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
Mapping the World with Night Lights
(Remote Sensing)

From NASA Earth Observations (2016)
https://earthobservatory.nasa.gov/features/NightLights
My lab has been streaming worldwide geotagged tweets since 2015. Over 8 billion geotagged tweets have been collected so far.
Geosocial data refers to the geographically referenced information generated by human activities 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-ibilities or ASIRS**
   - Accessibility
   - Scalability
   - Interoperability
   - Reproducibility
   - Shareability

2. Bias/representativeness

3. Privacy concerns

**Value**

- Mobility
- Trends
- Patterns
- Associations

**Volume**
- Large amounts of data
- Data intensive

**Veracity**
- Low data quality
- Noise and uncertainty

**Velocity**
- High data accumulation speed
- Computing intensive

**Variety**
- Coordinates, places
- Social media data
- Other movement data

**Space & Time**
- Year, month, day, hour, day of week
- Mobile location data
- Heterogeneous data sources
- Data integration
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


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 heterogeneous big data integration at various spatiotemporal scales (interoperability)

Enables scalable big data processing (scalability)

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.

- **Choropleth Map**
  - Spatial pattern for a specific time period

- **Daily Movement**
  - Temporal trend for a specific place

- **Flow Map**
  - OD Matrix

- **Subset/Download**
  - Cube as CSV

**Big Data Computing Cluster with 15 servers**

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

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

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

[Image of the ODT Flow Explorer interface]

http://gis.cas.sc.edu/GeoAnalytics/od.html
Intraflow for Spain (top line) and Argentina (bottom line) in 2019 and 2020;

- Inflow for New York County, U.S. in 2019 and 2020;
- Intraflow for a census tract in Columbia, South Carolina (mainly located within the USC) from 01/01/2019 to 02/24/2021;
- 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

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_uri`
  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.

[Link to GitHub repository](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

```python
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 target place within France
```

Obtain 2020 flow data using the ODT Flow API

```python
q = r'http://gis.cas.sc.edu/GeoAnalytics/REST'  # Set query url and parameters for the ODT Flow API
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
```

Monthly mobility change ratios of France regions

- Nationwide lockdown 03/16/2020
- End of lockdown 05/11/2020
- Second nationwide lockdown 10/28/2020

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. It has attracted over 5,000 visitors from 69 countries, served over 3.8 billion flow extractions.

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

---

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

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

Yuqin Jiang a, Zhenlong Li a and Xinyue Ye b

aDepartment of Geography, University of South Carolina, Columbia, USA; bDepartment of Geography, Kent State University, OH, USA

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

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.

Potential solutions
Thank you!

Questions/Comments?

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