Human Mobility Analytics with Big Geosocial Data
Challenges and Approaches

Zhenlong Li

Associate Professor, Department of Geography
Director, Center for GIScience and Geospatial Big Data
Co-Lead, Social Media Core of Big Data Health Science Center
Geoinformation and Big Data Research Laboratory (GIBD)
University of South Carolina
zhenlong@sc.edu
http://gis.cas.sc.edu/gibd

UNIVERSITY OF South Carolina

The 3rd ISDE International Lectures
03/21/2022

Center for GIScience and Geospatial Big Data
UNIVERSITY OF SOUTH CAROLINA

Geoinformation & Big Data Research Laboratory
GIBD
Making sense of geospatial big data.

Mapping the World with Night Lights (Remote Sensing)

From NASA Earth Observations (2016)

https://earthobservatory.nasa.gov/features/NightLights

Mapping the World with Geotagged Tweets
(Social Sensing)

~1.5 billion geotagged tweets
from July 1st, 2017 to June 30th, 2018

My lab has been streaming worldwide geotagged tweets since 2015. Over 8 billion geotagged tweets have been collected so far.

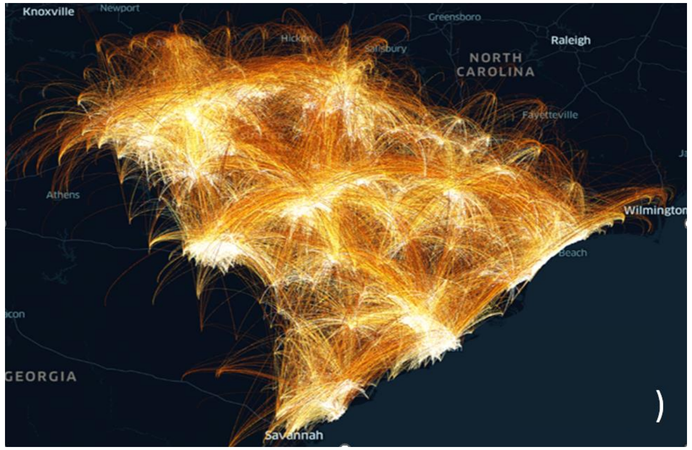
Geosocial data

Geosocial data refers to the geographically referenced information generated by <u>human activities</u> through

- social media platforms (e.g., Twitter, Weibo)
- mobile devices (with opt-in mobile apps, e.g., SafeGraph data)
- other location-aware applications (e.g., Taxi trip data, smart card data)

Digital “geographic footprint”

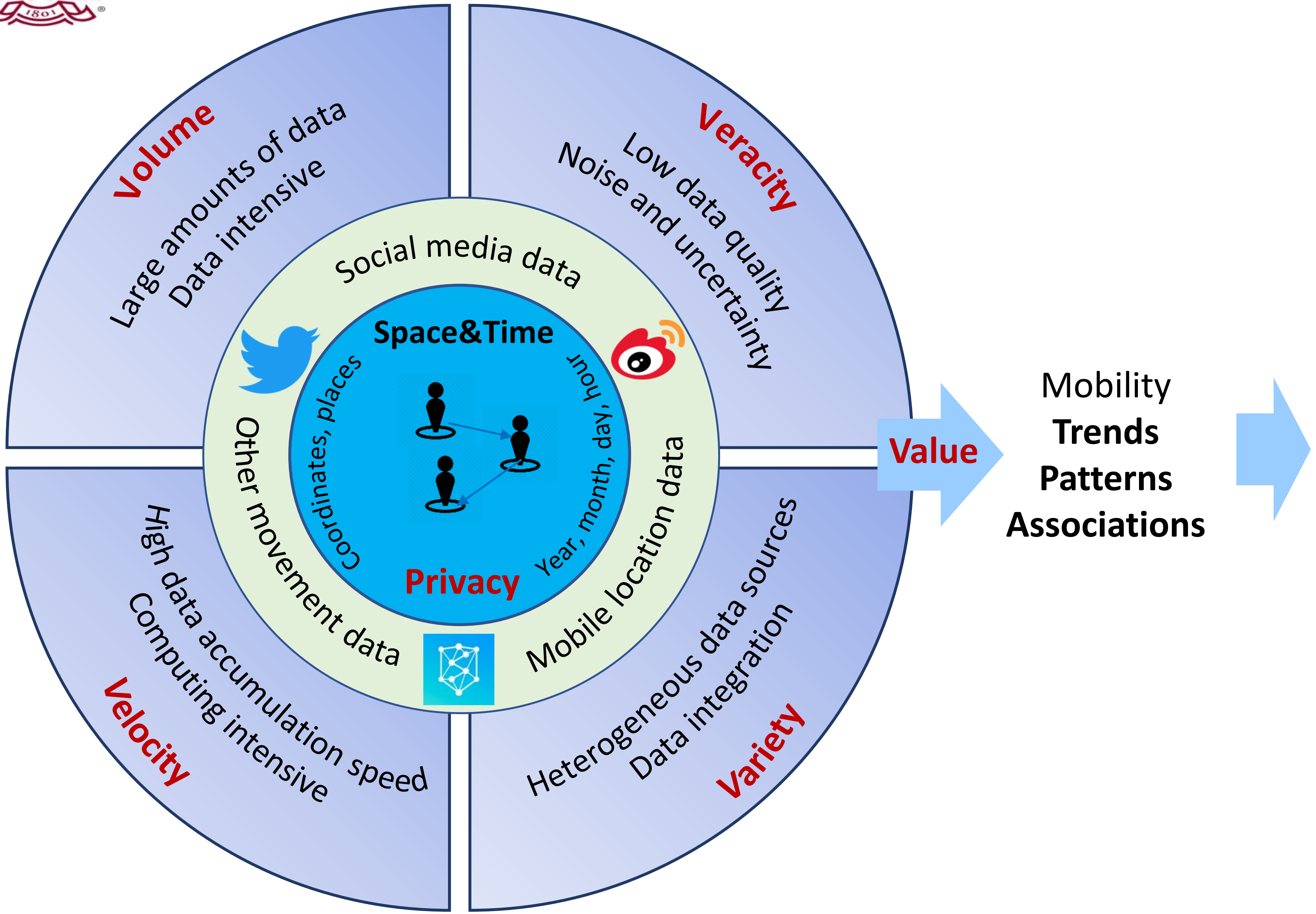
1.5 billion geotagged tweets in one year



Over 1 million visitation flows from home blockgroup to over 3,000 fast-food restaurants in SC in January 2019

1.3 billion taxi dropoffs in NYC metro area (Shekhar R., tinyurl.com/435kbnny)

Challenges of using big geosocial data for human mobility analytics



1. Computational challenges
Data management, processing, analysis, mining, modeling, and geo-visualization.

Five X-bilities or ASIRS
Accessibility
Scalability
Interoperability
Reproducibility
Shareability

2. Bias/representativeness

3. Privacy concerns

Origin-Destination-Time (ODT) Flow Platform
A scalable platform for integrating, analyzing, and sharing multi-source multi-scale human mobility data.

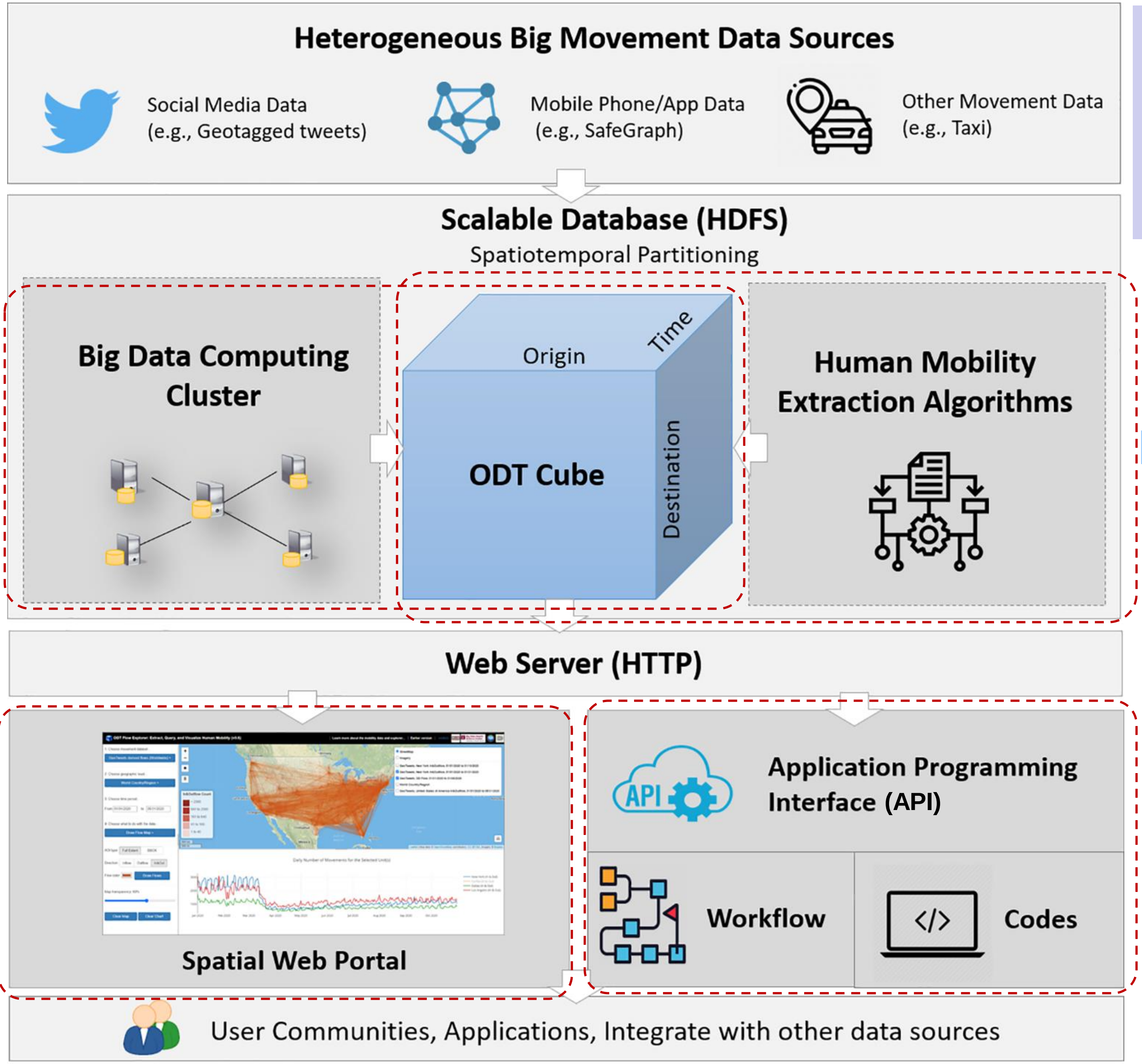
Understanding the bias/representativeness issues

Li Z., Huang X., Hu T., Ning H., Ye X., Huang B., Li X. (2021). ODT FLOW: Extracting, analyzing, and sharing multi-source multi-scale human mobility. *Plos One*, 16(8), e0255259. https://doi.org/10.1371/journal.pone.0255259

Li Z., Ning H., Jing F., Lessani N., (2023). Understanding the bias of mobile location data across spatial scales and over time: a comprehensive analysis of SafeGraph data in the United States, *Preprint*. https://tinyurl.com/2p9vw3ru

Architecture of the Origin-Destination-Time (ODT) Flow Platform

ODT Cube is a place-based data model designed to work with HPC to efficiently manage, query, and aggregate billions of OD flows at different spatiotemporal scales.



Enables scalable big data processing (**scalability**)

Enables heterogeneous big data integration at various spatiotemporal scales (**interoperability**)

Enables interactive data access and visual analytics (**accessibility**)

Enables reproduceable analysis workflow (**reproducibility, shareability**)

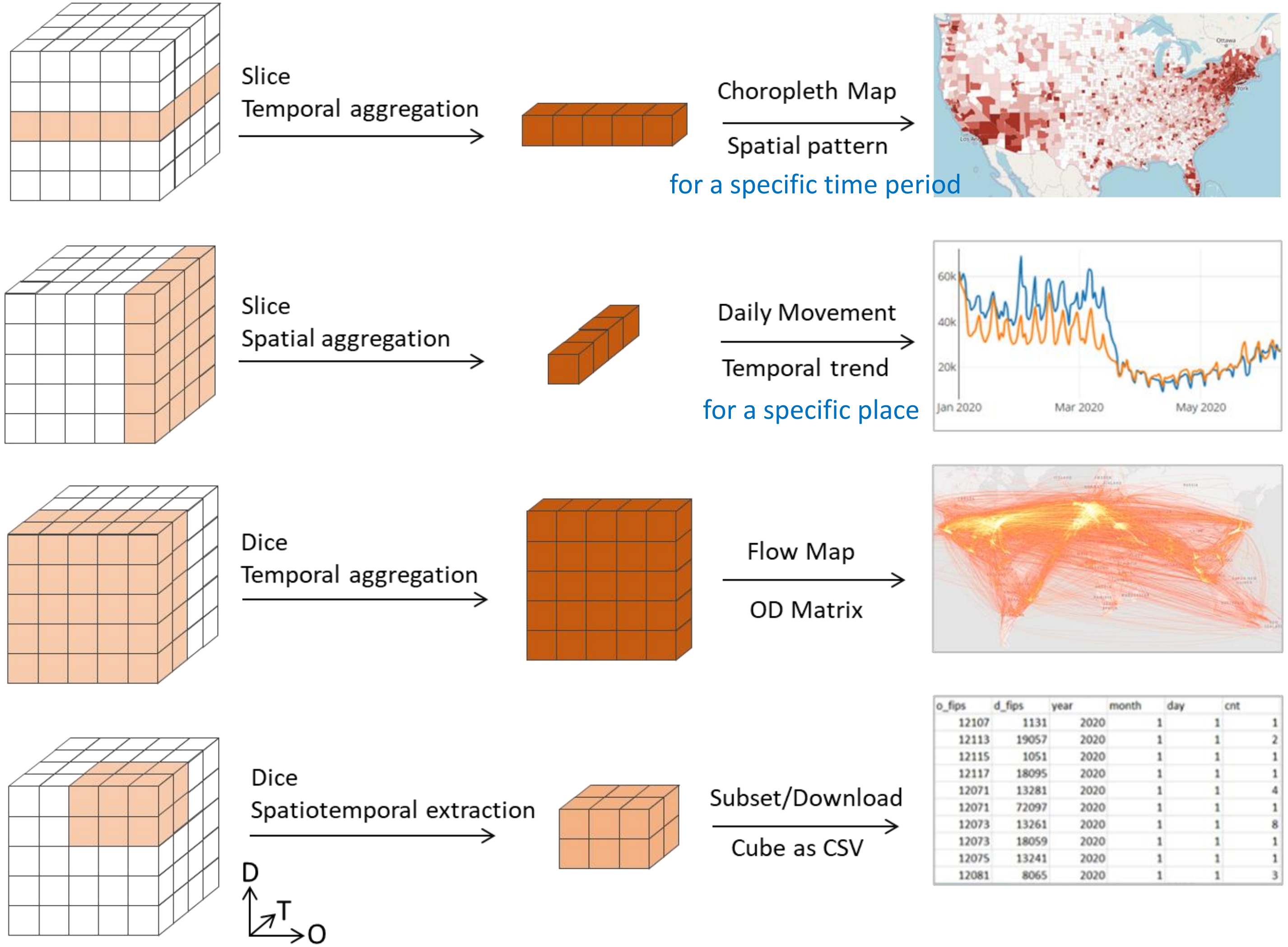
ODT-based human mobility analysis powered by high-performance computing

Four application scenarios illustrating how the **ODT Cube** coupled with **HPC and traditional data cube operations** can help analyze big mobility data.

Big Data Computing Cluster with 15 servers

(Apache Hadoop, Hive, Impala, Spark, and Esri GIS Tools for Hadoop)

Parallel query & visual analytics



o_fips	d_fips	year	month	day	cnt
12107	1131	2020	1	1	1
12113	19057	2020	1	1	2
12115	1051	2020	1	1	1
12117	18095	2020	1	1	1
12071	13281	2020	1	1	4
12071	72097	2020	1	1	1
12073	13261	2020	1	1	8
12073	18059	2020	1	1	1
12075	13241	2020	1	1	1
12081	8065	2020	1	1	3

ODT-based mobility data model enables us to handle different data sources in a unified way

- We computed the **daily OD flows** for 2019 and 2020 using worldwide geotagged tweets.
- We further computed the daily OD flows from mobile location data from SafeGraph.

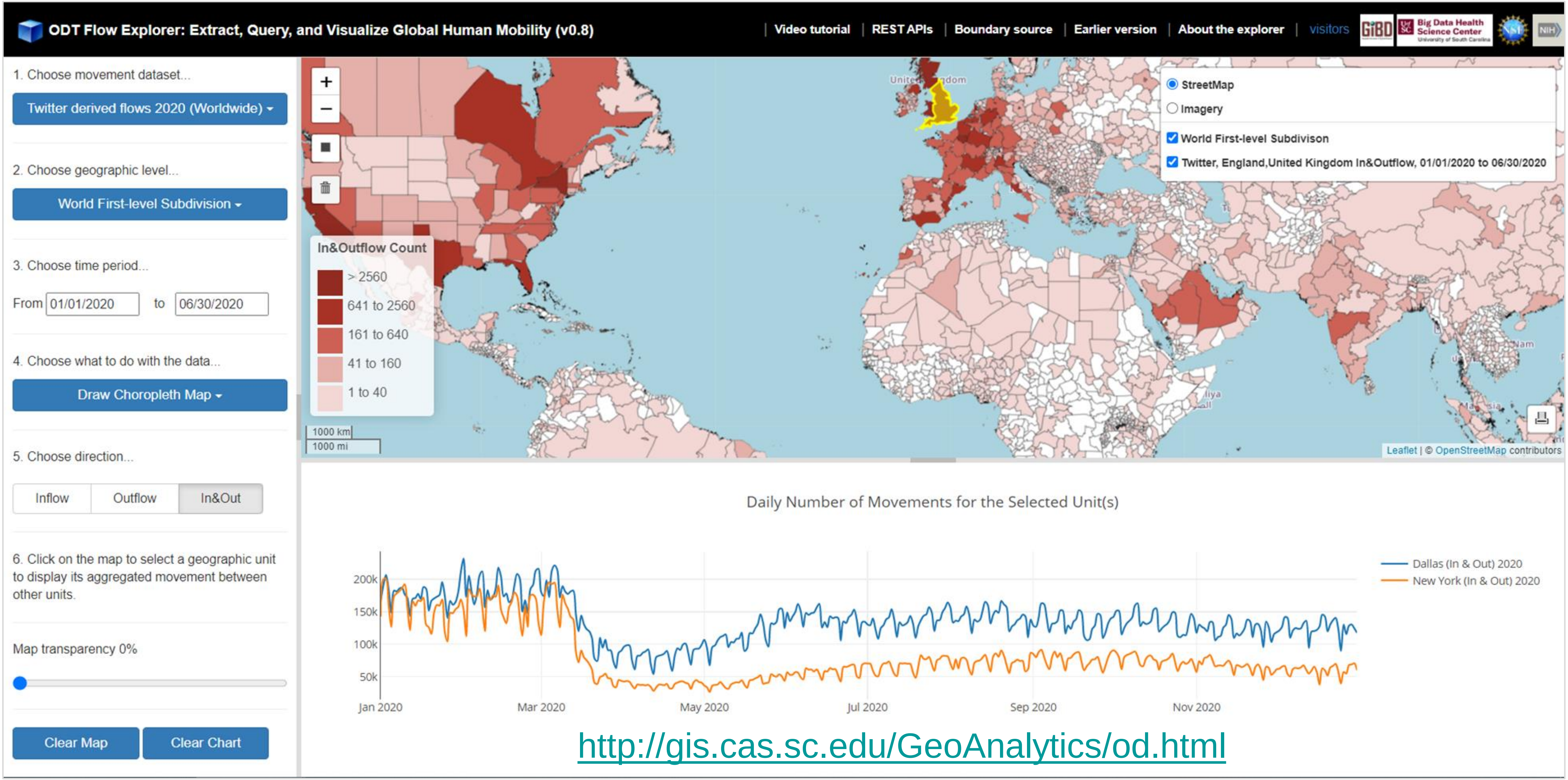
Statistics of the derived **daily flows** from Twitter data and SafeGraph data

	Twitter-derived OD Flow	**Cellphone-derived OD Flow**
Spatial coverage	Worldwide	U.S.
Temporal coverage	2019-2020 (daily)	2019-2021 (daily)
Original data records	2,695,552,594 geotagged tweets by 24,863,844 Twitter users	160,301,510 SafeGraph data records
Derived Entity-ODT	636,984,772	11,108,696,071
World country	1,253,291	—
World 1st level subdivision	9,333,761	—
U.S. state	809,741	1,958,450
U.S. county	**10,206,119**	439,790,381
U.S. census tract	—	**6,710,889,890**

[1]https://docs.safegraph.com/docs/social-distancing-metrics

ODT Flow Explorer: Interactive mobility data access and visual analytics

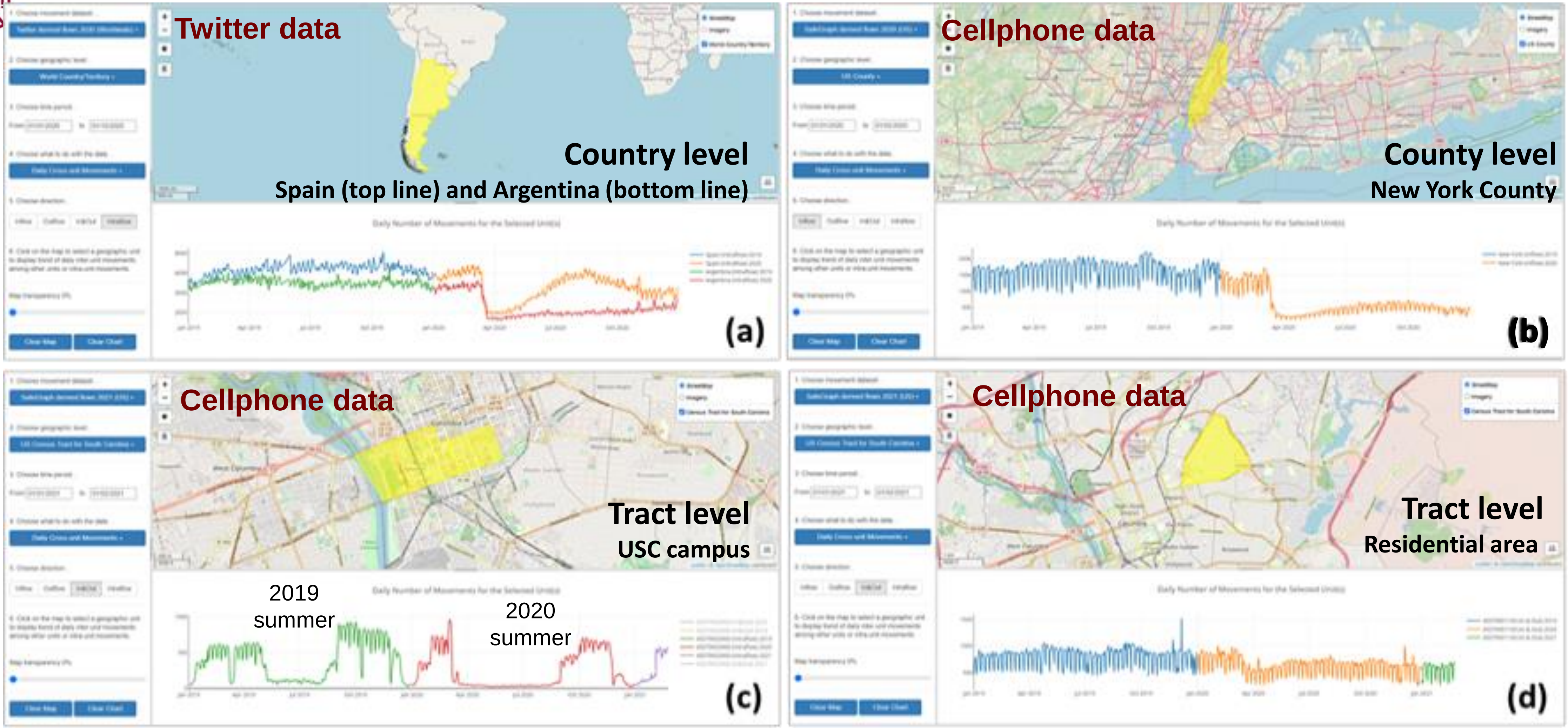
An interactive spatial web portal for on-demand querying, aggregating, and visualizing the billion-level OD flows.



http://gis.cas.sc.edu/GeoAnalytics/od.html

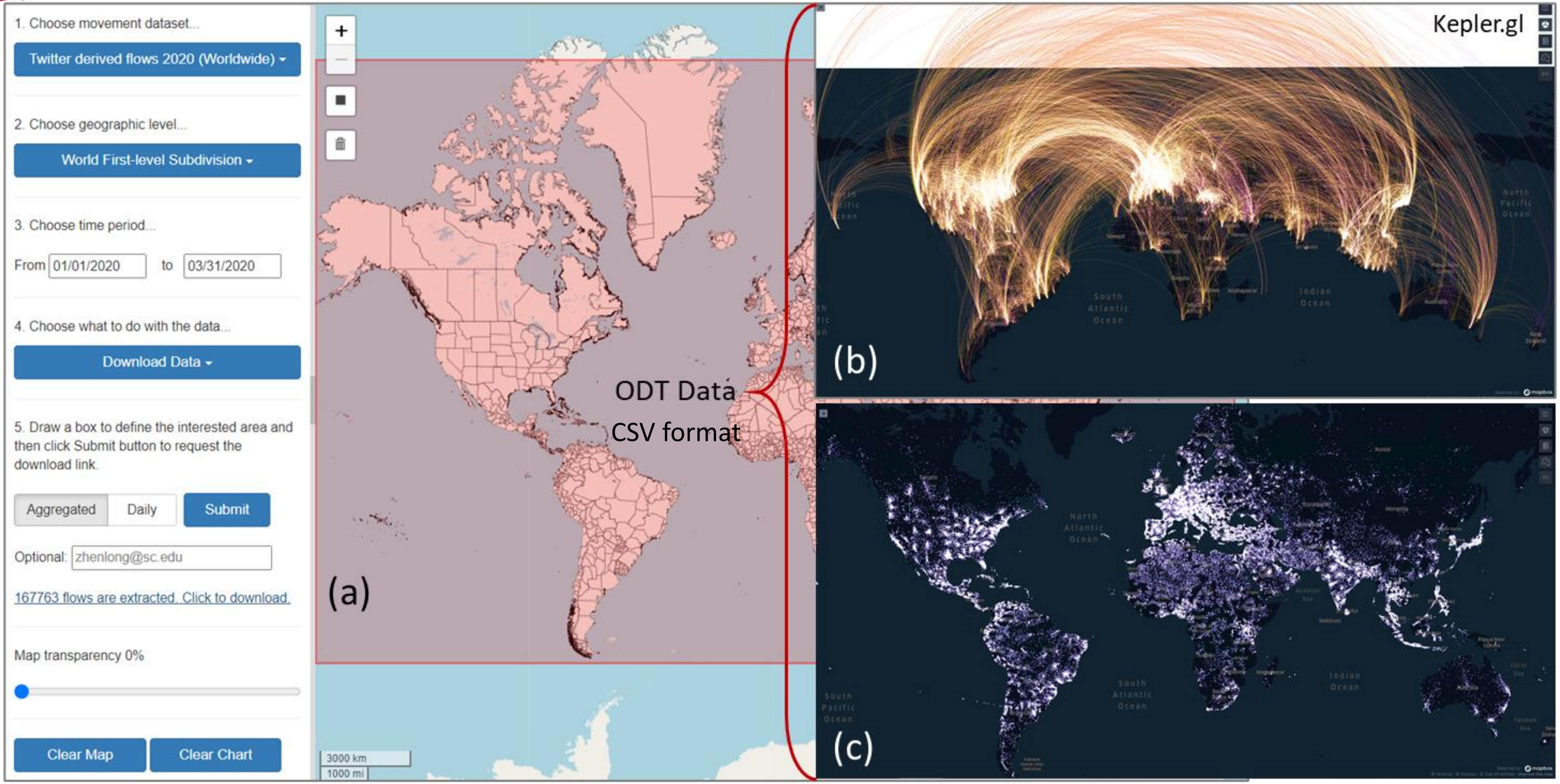
ODT Flow Explorer

Impact of the pandemic on daily population mobility at different spatial levels (2019-2020)



(a) Intraflow for Spain (top line) and Argentina (bottom line) in 2019 and 2020;
(b) Inflow for New York County, U.S. in 2019 and 2020;
(c) Intraflow for a census tract in Columbia, South Carolina (mainly located within the USC) from 01/01/2019 to 02/24/2021;
(d) Intraflow for a census tract in a residential area of Columbia from 01/01/2019 to 02/24/2021.

ODT Flow Explorer
Extract and download flow data with user-defined spatiotemporal constraints



ODT Flow REST API: Access flow data programmatically

ODT Flow REST APIs

Each API performs a specific task such as aggregating the flows for a selected place and downloading flow data for a selected geographic area. All APIs return data in CSV (comma-separated values) format. The API is specified in the "operation" parameter in the request (see examples below).

APIs

- **get_flow_by_place**

Return the aggregated movement between the selected place and other places.

- **get_daily_movement_by_place**

Return the daily inter-unit movements between the selected place and other places or the selected place's daily intra-unit movements.

- **get_daily_movement_for_all_places**

Return the daily movements for all places of a specific geographic level (currently return intra movement).

- **extract_odt_data**

Return the selected OD flows in either temporally aggregated format or daily format. The study area can be specified by a bbox. For SafeGraph daily flows, the days selected need be less than 31.

- **extract_odt_data_url**

Same as extract_odt_data, but returns a download URL and number of records instead of directly returning the csv data. Works better for extracting large amounts of flows.

extract_odt_data

Return the selected OD flows in either temporally aggregated format or daily format. The study area can be s selected need be less than 31.

In [11]:

```
# set the parameters of your interested data, including operation, scale, source, place..
params = {"operation": "extract_odt_data",
       "source": "twitter",
       "scale": "us_county",
       "begin": "04/01/2019" ,
       "end": "04/15/2019",
       "bbox": "-90,90,-180,180",
       "type": "daily"}

# obtain data using REST APIs
q = r'http://gis.cas.sc.edu/GeoAnalytics/REST'
r = requests.get(q, params=params)

# put the data into a Pandas DataFrame
df = pd.read_csv(StringIO(r.text))
df
```

Out[11]:

	o_place	d_place	year	month	day	cnt	o_lat	o_lon	d_lat	d_lon
0	21115	21115	2019	4	8	5	37.811	-82.816	37.811	-82.816
1	1099	1001	2019	4	7	1	31.523	-87.335	32.576	-86.681
2	36029	36121	2019	4	12	1	42.969	-78.582	42.867	-78.362
3	17109	17031	2019	4	9	1	40.460	-90.674	42.020	-87.772
4	51550	51550	2019	4	10	100	36.761	-76.289	36.762	-76.294
5	51041	51760	2019	4	12	13	37.441	-77.531	37.532	-77.493
6	49057	49011	2019	4	10	4	41.201	-111.990	41.128	-111.997
7	13121	39001	2019	4	2	1	33.740	-84.449	38.906	-83.347
8	18127	18167	2019	4	8	1	41.499	-87.067	39.486	-87.409
9	26125	42091	2019	4	8	1	42.491	-83.143	40.124	-75.458
10	39003	39095	2019	4	9	1	40.887	-83.899	41.657	-83.575
11	24013	24003	2019	4	3	1	39.577	-76.998	39.133	-76.625
12	37135	37101	2019	4	15	1	35.927	-79.087	35.723	-78.418

https://github.com/GIBDUSC/ODT_Flow

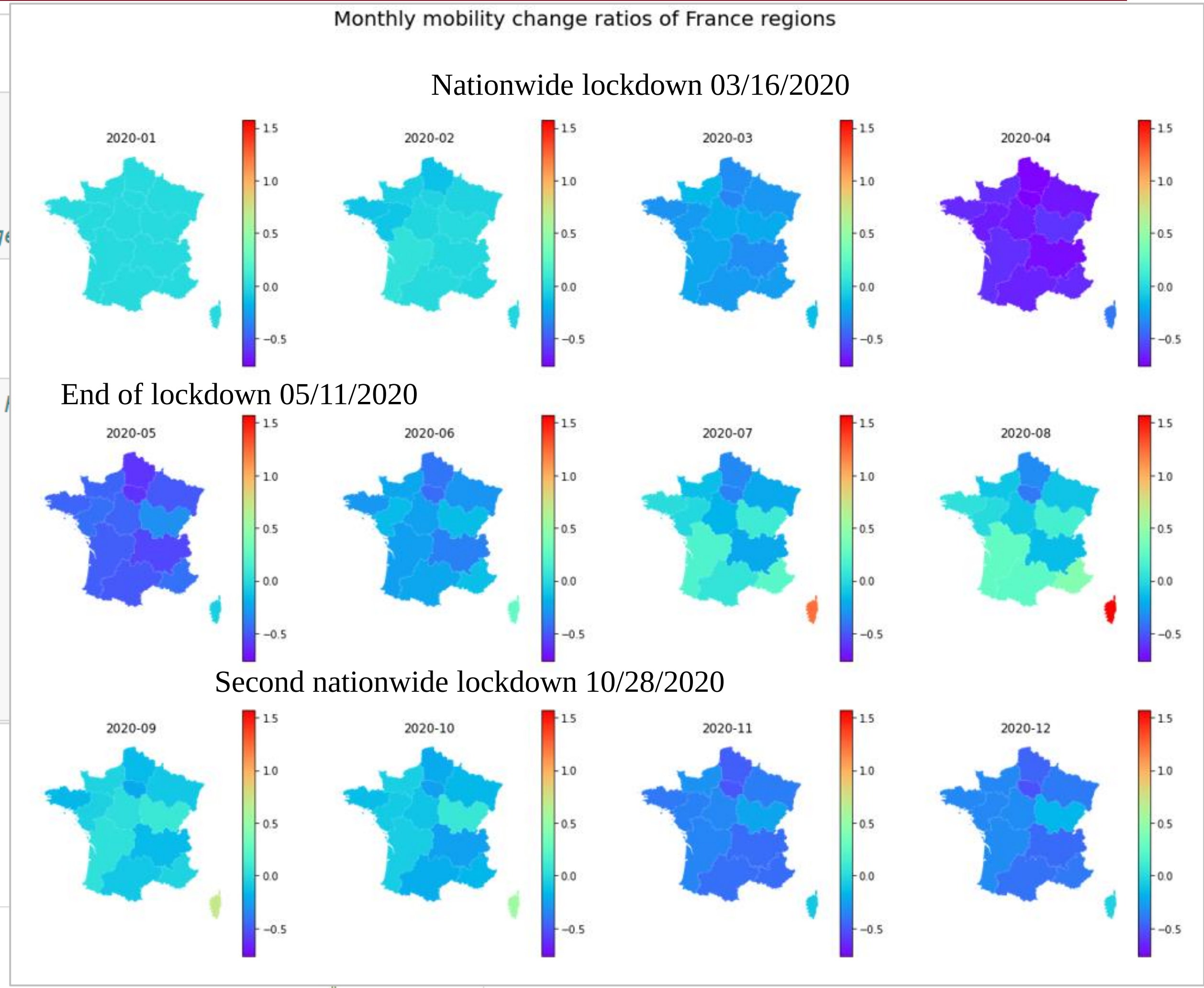
Use the ODT Flow API in Jupyter Notebook
Visual analytics of COVID-19 impact on human mobility in France in 2020

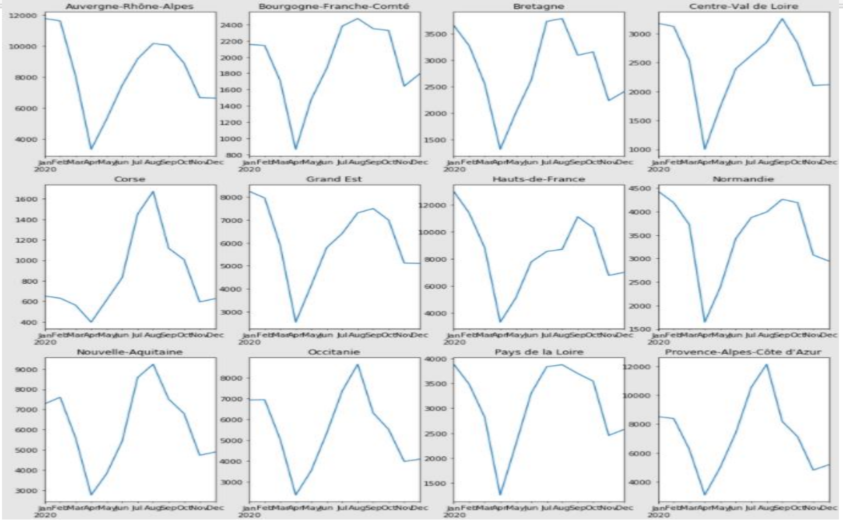
Read the boundary file

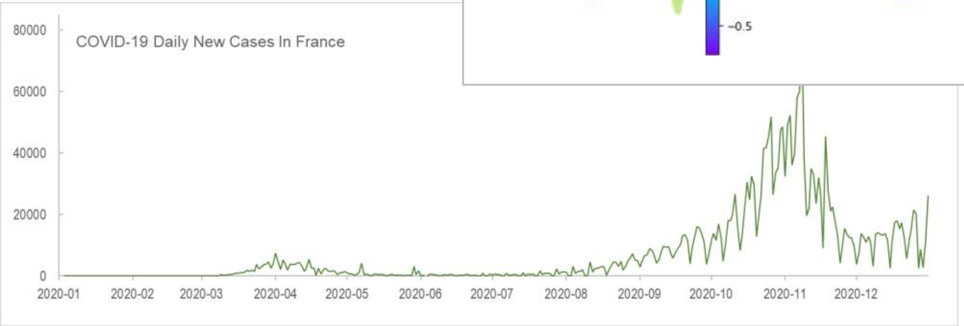
```
subdivision_file = r'gadm01_simplified/gadm36_1.shp'
gdf = gpd.read_file(subdivision_file)

target_place = r'FRA'   # set France as the target place (ISO code)
gdf_country = gdf[gdf['GID_1'].str[:3] == target_place] # Extract the boundary of the targe
```

Obtain 2020 flow data using the ODT Flow API

```
q = r'http://gis.cas.sc.edu/GeoAnalytics/REST'  #Set query url and parameters for the ODT F
params = {"operation": "get_daily_movement_for_all_places",
          "scale": "world_first_level_admin",
          "source": "twitter",
          "begin": "01/01/2020",
          "end": "12/31/2020"
     }
r = requests.get(q, params=params) #Submit request
df = pd.read_csv(StringIO(r.text))

df = df[df['place'].str[:3] == target_place] # Extract flows of the target place
```



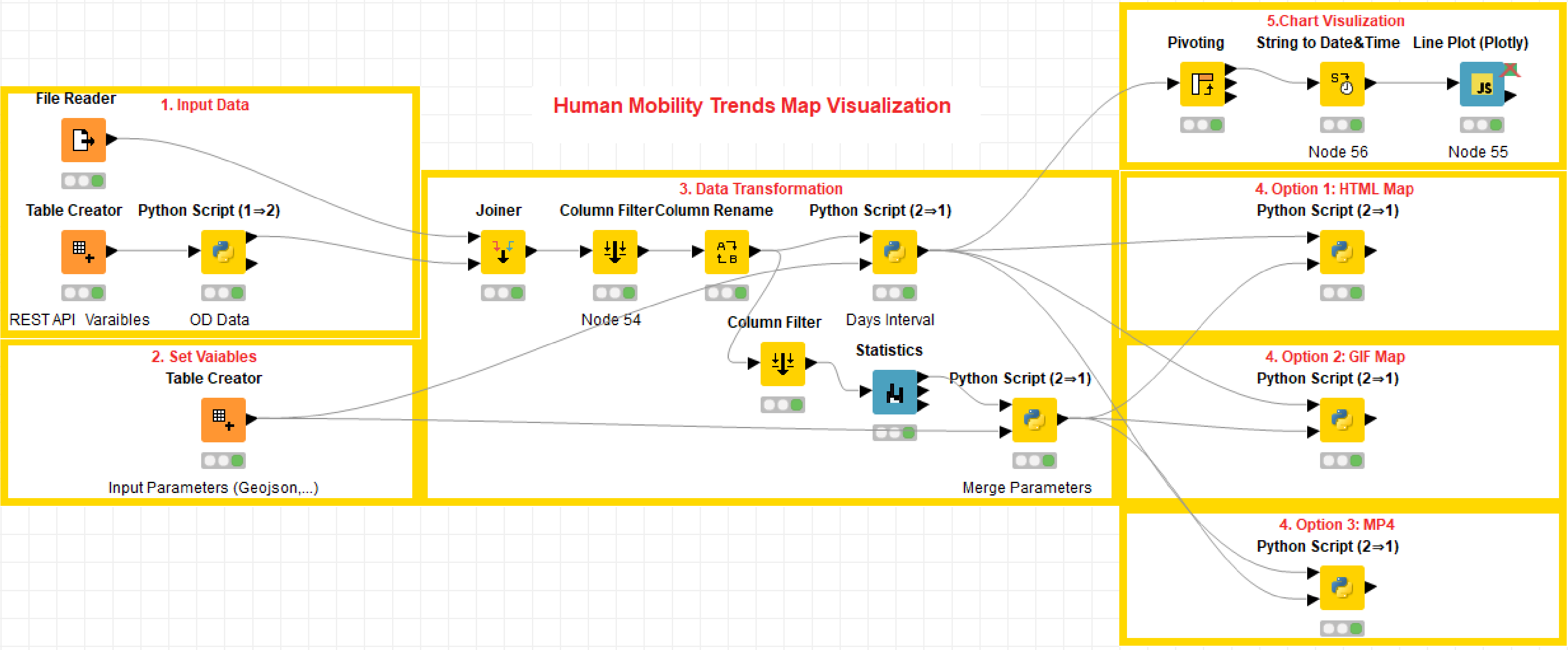


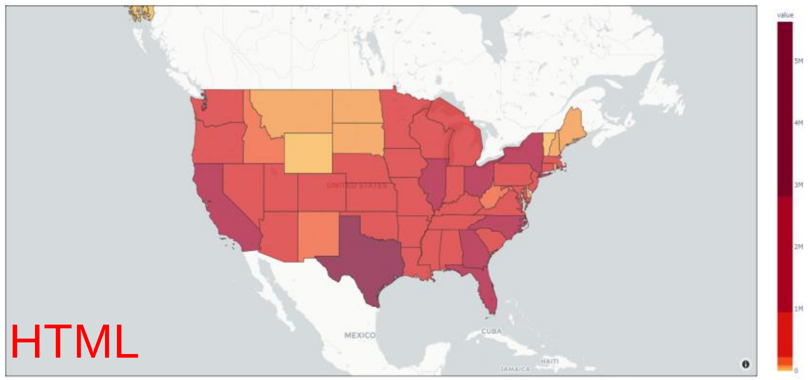


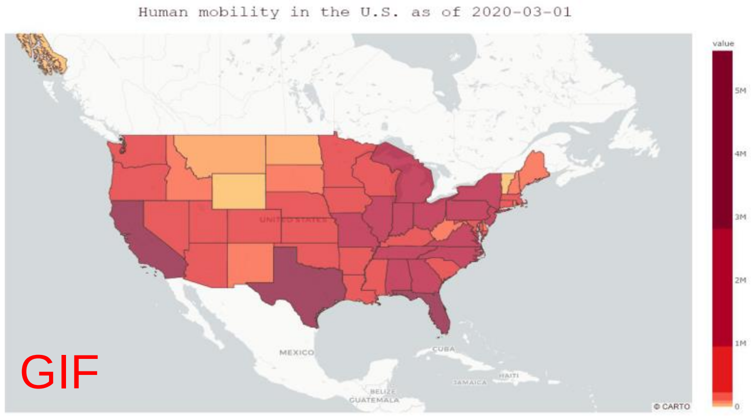
https://github.com/GIBDUSC/ODT_Flow/tree/main/API%20with%20Jupyter%20Notebook%20Case%20study%201

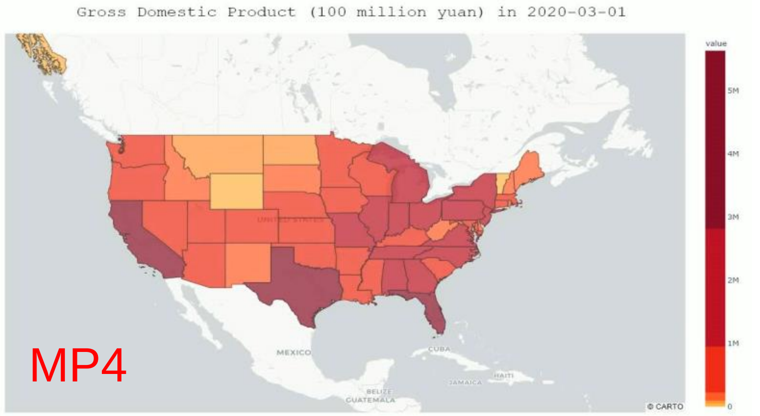
Use the ODT Flow API with Data Science Workflow Tool KINME (enable reproducibility)

Human Mobility Trends Visualization with Dynamic Map







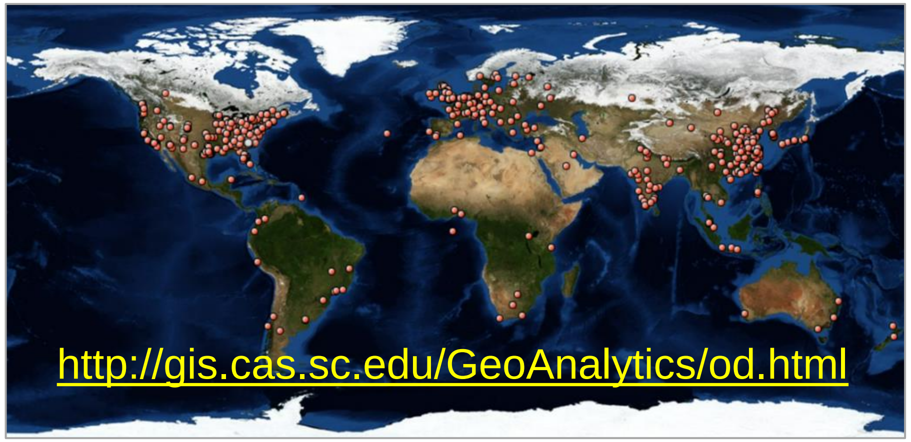


Credit: Dr. Tao Hu, Oklahoma State University

https://github.com/GIBDUSC/ODT_Flow/tree/main/KNIME%20workflow%20case%20studies

The ODT Flow Platform has been used by other researchers around the world

The ODT Flow Platform has attracted over **5,000** visitors from **69** countries, served over **3.8 billion** flow extractions.



WorldPop

UNIVERSITY OF Southampton

February 24th, 2021

Preliminary risk analysis of the international spread of new COVID-19 variants, lineage B.1.1.7, B.1.351 and P.1

Shengjie Lai, Jessica Floyd, Andrew Tatem

WorldPop, School of Geography and Environmental Science, University of Southampton, UK

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2022.2161426

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Evaluating COVID-19's impacts on Puerto Rican's travel behaviors

Lauren C. Carter and Ran Tao

School of Geosciences, University of South Florida, Tampa, FL, USA

Right Idea, Wrong Place? Knowledge Diffusion and Spatial Misallocation in R&D

97 Pages • Posted: 17 Feb 2023

Trevor Williams
Yale University, Department of Economics, Students

A fairness assessment of mobility-based COVID-19 case prediction models

Abdolmajid Erfani [1*], and Vanessa Frias-Martinez [2, 3]

[1] Department of Civil and Environmental Engineering, University of Maryland, 1173 Glenn Martin Hall, College Park, MD 20742, USA.

[2] College of Information Studies, University of Maryland, College Park, MD 20742, USA.

[3] University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA

We are continuing the development of ODT Flow Platform

1. Extending the spatial-temporal coverage of the flows extracted from Twitter and SafeGraph
   - Twitter-derived worldwide flows from 2015 to 2022
   - SafeGraph-derived US flows from 2018 to 2022
   - SafeGraph-derived Canada flows from 2018 to 2022

2. Expanding the movement data sources using the ODT model to integrate
   - NYC Taxi Trip data from 2009 to 2022
   - US Census migration mobility data (county and state) from 2000 to 2021

3. Developing more APIs for enhanced data sharing, access, analytics, and interoperability

Bias/Representativeness challenges

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE
2019, VOL. 46, NO. 3, 228–242
https://doi.org/10.1080/15230406.2018.1434834

Taylor & Francis
Taylor & Francis Group

Check for updates

Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level

Yuqin Jiang [a], Zhenlong Li [a] and Xinyue Ye [b]

[a]Department of Geography, University of South Carolina, Columbia, USA; [b]Department of Geography, Kent State University, OH, USA

https://www.tandfonline.com/doi/abs/10.1080/15230406.2018.1434834

Understanding the bias of mobile location data across spatial scales and over time: a comprehensive analysis of SafeGraph data in the United States

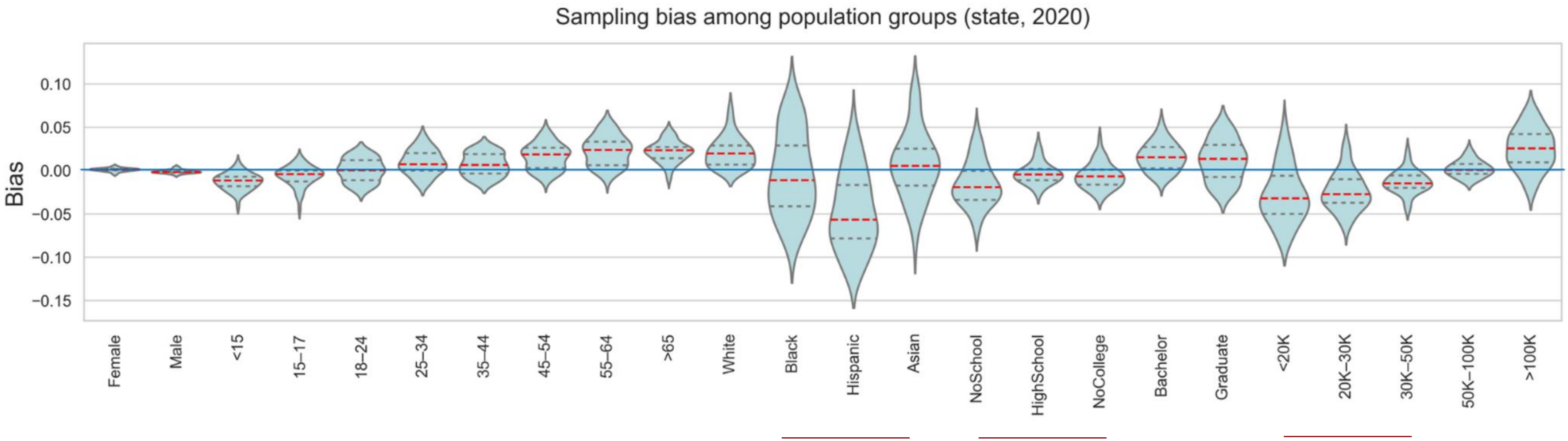
Zhenlong Li*, Huan Ning, Fengrui Jing, M.Naser Lessani

Geoinformation and Big Data Research Laboratory
Department of Geography, University of South Carolina, USA
*zhenlong@sc.edu

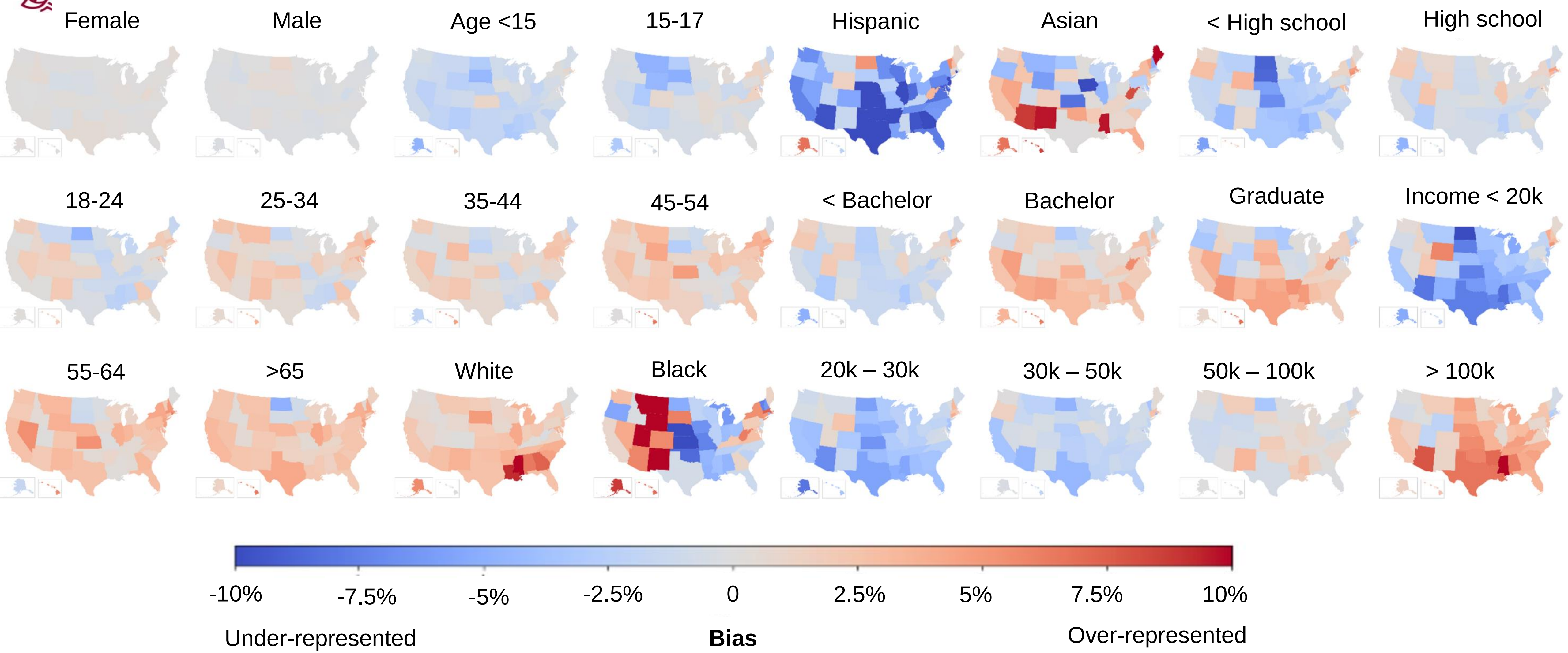
03/2023

Preprint:
https://tinyurl.com/2p9vw3ru

Demographic and socioeconomic bias of SafeGraph mobile location data (2020)



Demographic and socioeconomic bias of SafeGraph mobile location data (2020)



Potential solutions

1. **Conduct sensitivity analyses** to assess the impact of sampling bias on the results.
2. **Apply statistical weighting method** to adjust the data to reflect the true distribution of the population of interest.
3. **Combine with other data sources** to provide additional information about the characteristics of the population.

Acknowledgement

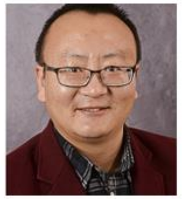
GIBD Lab Core Faculty


Dr. Susan Cutter


Dr. Cuizhen Wang

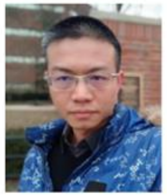

Dr. Michael Hodgson


Dr. Fengrui Jing
(Postdoc Associate)


Dr. Zhenlong Li

Students
(partial list)


Naser Lessani
PhD


Huan Ning
PhD


Seth Church
PhD


Alex Fulham
MS


Breonna Roden
MS

Alumni
(partial list)


Dr. Xiao Huang
(UARK)


Dr. Yago Martin
(UCF)


Dr. Yuqin Jiang
(Texas A&M)

Collaborators
(partial list)

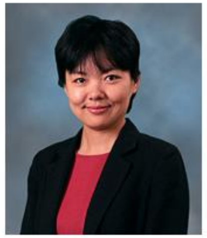

Dr. Xiaoming Li
(USC)


Dr. Dwayne Porter
(USC)


Dr. Xinyue Ye
(Texas A&M)


Dr. Chris Emrich
(UCF)


Dr. Bankole Olatosi
(USC)


Dr. Shan Qiao
(USC)

Sponsors

Thank you !

Questions/Comments?

http://gis.cas.sc.edu/cegis