社会地理计算联合研究中心
Geocomputation Center for Social Science

Center for Geographic Analysis
Harvard University

China Data Institute

Descartes Lab Social Distancing Index In COVID-19 Research

Xiaokang Fu, Wuhan University, whistle_ant@whu.edu.cn
the Mobility Research Group of the ChinaDataLab
the Geo-computation Center for Social Sciences, Wuhan University
Chair: Mengxi Zhang, Ball State University

9PM-10PM, 04/30/2021（US EDT）

Outline

- ❑ Background
- ❑ Data Sources
- ❑ Data Depository
- ❑ Related research
- ❑ Data Workflows
- ❑ References

Background

The COVID-19 pandemic poses unprecedented challenges for countries around the world.
Many studies have shown that mobility data can provide significant supports for public health actions across the pandemic. Users'locations are usually collected from public transit, mobile operator, and mobile phone applications.

Background

In the COVID-19 studies, there are four common ways to estimate human Mobilities：

1） including public transit-based transportation flow (e.g., IATA)；

2） GPS-based social activities (e.g., Google Mobility Report) ；

3） mobility index (e.g., Baidu Mobility Index)；

4） social media derived mobility index (e.g., Geotagged Tweets).

Hu, Tao, et al. "Human Mobility Data in the COVID-19 Pandemic: Characteristics, Applications, and Challenges." *ResearchGate Preprint, DOI* 10 (2021).

Human Mobility Data in the COVID-19 Pandemic: Characteristics, Applications, and Challenges

Tao Hu[1], Siqin Wang[2,*], Bing She[3], Mengxi Zhang[4], Xiao Huang[5,*], Yunhe Cui[6], Jacob Khuri[7], Yaxin Hu[8], Xiaokang Fu[9], Xiaoyue Wang[9], Peixiao Wang[9], Xinyan Zhu[9], Weihe Wendy Guan[1], Shuming Bao[10], Zhenlong Li[11]

1. Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA. taohu@g.harvard.edu (T.H.); wguan@cga.harvard.edu (W.G.)
2. School of Earth and Environmental Sciences, The University of Queensland, St Lucia, Australia, 4067. s.wang6@uq.edu.au (S.W.);
3. Institute for Social Research, University of Michigan, MI, USA, coolnanjizhou@gmail.com (B.S.);
4. School of Health, Ball State University, Muncie, Indiana 47304, USA; mzhang2@bsu.edu (M.Z);
5. Department of Geosciences, University of Arkansas, Fayetteville, AR, 72701, USA. xh010@uark.edu (X.H.);
6. University of Connecticut, Department of Geography, Storrs, CT 06269, USA; yunhe.cui@uconn.edu (Y.C.);
7. Independent Public Health Researcher, California, USA; jacobkhuri@outlook.com (J.K.);
8. School of Resource and Environmental Sciences, Wuhan University, Wuhan, Hubei 430079, China, yaya_cindy@whu.edu.cn (Y. H.);
9. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei 430079, China, whistle_ant@whu.edu.cn (X.K.); 2017302590218@whu.edu.cn (X. W.); xinyanzhu@whu.edu.cn (X.Z.);
10. China Data Institute, Ann Arbor, MI, USA; sbao@umich.edu (S.B.)
11. Geoinformation and Big Data Research Laboratory, Department of Geography, University of South Carolina, SC, USA; zhenlong@sc.edu (Z.L.)

Abstract: The COVID-19 pandemic poses unprecedented challenges around the world. Many studies indicate that human mobility data provide significant support for public health actions during the pandemic. Researchers have applied mobility data to explore spatiotemporal trends over time, investigate associations with other variables, and predict or simulate the spread of COVID-19. Our objective was to provide a comprehensive overview of mobility data to guide researchers and policymakers in conducting data-driven evaluations and decision-making for the COVID-19 pandemic and other infectious disease outbreaks. We summarized the mobility data usage in COVID-19 studies by reviewing recent publications on COVID-19 and human mobility from a data-oriented perspective. We identified three major sources of mobility data:

Table 1. Summary of human mobility datasets in recent COVID-19 studies

Data Category	Name and Provider	Region and scale	Available Time	OD Flow	Availability	Strengths	Weaknesses	Selected References
Index-based Mobility Data	Cuebiq Mobility Index	U.S. at multiple geographic levels	01/01/2020	No	Submit Application	Available in DMA level; index allows counties to be compared to one another	Only available in U.S.	Fraiberger et al., 2020 Pepe et al., 2020
	Baidu Mobility Index	China/city and province	1/1/2020 ~ 5/7/2020 & 9/3/2020 ~ presen	Yes	Public	inter/intro-city mobility index	Not publicly accessible after May 7 2020, only available for Mainland China	Ze-Liang et al., 2020 Liu et al., 2020 Xu et al., 2020
	Descartes Lab Mobility Index	U.S./county and state	03/01/2020 – 06/06/2020	No	Submit Application	accurate positioning data (m50 score based on normalization methods)	Inter-city index not covered; only freely available in a certain period of time and scale	Warren et al., 2020 Gao et al., 2020; Chen et al., 2020
	Unacast Social Distancing Index	U.S.	02/24/2020 ~ present	No	Submit Application	Granular data, available down to specific data points; bias correction based on classifications of businesses	Since data is coming from third party sources, people may have to agree to consent on those sources	Brodeur et al., 2021
	University of Maryland Mobility Metrics and Social Distancing Index	U.S./county and state	01/01/2020 ~ present	No	Submit Application	Integrated and cleaned location data from multiple sources; be highly representative	Only available in the U.S.	Zhang et al., 2020 Lee et al., 2020 Ghader, et al., 2020
	Camber Systems Social Distancing Reporter	U.S./county	08/01/2020 ~ present	No	Submit Application	Integrating multiple data sources; less biased and more representative; easy to interpret	subject to calibration; only available in U.S. county level; no data before August 2020	Jeffrey et al., 2020

Data Sources

- Descartes Lab Social Distancing Index data
  - https://github.com/descarteslabs/DL-COVID-19
  - https://www.descarteslabs.com/mobility/#overview

Github data repository

a6df4cc 13 hours ago 402 commit

README		13 months ag
DL-us-m50.csv	Data through 2021-04-19	13 hours ago
DL-us-m50_index.csv	Data through 2021-04-19	13 hours ago
DL-us-mobility-daterow.csv	Data through 2021-04-19	13 hours ago
DL-us-mobility.ndjson	Data through 2021-04-19	13 hours ago
DL-us-samples.csv	Data through 2021-04-19	13 hours ago
LICENSE	Add CC BY 4.0 License	13 months ago
README.md	Data through 2021-04-14	6 days ago

README.md

Data for Mobility Changes in Response to COVID-19

[U.S. mobility Data (ndjson) | U.S. mobility Data (csv) | U.S. m50_index Data (alternate csv)]

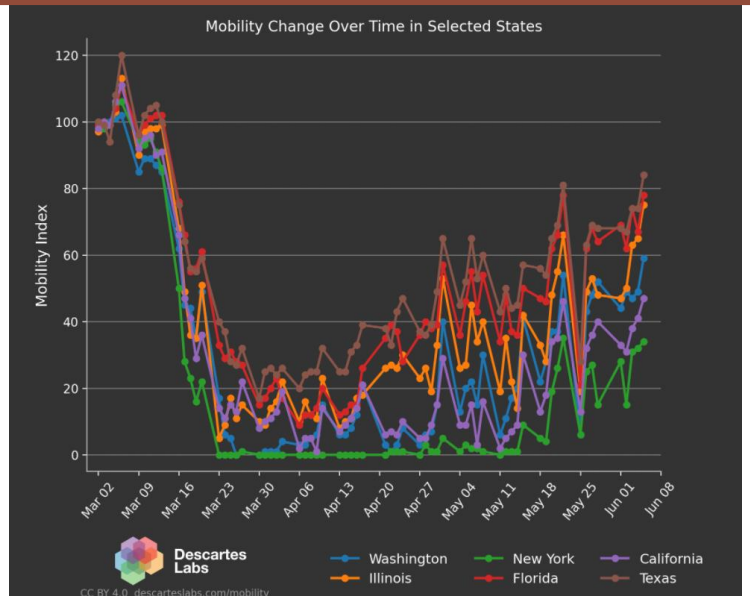
Descartes Labs is releasing mobility statistics (representing the distance a typical member of a given population moves in a day) at the US admin1 (state) and admin2 (county) level. A technical report describing the motivation behind this work with methodology and definitions is available at arxiv.org/pdf/2003.14228.pdf. We intend to update the data in this repository regularly.

Note: Data for 2020-04-20, 2020-05-29, 2020-10-08, 2020-12-11 through 2020-12-18, 2021-01-08 through 2021-01-14, 2021-04-07 and 2021-04-12 did not meet quality control standards, and was not released.

Mobility Data

NDJSON format data can be found in the DL-us-mobility.ndjson file.

Mobility Change Over Time in Selected States



Data Sources

- **Descartes Lab Social Distancing Index Data：**
  - **Region and scale：** US at state, county
  - **Update frequency：** Daily
  - **Variables:** The distance a typical member of a given population moves in a day.
  - **Availability:** Open data

https://arxiv.org/pdf/2003.14228.pdf

Mobility Changes in Response to COVID-19

Michael S. Warren, Samuel W. Skillman
Descartes Labs
100 North Guadalupe Street
Santa Fe, NM 87501, USA

v1.0-4-ga6898c6

arXiv:2003.14228v1 [cs.SI] 31 Mar 2020

***Abstract*—In response to the COVID-19 pandemic, both voluntary changes in behavior and administrative restrictions on human interactions have occurred. These actions are intended to reduce the transmission rate of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). We use anonymized and/or de-identified mobile device locations to measure mobility, a statistic representing the distance a typical member of a given population moves in a day. Results indicate that a large reduction in mobility has taken place, both in the United States and globally. In the US, large mobility reductions have been detected associated with the onset of the COVID-19 threat and specific government directives. Mobility data at the US admin1 (state) and admin2 (county) level have been made freely available under a Creative Commons Attribution (CC BY 4.0) license via the GitHub repository** github.com/descarteslabs/DL-COVID-19.

I. Introduction

As of March 28, 2020, the coronavirus pandemic COVID-19 is affecting 200 countries and over 600,000 individuals [1, 2]. Efforts to prevent spread of the virus include travel and work restrictions, quarantines, curfews, cancellations and postponements of events, and facility closures. These policies aim to reduce the probability of contact between infected and non-infected persons to minimize disease transmission [3]. We hypothesize that the effect of these interventions are measurable in the aggregate geospatial statistics of people's daily movement. Measuring changes in mobility is a critical input for forecasting disease spread and assessing the effectiveness of containment strategies [4].

Geolocation reports from smartphones and other mobile devices offer a mechanism to sample the movement of individuals. There are significant privacy concerns related to the availability of such data [5]. Standard practice is to make such traces anonymous, where the true identities of nodes are replaced by random identifiers. However, the privacy concern remains. Nodes are open to observations in public spaces, and they may voluntarily or inadvertently disclose partial knowledge of their whereabouts [6]. Some studies have shown that sharing anonymized location data can lead to privacy risks and that, at a minimum, the data needs to be coarse in either the time domain or the space domain [7]. We do not and will not use mobile device location data to identify individuals. All analysis we perform is statistically aggregated, removing the ability to characterize the behavior of any single device.

II. Method

Cloud computing with associated high-bandwidth storage capacity, combined with recent advances in machine learning, is enabling understanding of the world at a scale and level of granularity never before possible [8]. Mobile devices exemplified by the smartphone are an obvious result of this technological advance. In this work, we analyze a commercially available mobile device location dataset using cloud computing resources. We heavily leverage our festivus virtual file system layer [9] for this analysis, reading over 50 TB of input data and utilizing about 50,000 CPU hours of computation.

Raw location data from commercial data providers is generally provided in a compressed, comma-delimited text format, with one position report per line. Reports contain, at a minimum, an anonymized node id, epoch time (seconds elapsed since 00:00:00 UTC on January 1, 1970), latitude, longitude and an estimate of position accuracy. Data is obtained in multiple files to facilitate parallel processing, and a typical volume of compressed data is 100 GB per day. Data is delivered once per day.

Our analysis algorithm is implemented with the Python programming language [10, 11] in a parallel task-based execution environment. The first step in analysis is to eliminate reports with estimated position accuracy that exceeds some threshold. We use a threshold of 50 meters. Median reported accuracy across all position reports is in the range of 15-20 meters.

In order to make daily statistics more interpretable, the reported timestamp is converted to local time. If a time conversion is not performed, then the statistics in each area would be analyzed in UTC, and would not correspond to a "day" in the local timezone. For instance, a day in UTC from 00:00 to 23:59 would correctly correspond to a local day (midnight to midnight) in the UK, but would be noon to noon in New Zealand. This would lead to analysis artifacts (e.g. a day that combines Sunday afternoon with Monday morning in New Zealand cannot be consistently compared to midnight Sunday through midnight Monday in the UK). Our initial conversion to local time uses an approximate local solar time with a time zone offset of,

$$tzhours = \left\lfloor \frac{longitude}{15} \right\rceil$$

Data Sources

Index

- m50: The median of the max-distance mobility for all samples in the specified region.
- **m50_index**: The percent of normal m50 in the region, with normal m50 defined during 2020-02-17 to 2020-03-07

$$m50_index = 100\frac{m50}{m50_{norm}},$$

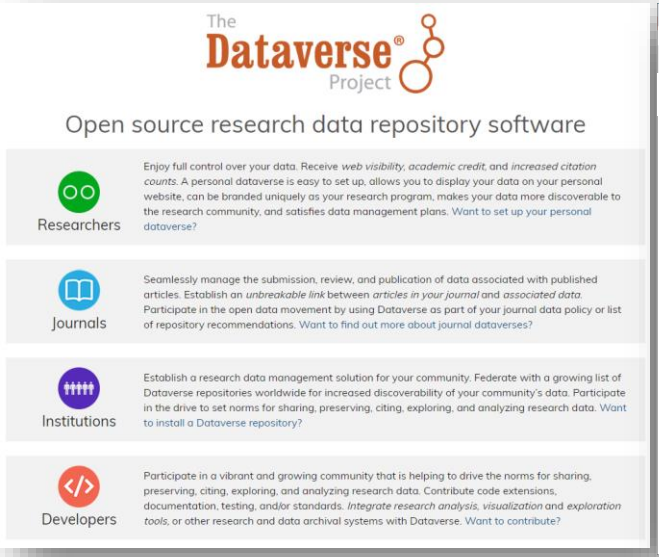
Pros

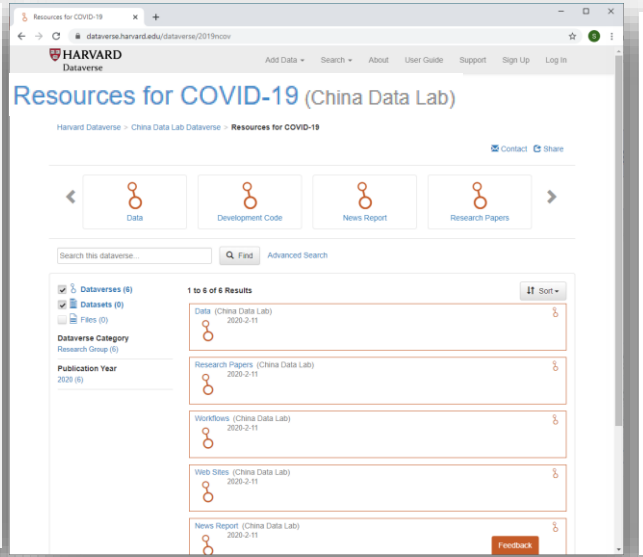
- accurate positioning data (m50 score based on normalization methods)

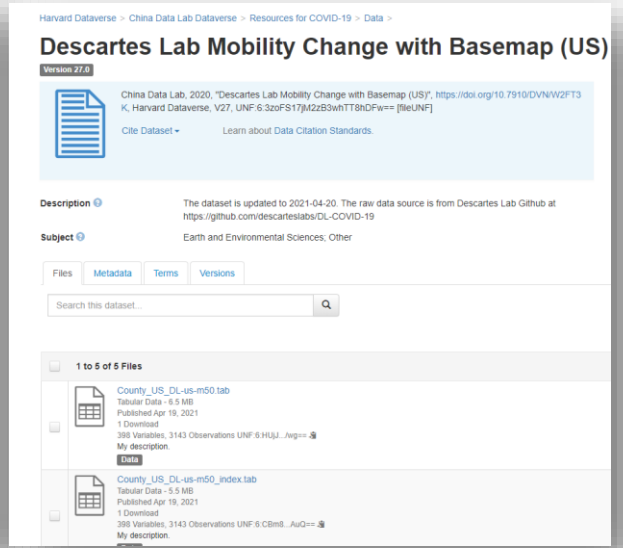
Cons

- Inter-city index not covered;
- only freely available in a certain period of time and scale

Data Depository: dataverse.harvard.edu







Related research

Clustering Patterns Connecting COVID-19 Dynamics and Human Mobility Using Optimal Transport

Nielsen, F., Marti, G., Ray, S., & Pyne, S. (2021). Clustering Patterns Connecting COVID-19 Dynamics and Human Mobility Using Optimal Transport. *Sankhya B*. https://doi.org/10/gjrfs5

Sankhyā B : The Indian Journal of Statistics
https://doi.org/10.1007/s13571-021-00255-0
© 2021, Indian Statistical Institute

Clustering Patterns Connecting COVID-19 Dynamics and Human Mobility Using Optimal Transport

Frank Nielsen
Sony Computer Science Laboratories Inc, Tokyo, Japan
Gautier Marti
Independent Researcher, Abu Dhabi, United Arab Emirates
Sumanta Ray
Centrum Wiskunde, Informatica, Amsterdam, The Netherlands
Saumyadipta Pyne
University of Pittsburgh, Pittsburgh, USA

Abstract

Social distancing and stay-at-home are among the few measures that are known to be effective in checking the spread of a pandemic such as COVID-19 in a given population. The patterns of dependency between such measures and their effects on disease incidence may vary dynamically and across different populations. We described a new computational framework to measure and compare the temporal relationships between human mobility and new cases of COVID-19 across more than 150 cities of the United States with relatively high incidence of the disease. We used a novel application of Optimal Transport for computing the distance between the normalized patterns induced by bivariate time series for each pair of cities. Thus, we identified 10 clusters of cities with similar temporal dependencies, and computed the Wasserstein barycenter to describe the overall dynamic pattern for each cluster. Finally, we used city-specific socioeconomic covariates to analyze the composition of each cluster.

AMS (2000) *subject classification.* Primary; 37Mxx Secondary; 37M10.
Keywords and phrases. Clustering, Optimal transport, Wasserstein distance, Time series, Mobility, COVID-19.

1 Introduction

The year 2020 marks the centenary of birth of Professor Calyampudi Radhakrishna Rao, on which we congratulate this living legend in the field of statistics, and wish him a longer, healthy life. The same year will also be remembered for the occurrence and phenomenal spread of the COVID-19 pandemic that has profoundly impacted all aspects of human life globally.

Related research

Overview

- Social distancing and stay-at-home are among the few measures that are known to be effective in checking the spread of a pandemic such as COVID-19 in a given population.
- **The patterns of dependency** between such measures and their effects on disease incidence **may vary dynamically and across different populations**.
- Described a **new computational framework** to **measure and compare the temporal relationships** between human mobility and new cases of COVID-19
  - Used a **novel application of Optimal Transport for computing the distance** between the normalized patterns induced by bivariate time series for each pair of cities.
  - **Identified 10 clusters** of cities with similar temporal dependencies, and computed the Wasserstein barycenter to describe the overall dynamic pattern for each cluster.
  - Used city-specific socioeconomic covariates to analyze **the composition of each cluster**.

Related research

- Data:
- 1. COVID-19 incidence and population data
  - from the Johns Hopkins Coronavirus Resource Center(https://coronavirus.jhu.edu/)
  - from the online resources of the U.S. Census Bureau and the U.S. Centers for Disease Control and Prevention (CDC) (https://www.census.gov/quickfacts/, https://svi.cdc.gov/)
- 2. Human mobility index data
  - Descartes Labs https://github.com/descarteslabs/DL-COVID-19

Related research

The overall workflow

$$\Delta M(t) = M(t) - M(t-1) \tag{2.1}$$

$$M'(t) = \{(M(t) - M(t-1)) + 0.5 * (M(t+1) - M(t-1))\}/2. \tag{2.2}$$

Algorithm The workflow of the analytical framework.

Input: For each of $k(= 151)$ given cities, a bivariate time series: mobility (M) and new cases (N) for each date (t) over a fixed time-interval (March 1 – May 31, 2020).

Steps of the Analysis: .

1: As measures of mobility, along with M, also consider its variants ΔM and M' computed with Eqs. 1 and 2.

2: Performed normalized ranking of variables ($M/\Delta M/M'$, N and t) to represent each city as a discrete set of ranked points in unit cube ($[0, 1]^3$)

3: Compute optimal transport (OT) distance between the point sets representing each pair of cities.

4: Cluster the cities based on the OT distance matrix. Three different hierarchical clusterings $HC1$, $HC2$ and $HC3$ were obtained based on Ward's linkage method and 3 variants of mobility: M, ΔM, and M' respectively.

5: Apply HCMapper to compare the dendrograms of different clusterings ($HC1$, $HC2$ and $HC3$). Select the clustering ($HC3$) that yields the most spatially homogeneous clusters.

6: Compute Wasserstein barycenter for each cluster of the selected clustering ($HC3$).

7: Analyze the composition of the clusters by applying random forest classifier on 15 city-specific covariates as feature set. Identify the contributions of the covariates to discriminate among the clusters.

Related research

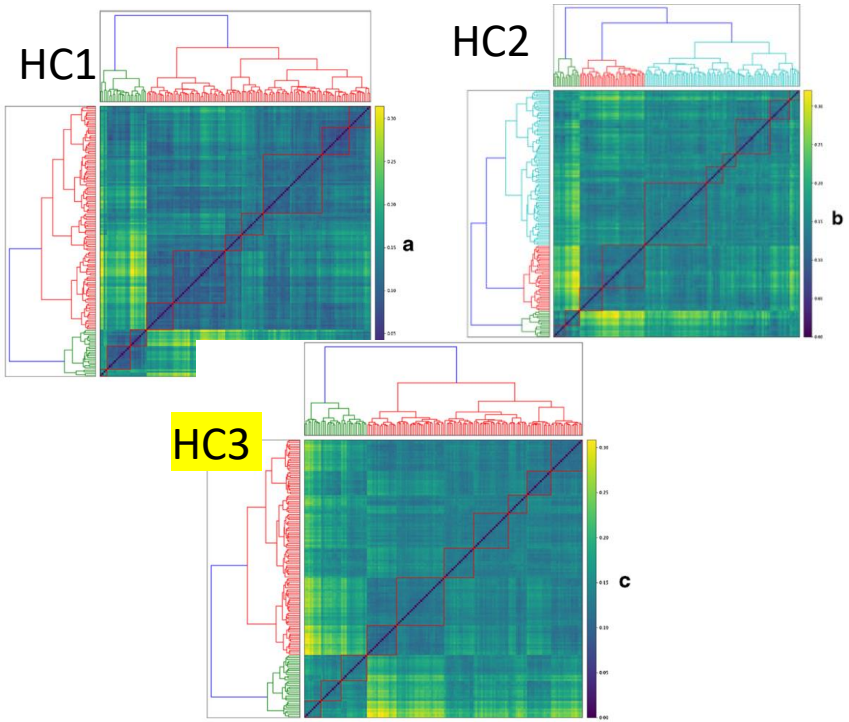


Figure 1: The dendrograms show 3 hierarchical clusterings of cities **a** $HC1$ (N, M, t), **b** $HC2$ $(N, \Delta M, t)$, and **c** $HC3$ (N, M', t) using Ward's linkage. Based on visual inspection of the seriated distance matrix, 10 clusters were identified in each case, as shown on the heatmaps

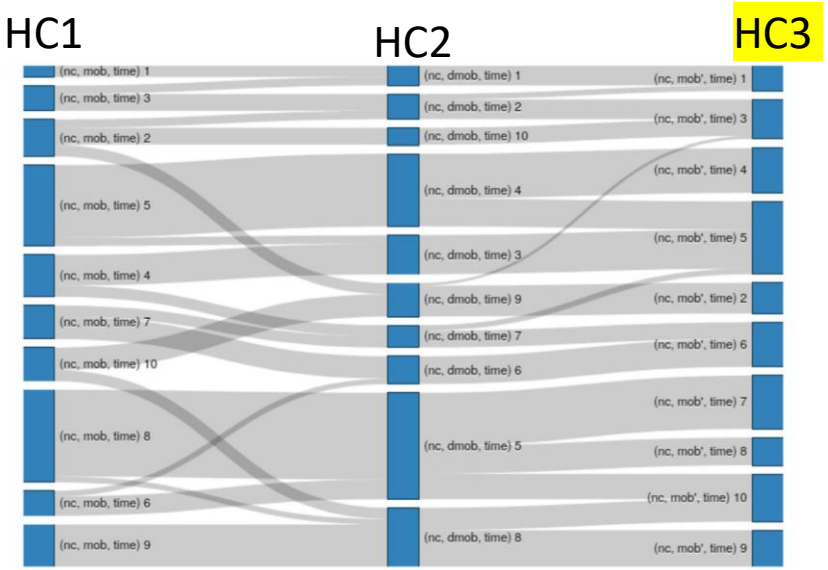


Figure 2: HCMapper is used for comparing 3 hierarchical clusterings: $HC1(N, M, t)$, $HC2(N, \Delta M, t)$ and $HC3(N, M', t)$. The cluster sizes and divergences across the clusterings are shown with blue rectangles and grey edges respectively

Among the clusterings, HC3 appeared to have clusters of consistent sizes, and also the fewest singularities and divergences.

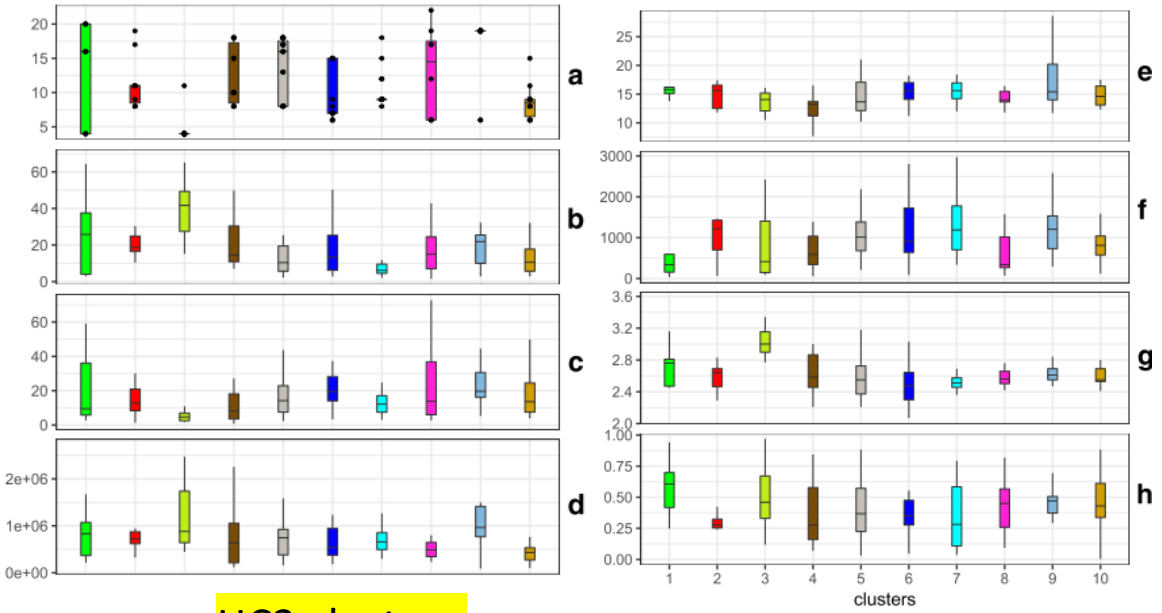
Related research

High spatial correlation (Moran's I giving p-value of 0)

Figure 3: The geographic distribution of the 10 clusters of COVID-19 affected U.S. cities as identified by $HC3$ are shown. The county corresponding to each city is mapped in its cluster-specific color

Related research

The boxplots show the differences across the identified 10 clusters of cities identified by HC3 in terms of the values of the 8 most significant covariates: a Reaction Time (RT), b Hispanic percent, c Black percent, d population size, e senior percent, f population density 2010, g persons per household, and h SVI ses.



HC3 clusters

Related research

Using Random Forest (RF) classification, we identified **socioeconomic characteristics** (based on the covariates) of each of the cities that could **discriminate among the assigned cluster labels**.

- Reaction time is the most significant feature in both, and the 3 least significant features also appear in the same order.
- The ranks hardly change for most features, and some like Hispanic percent and persons per household are highly correlated.

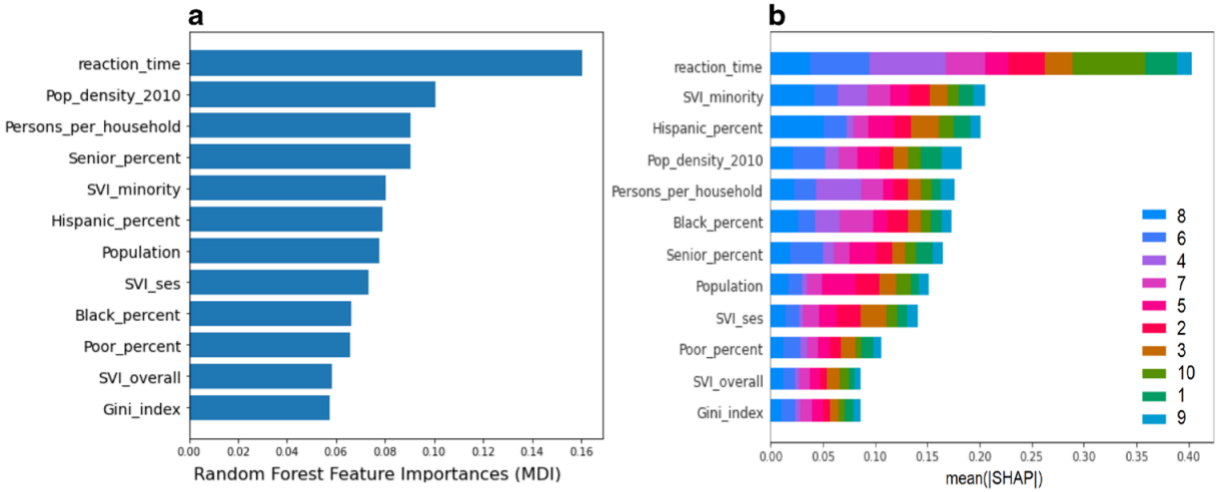


Figure 6: The most significant of the static city-specific covariates in discrimination of the 10 clusters identified by $HC3$. The contributions towards each cluster are measured by **a** the embedded method of RF classifier (MDI), and **b** the mean Shapley values for each covariate

Related research

Conclusion

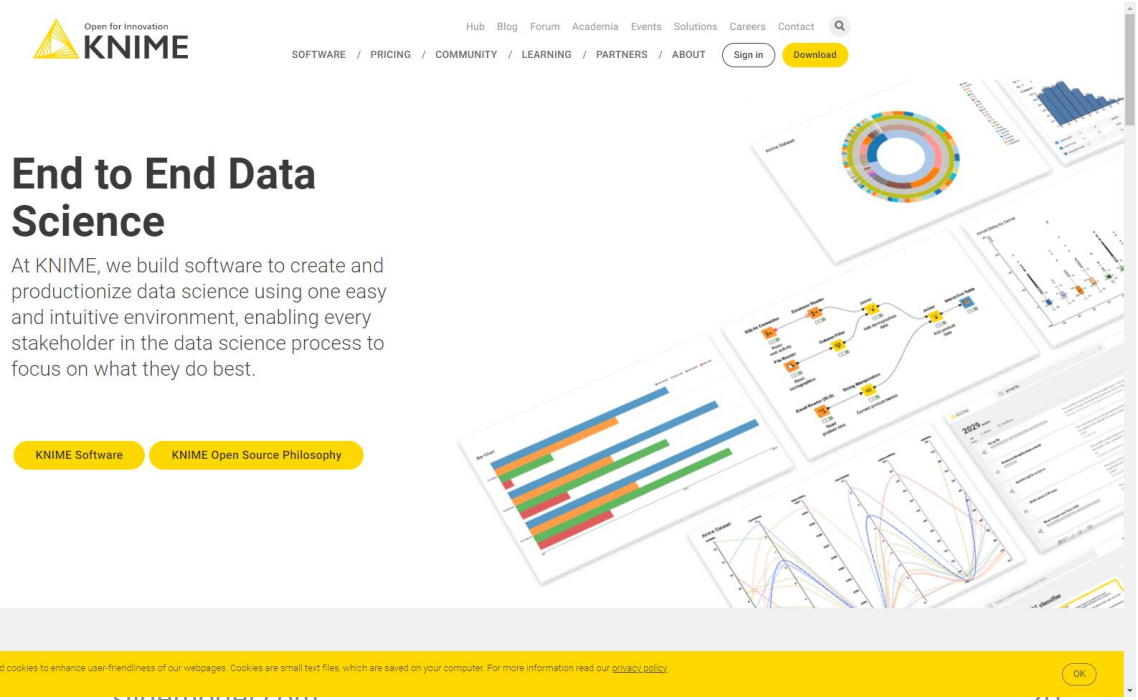
In this study, we have shown that such dependencies not only change over time but across locations and populations, and are likely to be determined by underlying socioeconomic characteristics.

Our analytical approach is particularly relevant considering the high socioeconomic costs of such measures.

The Flowchart for Data Integration & Visualization

KNIME

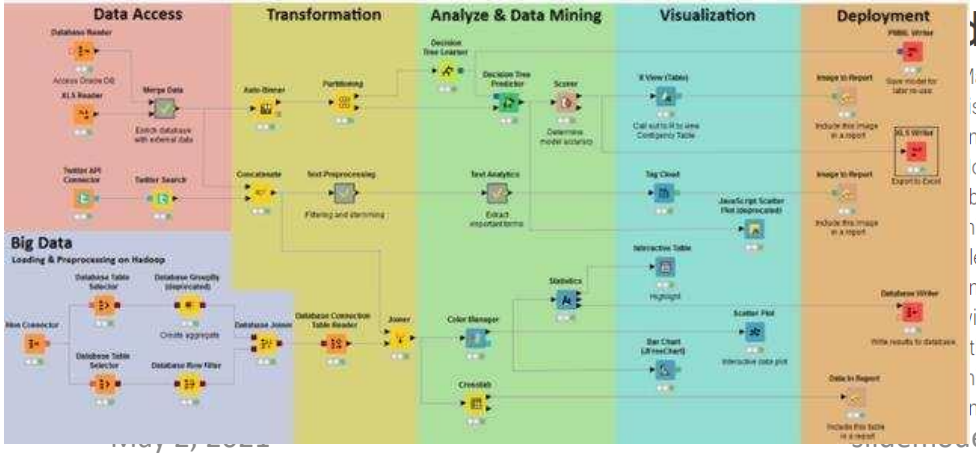
https://www.knime.com/



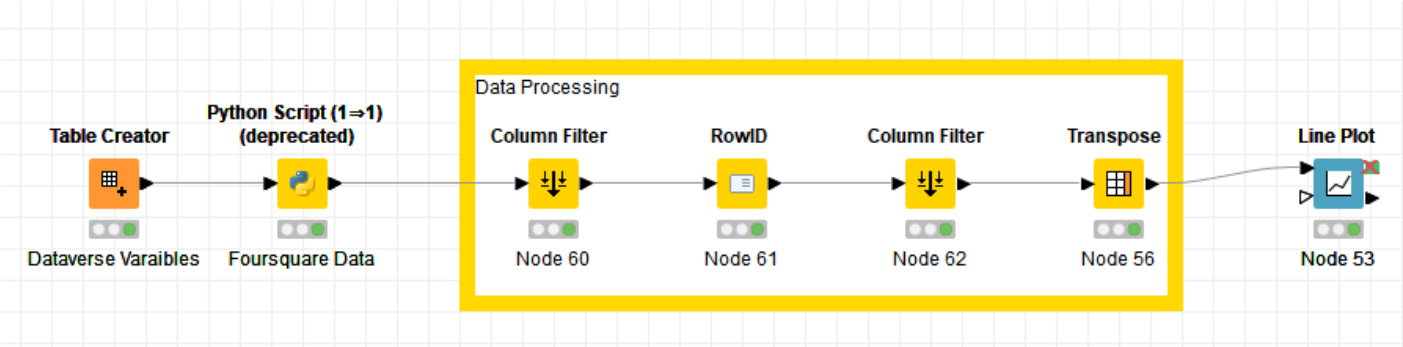
The Flowchart for Data Integration & Visualization

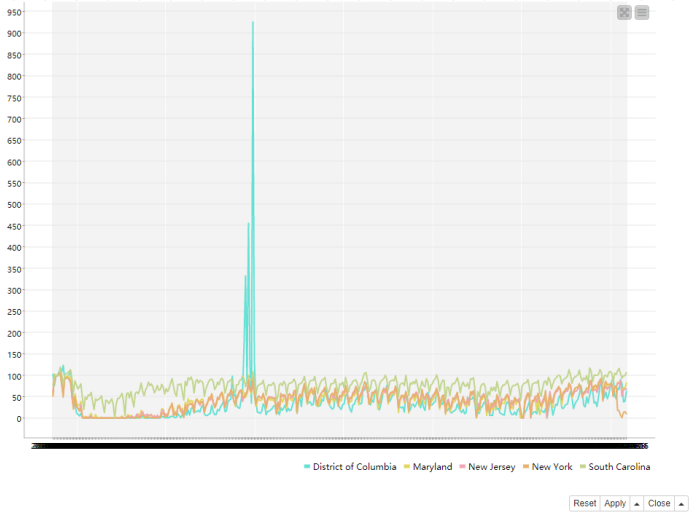
KNIME | Open for innovation

1.The freedom to use the tools you like
2.Retain quality and accuracy in your analytics
3.Spend your time where it makes the most impact
4.Easily deploy and scale your work



Descartes Lab Social Distancing Index Access and Analysis with KNIME





Descartes Lab Social Distancing Index Access and Analysis with KNIME

Step 1: **Download KNIME 4.3.2 from** https://www.knime.com/downloads/download-knime

Step 2: **Download workflow (02_Descartes_Lab_Social_Distancing_Index_Data_Connection.knwf)** from **Google Drive**

Step 3: **Open KNIME** from local PC or China Data Lab Cloud Platform

Step 4: **Import KNIME workflow** file (**02_Descartes_Lab_Social_Distancing_Index_Data_Connection.knwf**)

Step 5: **Configure "Input Data"** for each

Step 6: Click **Run** function from the top menu

Step 7: **Display the outputs**:

- **Table view** with state level mobility index over time

*******Before Running the Workflow*******

1. Please install Anaconda as the Python environment
2. Configure the python path in KNIME
3. Install pyDataverse library in Anaconda
   > pip install pyDataverse==0.2.1

References

Hu, Tao, et al. "Human Mobility Data in the COVID-19 Pandemic: Characteristics, Applications, and Challenges." *ResearchGate Preprint, DOI* 10 (2021).

Warren, Michael S., and Samuel W. Skillman. 2020. “Mobility Changes in Response to COVID-19.” ArXiv:2003.14228 [Cs], March. http://arxiv.org/abs/2003.14228.

Li, Zhenlong, al. "ODT FLOW: A Scalable Platform for Extracting, Analyzing, and Sharing Multi-source Multi-scale Human Mobility." *arXiv preprint arXiv:2104.05040* (2021).

Zhang, M., Wang, S., Hu, T., Fu, X., Wang, X., Hu, Y., ... & Bao, S. (2021). Human mobility and COVID-19 transmission: a systematic review and future directions. *medRxiv*.

M. Monod et al., Science 371, eabe8372 (2021). DOI: 10.1126/science.abe8372

J. Mossong et al., Social contacts and mixing patterns relevant to the spread of infectious diseases. PLOS Med. 5, e74 (2008). doi: 10.1371/journal.pmed.0050074; pmid: 18366252.

J. Zhang et al., Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. Science 368, 1481–1486 (2020). doi: 10.1126/science.abb8001;pmid: 32350060.