2020-06-15.tar.gz (3.1GB)	1a359ff6	c59b59df

Derivable Index

- **subject**
- **Organization**
- **Index values**

https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html

Other Data Sources

Category	Provider	Name	Source
Mobility	Facebook, Camber, Harvard University, and so on.	Facebook Data for Good Mobility Dashboard	https://covid19.cambersystems.com/
Policy	International Monetary Fund	IMF COVID-19 Policy Tracker	https://www.imf.org/en/Topics/imf-and-covid19/Policy-Responses-to-COVID-19
Health	*Nuclear Threat Initiative (NTI), JHU and Economist Intelligence Unit (EIU)*	Global Health Security (GHS) Index	https://www.ghsindex.org/
Survey	Premise	COVID-19 Global Impact Study	https://www.premise.com/covid-19/
…	…	…	…

Data Sharing and Benchmark on CDL

- **Cases**
  - ✓ World COVID-19 Case Data (country)
  - ✓ US COVID-19 Cases Data (state/county/metro)
  - ✓ China COVID-19 Cases Data (city/province)
  - ✓ Canada, Austria, Italy, India … (GMU-STC)
- **Health Care Data**
  - ✓ Health facilities
  - ✓ Critical care staff
  - ✓ Chronic Disease Risk Factors
- **Mobility**
  - Transportation
    - Flight
    - Train
  - Mobility Index
    - Baidu mobility
    - Google mobility report
    - Apple mobility report
    - Foursquare
  - Social Distancing Index
    - Descartes Labs
    - University of Maryland
- **Environment**
  - ✓ Air Quality Data
  - ✓ Meteorological Data
- **Policy and regulations**
  - ✓ Chronicles
  - ✓ US policy index (GMU-STC)
  - ✓ China policy index (in progress)
- **Census**
  - ✓ Demographic
  - ✓ Economic

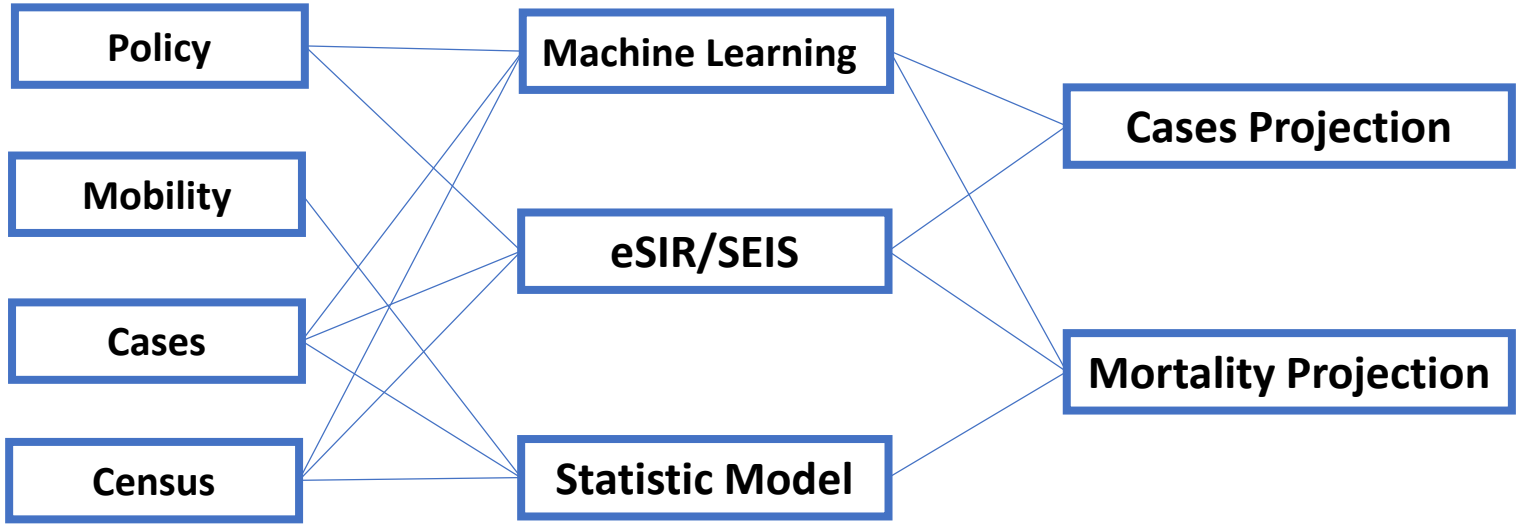
Data Sources for Derivable Index

- Media
  - ✓ Tweets and Geotagged Tweets
  - ✓ Global news
- Remote Sensing
  - ✓ Remote sensing images (GMU-STC)
  - ✓ Night-light images (GMU-STC)
- Publication

Model Replication and Extension on CDL

- ❑ Create benchmark datasets for COVID-19 modeling comparison studies
- ❑ Build COVID-19 models based on workflows for replicability, reproducibility and extension
- ❑ Integrate COVID-19 data, models, and tools
- ❑ Establish COVID-19 ecosystem for collaborative research, educating and training

ID	Group	Model Names	Methods	Replicable	Code	Data	Doc
1	Columbia University and University of North Carolina	Columbia-UNC	Statistical survival-convolutional model	Normal	YES	YES	YES
2	Qi-Jun Hong	QJHong	Machine learning using case data and mobility data	Hard	YES	YES	NO
3	Youyang Gu	YYG	SEIS mechanistic model	Hard	YES	NO	YES
4	Michigan University	eSIR	extended state-space SIR epidemiological model	Normal	Yes	YES	YES

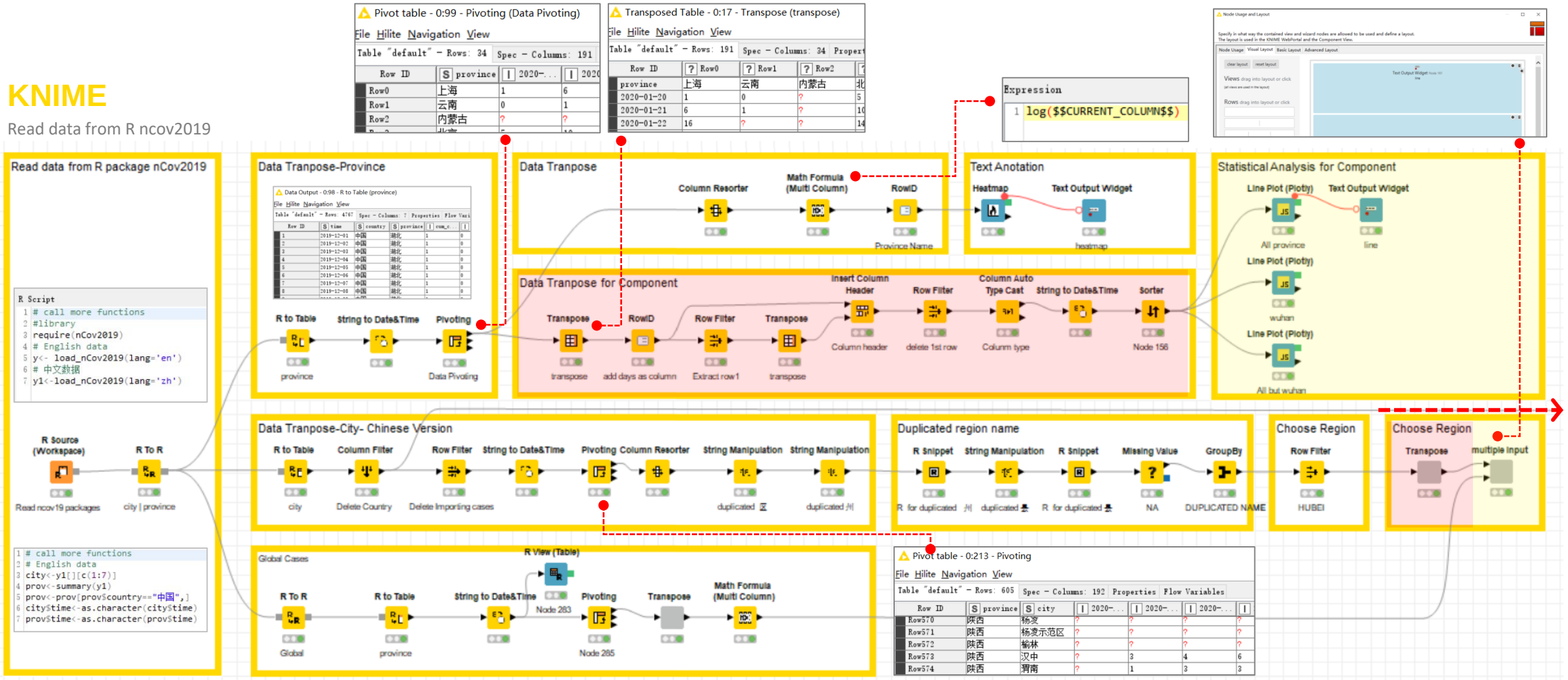


COVID-19 Model on CDL Cloud Platform

- University of Washington
- SEIR Model
- https://github.com/lilywang1988/eSIR

KNIME

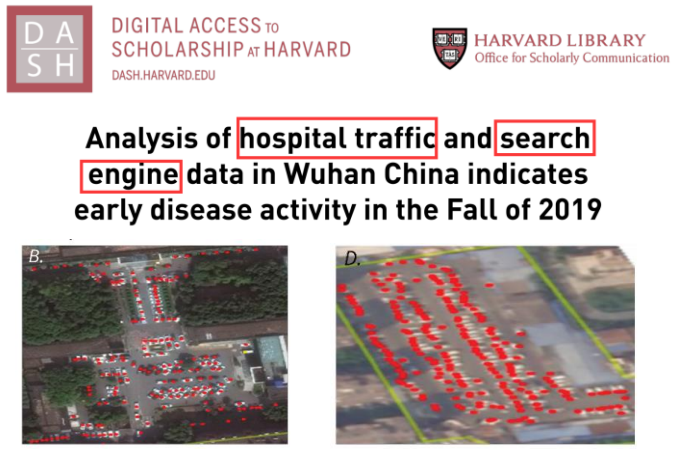
Read data from R ncov2019



Summaries and Discussions

1. How to use derivable data sources?

Unstructured big data, such as Social media, search engines, sensors, and crowd-sourcing data, provides many derivable indexes that adapt to many different needs, with wide coverage and strong timeliness.



2. How to promote data reliability?

- The confirmed cases and mortality number are underestimated
- Multi-sources data integration and cross validation

Data Quality of Chinese Surveillance of COVID-19: Objective Analysis Based on WHO's Situation Reports

Alvaro Javier Idrovo, MD, PhD[1] and Edgar Fabián Manrique-Hernández, MD[1]

3. How to break data limitations and fully understand system?

COVID-19 Panorama Puzzle: build a component-based research framework, easy to associate, integrate, replicate, reproduce and expand

"It is too early to be publishing papers on the effects of the COVID-19 pandemic in top academic journals like the JPE, particularly quantitative papers"—Greg Kaplan (Editor of the Journal of Political Economy)

https://www.shutterstock.com/video/clip-11746385-success-puzzle-white-background-videos-1-file

Data References

- COVID-19 Cases Data
  - https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/
  - https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html
  - https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
  - https://covidtracking.com/
  - https://usafacts.org/
  - https://github.com/nytimes/covid-19-data
- Health Data
  - https://coronavirus-resources.esri.com/datasets/definitivehc::definitive-healthcare-usa-hospital-beds
  - https://npiregistry.cms.hhs.gov/
  - https://chronicdata.cdc.gov/browse?category=Behavioral+Risk+Factors
  - https://data.census.gov/cedsci/
- Mobility Data
  - https://opensky-network.org/
  - http://xxgk.mot.gov.cn
  - https://www.transit.dot.gov/ntd/ntd-data
  - https://www.google.com/covid19/mobility/
  - https://www.apple.com/covid19/mobility
  - https://visitdata.org
  - https://www.safegraph.com/covid-19-data-consortium

Data References

- Mobility Metrics and Social Distancing Index
  - https://qianxi.baidu.com
  - https://github.com/descarteslabs/DL-COVID-19
  - https://covid19.cambersystems.com/
  - https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
  - https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-day
- Environment
  - https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-day
- Policy
  - https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
- Media
  - https://github.com/echen102/COVID-19-TweetIDs
  - https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset
  - https://doi.org/10.7910/DVN/2TOFZS/TIOZC8
  - https://www.gdeltproject.org
- Remote Sensing
  - https://ladsweb.modaps.eosdis.nasa.gov/search/order/1/VNP46A1--5000
  - https://disc.gsfc.nasa.gov/datasets/OMTO3G_003/summary?keywords=omi
- Literature
  - https://www.semanticscholar.org/cord19
  - https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov

References

Source: Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of The Total Environment, 140033.

Guan et al. (2020)	2/28	China	Confirmed cases of COVID-19	Geographical characteristics of those infected by COVID-19
Chen et al. (2020a) and Chen et al. (2020b)	2/28	China	Confirmed cases	Distribution of the contagion cases and their correlation with the emigration of the Wuhan population in the initial stage of the epidemic
Huang et al. (2020)	3/10	China	Confirmed cases	Spatiotemporal analysis of COVID-19 and its relationship with epidemiological characteristics, control of measures taken and their effects.
Arab-Mazar et al. (2020)	3/14	Iran	Confirmed cases	Spatiotemporal analysis of COVID-19 at the national and provincial levels
Giuliani et al. (2020)	3/20	Italy	Confirmed cases	Spatiotemporal analysis of COVID-19 at the national and provincial levels
Zhou et al. (2020)	3/20	China	Data mining and confirmed cases	Reflections on the use of GIS with big data and spatiotemporal analysis of COVID-19
Rezaei et al. (2020)	3/24	South Korea	Confirmed cases	Spatiotemporal analysis of COVID-19 at the national and provincial levels
Zhang et al. (2020)	3/26	China	Confirmed cases	Comparison of spatiotemporal evolution between COVID-19 and SARS 2003
Dagnino et al. (2020)	3/27	Brasil	Confirmed cases	Spatiotemporal analysis of COVID-19 in Rio Grande do Sul
Ahmadi et al. (2020a) and Ahmadi et al. (2020b)	3/31	Iran	Confirmed cases	Spatiotemporal analysis and call for the need to do more GIS studies locally
Xiong et al. (2020)	4/5	China	Confirmed cases	Pearson's correlation methods for spatiotemporal analysis
Gross et al. (2020)	4/5	China	Confirmed cases	Levy's flight to explain the spatiotemporal dynamics of the pandemic
Desjardins et al. (2020)	4/8	USA	Confirmed cases	Prospective space-time statistics to identify active and emerging COVID-19 groups at the county level
Rossman et al. (2020)	4/9	Israel	Polls and confirmed cases	Online questionnaire aimed at the geographical identification of possible symptomatics
De Ángel Solá et al. (2020)	4/10	Caribe basin	Confirmed cases	Predicting the global spread of COVID-19 based on geographic and climatic data
Team CC-R (2020)	4/10	USA	Confirmed cases	Geographical characteristics and spatiotemporal analysis of infections
Orea and Álvarez (2020)	4/13	Spain	Confirmed cases	Analysis by provinces of the effectiveness of quarantine on the spread of the pandemic
Murugesan et al. (2020)	4/14	India	Confirmed cases	Spatiotemporal analysis of COVID-19 at the national and provincial levels
Tang et al. (2020)	4/15	China	Confirmed cases	Poisson segmented model for the analysis of changing patterns in different geographic areas
Silva et al. (2020)	4/16	Brasil	Confirmed cases	Spatiotemporal analysis of COVID-19 in Estado da Bahia
Buzai (2020)	4/17	Argentina	Confirmed cases	Spatiotemporal analysis and reflections on health geography
Santana Juárez (2020)	4/17	Mexico	Confirmed cases	Spatiotemporal analysis of COVID-19 at the national and provincial levels
Saha et al. (2020)	4/26	World	Confirmed cases	Spatiotemporal analysis and reflections on usefulness of GIS in the pandemic

References

Source: Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of The Total Environment, 140033.

Luo et al. (2020)	2/17	China	Confirmed cases	Correlation between the number of contact cases and absolute humidity of the geographical location
Sajadi et al. (2020)	3/9	World	Confirmed cases	Analysis of the temperature, humidity and latitude of 8 cities with high infections and extrapolation to the rest of the world
Wang et al. (2020)	3/10	World	Confirmed cases	Analysis of temperature and humidity in Chinese cities with more than 40 infections and extrapolation to the rest of the world
Ma et al. (2020)	3/18	China	Confirmed cases	Correlation between deaths from COVID-19 in Wuhan and climate data and environmental pollution
Tosepu et al. (2020)	4/3	Indonesia	Confirmed cases	Correlation between climate and COVID-19 in Jakarta
Liu et al. (2020)	4/5	China	Confirmed cases	Correlation between climate and COVID-19 in 30 Chinese cities
Keshavarzi (2020)	4/6	World	Confirmed cases	Search for climatic and geographic evidence to explain the pandemic
Baker et al. (2020)	4/7	World	Confirmed cases	Climate-dependent epidemic model to simulate the COVID-19 pandemic in different scenarios
Hasan and Mahfujul Haque (2020)	4/7	World	Confirmed cases	Correlation between climate and COVID-19 in temperate and tropical countries
Bariotakis et al. (2020)	4/9	World	Confirmed cases	Predicting the global spread of COVID-19 based on climate data.
De Ángel Solá et al. (2020)	4/10	Caribe basin	Confirmed cases	Predicting the global spread of COVID-19 based on geographic and climatic data
Ahmadi et al. (2020a) and Ahmadi et al. (2020b)	4/13	Iran	Confirmed cases	Correlation between demographic and climatic characteristics and the spread of COVID-19
Qi et al. (2020)	4/16	China	Confirmed cases	Analysis of average daily temperature and relative humidity against cases of contagion
Bashir et al. (2020)	4/18	USA	Confirmed cases	Correlation between climatic characteristics and the spread of the virus in New York
Gupta et al. (2020)	4/19	USA, India	Confirmed cases	Correlation between climatic characteristics and the spread of the virus in the USA, and extrapolation of the method to India

References

Source: Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of The Total Environment, 140033.

De Kadt et al. (2020)	3/20	South Africa	Socio-economic characteristics and urban structure	For Gauteng city, 2 risk maps: to maintain social distance and preventive hygiene, and health and social vulnerability during the outbreak
Ahmadi et al. (2020a) and Ahmadi et al. (2020b)	3/31	Iran	Confirmed cases	Emphasises the need for more GIS studies at local scales
Gibson and Rush (2020)	4/6	South Africa	Socio-economic characteristics and urban structure	Analysis of the feasibility of social distancing in the slums of Cape Town.
Kuchler et al. (2020)	4/6	USA, Italy	Data mining and confirmed cases	Identification of Facebook friendships in Westchester (NY) and Lodi (Lombardia) and correlations with the pandemic
Allcott et al. (2020)	4/7	USA	Data mining and polls	Correlation between the ruling party in each county, social behaviour and confirmed COVID-19 cases
Lakhani (2020a) and Lakhani (2020b)	4/8	Australia	Socio-economic characteristics and urban structure	Analysis of the spatial distribution of health services and the population over 65 years of age in Melbourne
Padula and Davidson (2020)	4/9	World	Confirmed cases	Correlation between number of nurses per country and COVID-19 mortality
Lakhani (2020a) and Lakhani (2020b)	4/11	Australia	Socio-economic characteristics and urban structure	Characteristics of rural areas and their ageing population with respect to health services to face the pandemic
Minetto et al. (2020)	4/16	North Korea, Russia, Germany, USA, China	Remote sensing	Changes in population and economic dynamics through the comparison of satellite images prior to and during the pandemic
Kuupiel et al. (2020)	4/16	Ghana	Socio-economic characteristics and urban structure	Geographical accessibility analysis of Upper East Region inhabitants to point of care tests
Samuelsson et al. (2020)	4/18	World	Polls	Discussion on the characteristics of current cities and post-pandemic urban strategies
Coccia (2020)	4/20	Italy	Confirmed cases	Correlation between geoenvironmental and demographic characteristics of 55 cities and their relationship with the dynamics of COVID-19
Mollalo et al. (2020)	4/22	USA	Confirmed cases	Correlation between socioeconomic, environmental, topographic and demographic themes using a family of spatial regression and autoregressive models
Oto-Peralías (2020)	4/24	Spain	Confirmed cases	Correlation between geographic and socioeconomic variables to explain the great disparity of those infected between the provinces

Acknowledgement

- **Collaborators on COVID-19 Data Collection**
  - Harvard University, Cambridge, USA;
  - Wuhan University, Wuhan, China;
  - George Mason University, Fairfax, USA;
  - China Data Institute, Ann Arbor, US;
  - Ball State University, Muncie, US;
  - Shandong University of Science and Technology, Qingdao, China;
  - University of South Florida, Tampa, US;
  - University of Michigan, Ann Arbor, US;
  - East China Normal University, Shanghai, China;
  - Peking University, Beijing, China;
  - Fuzhou University, Fuzhou, China;
  - Hefei University of Technology, Hefei, China;
  - University of Washington, Seattle, US;
  - University of California, San Diego, USA
- **Harvard Dataverse Team**
- **Harvard CVT (Coronavirus Visualization Team)**

Websites and Contact

Resources for Coronavirus Study

http://chinadatalab.net

The Cloud for COVID-19 Study

Http://chinadatalab.org

spatialdatalab@lists.fas.harvard.edu