Social Media (Twitter) Big Data, Human Mobility, and COVID-19

Zhenlong Li (李振龙)

Geoinformation and Big Data Research Laboratory (GIBD)

Big Data Health Science Center

Hazards and Vulnerability Research Institute

Department of Geography, University of South Carolina, USA

zhenlong@sc.edu

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

06/29/2020 EDT

Geoinformation & Big Data Research Laboratory

GIBD

Making sense of geospatial big data.

Geospatial Big Data Research at GIBD Lab

Big Social Sensing Data

Big Remote Sensing Data

Big Health Data

Big Climate Data

Innovations

Methods, Algorithms, Framework, Tools

Big Data Computing/Analytics

Disaster Management

Human Mobility

Public Health (HIV, suicide..)

Climate Analysis

Sponsors:

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

Powered by Big Data Computing Facilities at GIBD

GeoRapider: Geospatial Big Data Computing Cluster with 15 Computer Servers (and growing)

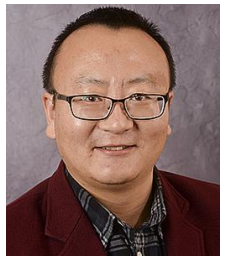
housed at the Research Computing Center, USC

AI (Deep Learning) Workstations

powered by high-end NVIDIA Titan XP GPUs

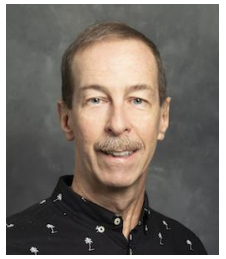
Enabled by a Strong Interdisciplinary Team

Lab Core Faculty

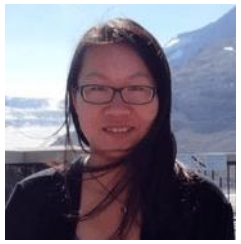

Dr. Zhenlong Li

Dr. Susan Cutter

Dr. Cuizhen Wang


Dr. Michael Hodgson

Students


Yuqin Jiang

Grayson Morgan

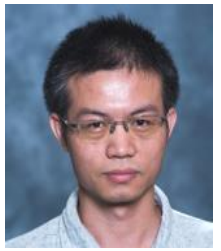
Qian Huang

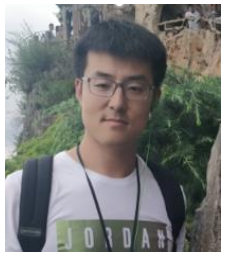
Finn Hagerty

Alumni

Dr. Xiao Huang (UARK)

Dr. Yago Martin (UCF)


Huan Ning (NJIT)

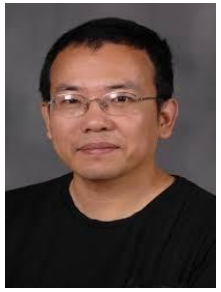

Dong Xu (ECNU)

Collaborators
(not a full list)

Dr. Xiaoming Li (USC)

Dr. Amir Karami (USC)

Dr. Dwayne Porter (USC)


Dr. Xinyue Ye (Texas A&M)

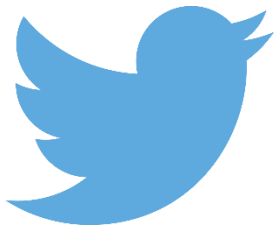
Dr. Chris Emrich (UCF)

Dr. Qunying Huang (UWM)

Dr. Sathish Kumar (CCU)

Streaming Geotagged Tweets

- We have been streaming worldwide geotagged tweets since 2015
  - **Over 4 million** geotagged tweets is being streamed **every day.**
  - **Over 7 billion** geotagged tweets have been collected so far (stored in a highly secured server with Duo Two Factor Authentication).



Real-time streaming

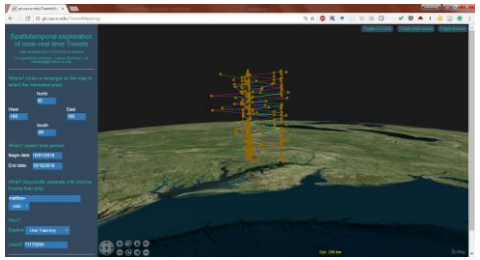
Twitter API

GeoRapider

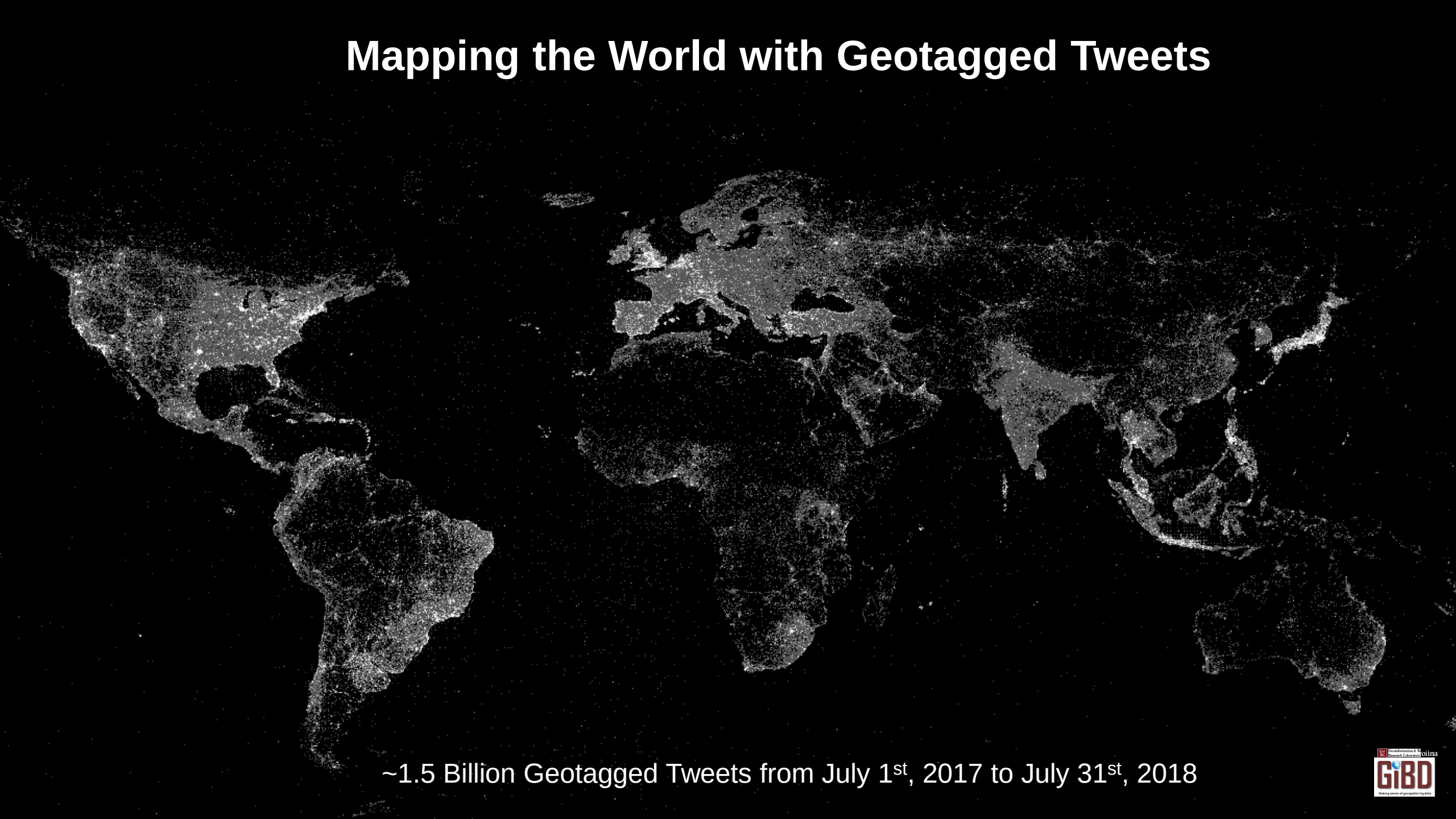
High Performance in-memory Computing cluster
Hadoop Distributed File System

Firewall

Parallel query analytics, mapping



Mapping the World with Geotagged Tweets



~1.5 Billion Geotagged Tweets from July 1st, 2017 to July 31st, 2018

Geotagging

Twitter allows users to share locations

- **Coordinates**: latitude and longitude

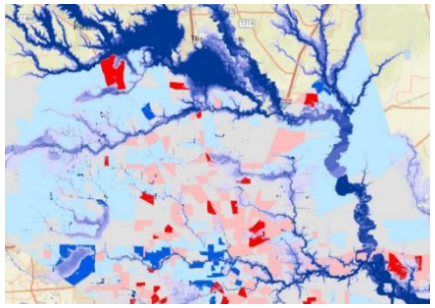
- **Places:** At different levels: POI, neighborhood, city, state, country

e.g., Columbia, SC

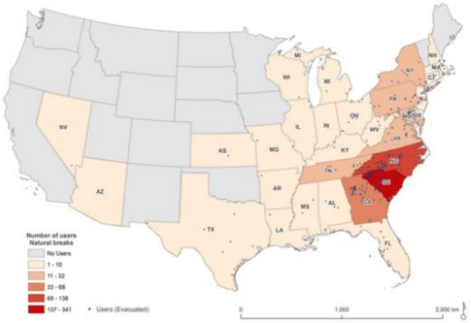
When individuals send a tweet, they leave a digital “geographic footprint” .

What studies have we done with these massive data?

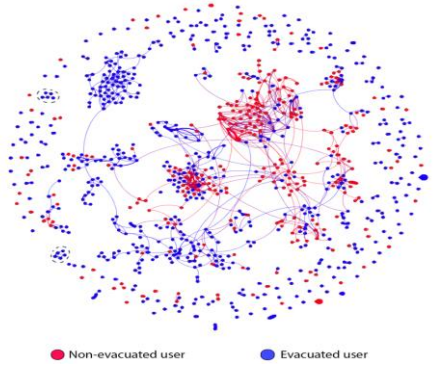
Rapid flood mapping for situational awareness



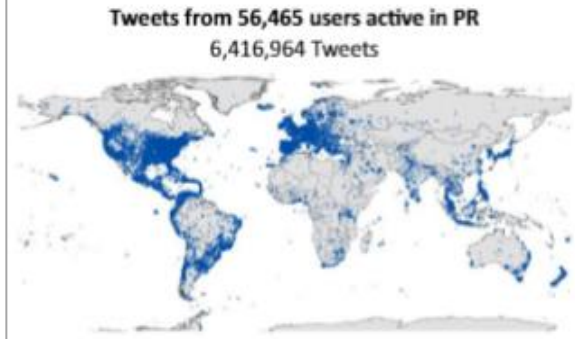
Gauge hurricane evacuation compliance



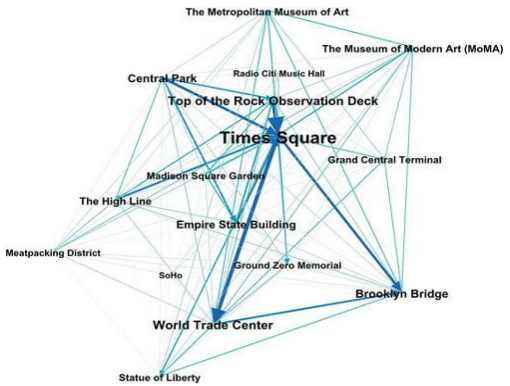
Understand evacuation decision making



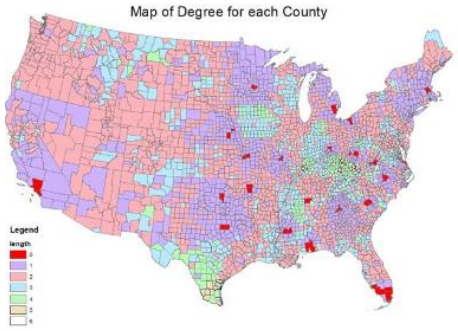
Population migration caused by disasters



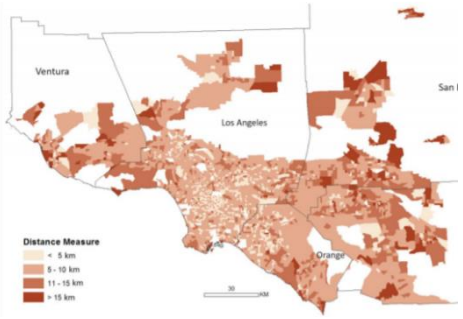
Study tourist movement patterns



Measure inter-city spatial networks



Measure and model activity space



Human mobility during disruptive events

Find out more at http://gis.cas.sc.edu/gibd/research

How do we handle the big data?

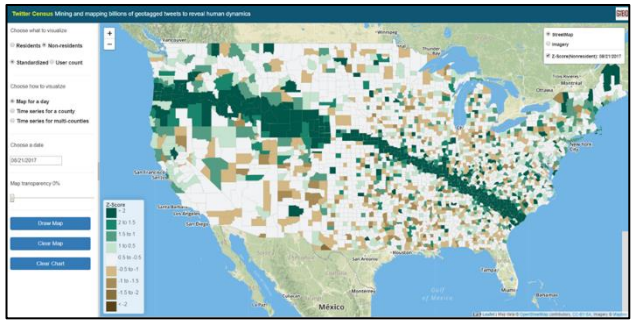
Big Data Computing Platform

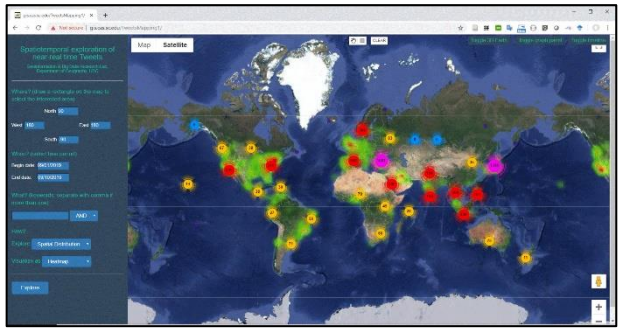
Streaming, storage, management, processing of billions of tweets (and other big data)

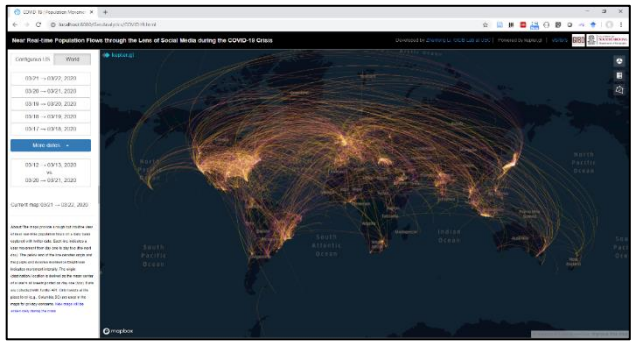
- **Hadoop Distributed File System**
  - Scalable
  - Fault tolerance
- **Spatiotemporal indexing**
  - Spatial index: Quadtree, R-tree
  - Temporal index: Year, month, day, hour
- **Parallel query processing**
  - Impala (in-memory, interactive)
  - Spark (Spark ML)
  - Hive (MapReduce, spatial analysis)
  - jupyter notebook

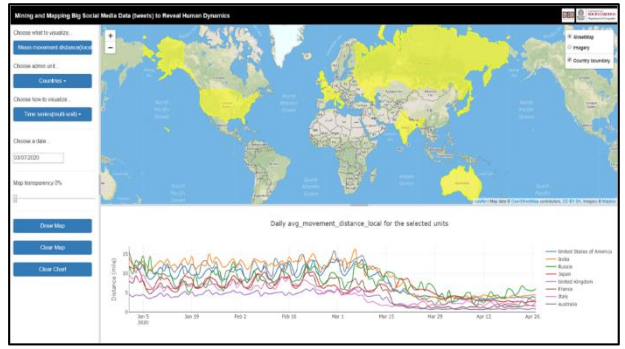
Geovisual Analytics Tools

Web-based query and visual analytical tools for exploratory analysis



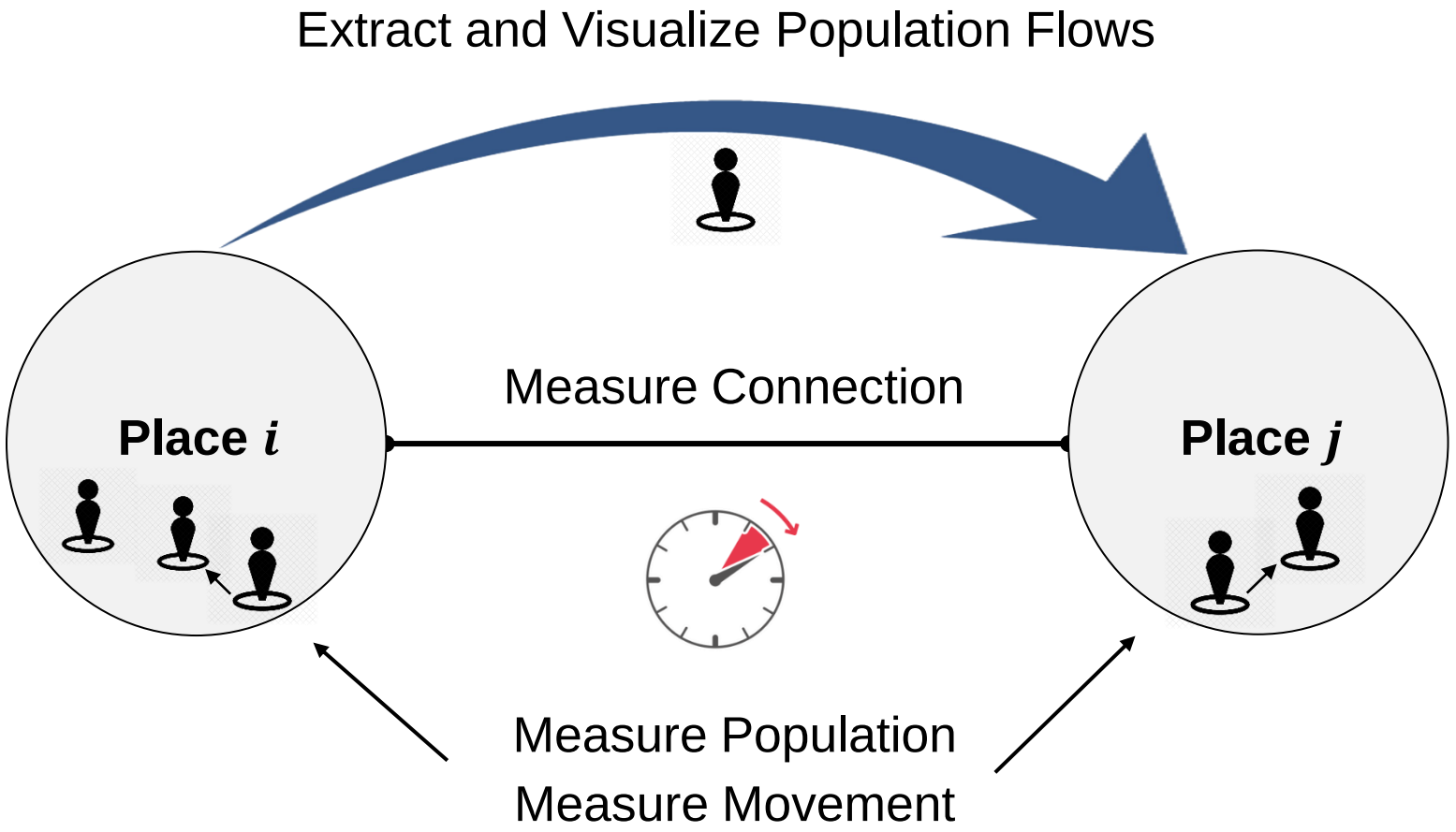






Computing Cluster

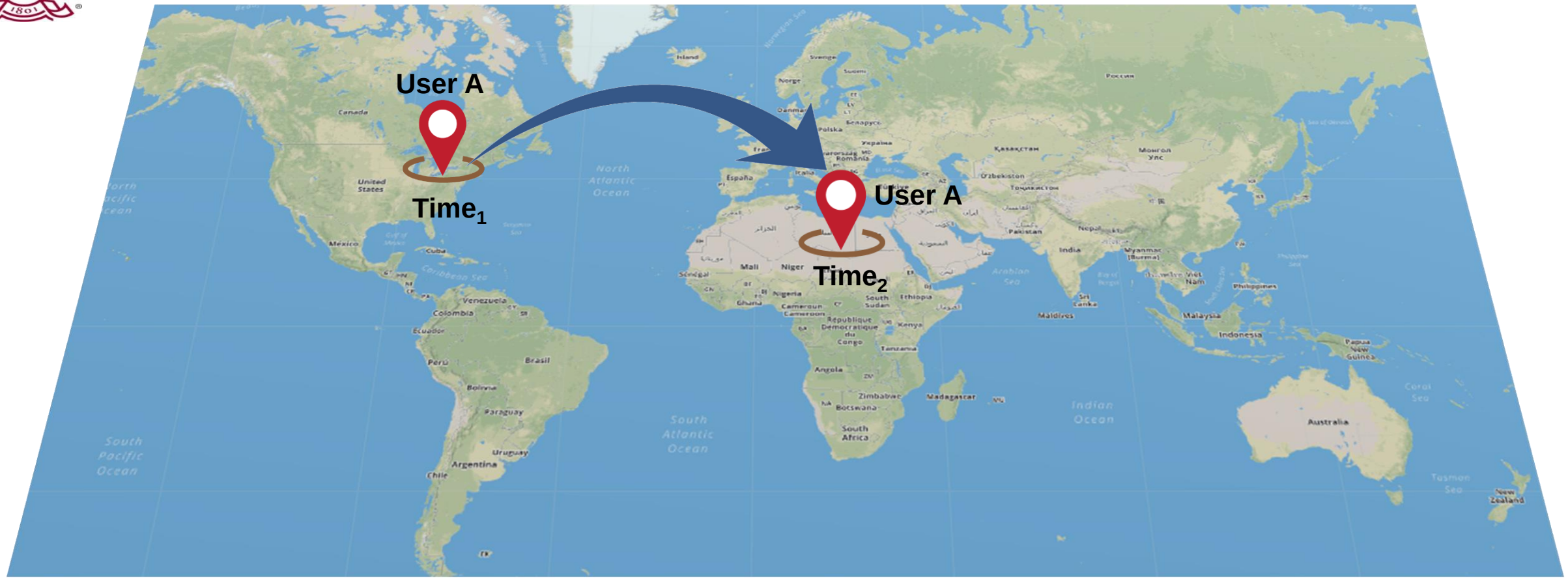
What human mobility information can we extract from social media data for COVID-19?



Place can be: **community, city/county, state, country, or grid**

Time period can be: **hour, day, week, month, year**

Extract daily population flows from millions of twitter users

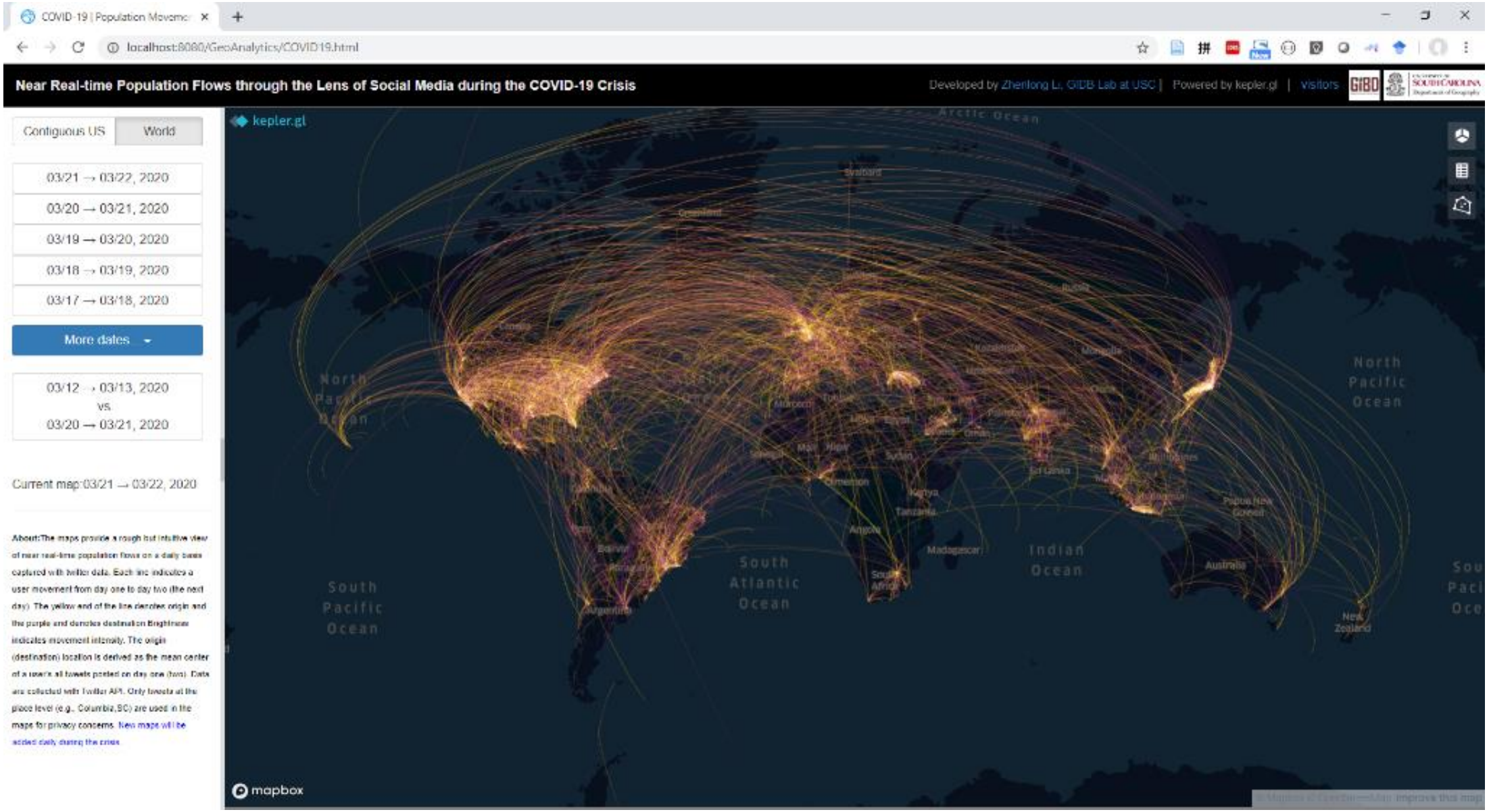


We have about **1.2 million unique twitter users** on a daily basis (excluding the non-human twitter accounts).

The changes of daily population flows in March 2020

Outgoing population flows from Italy to other parts of the world between March 1 and March 11, 2020

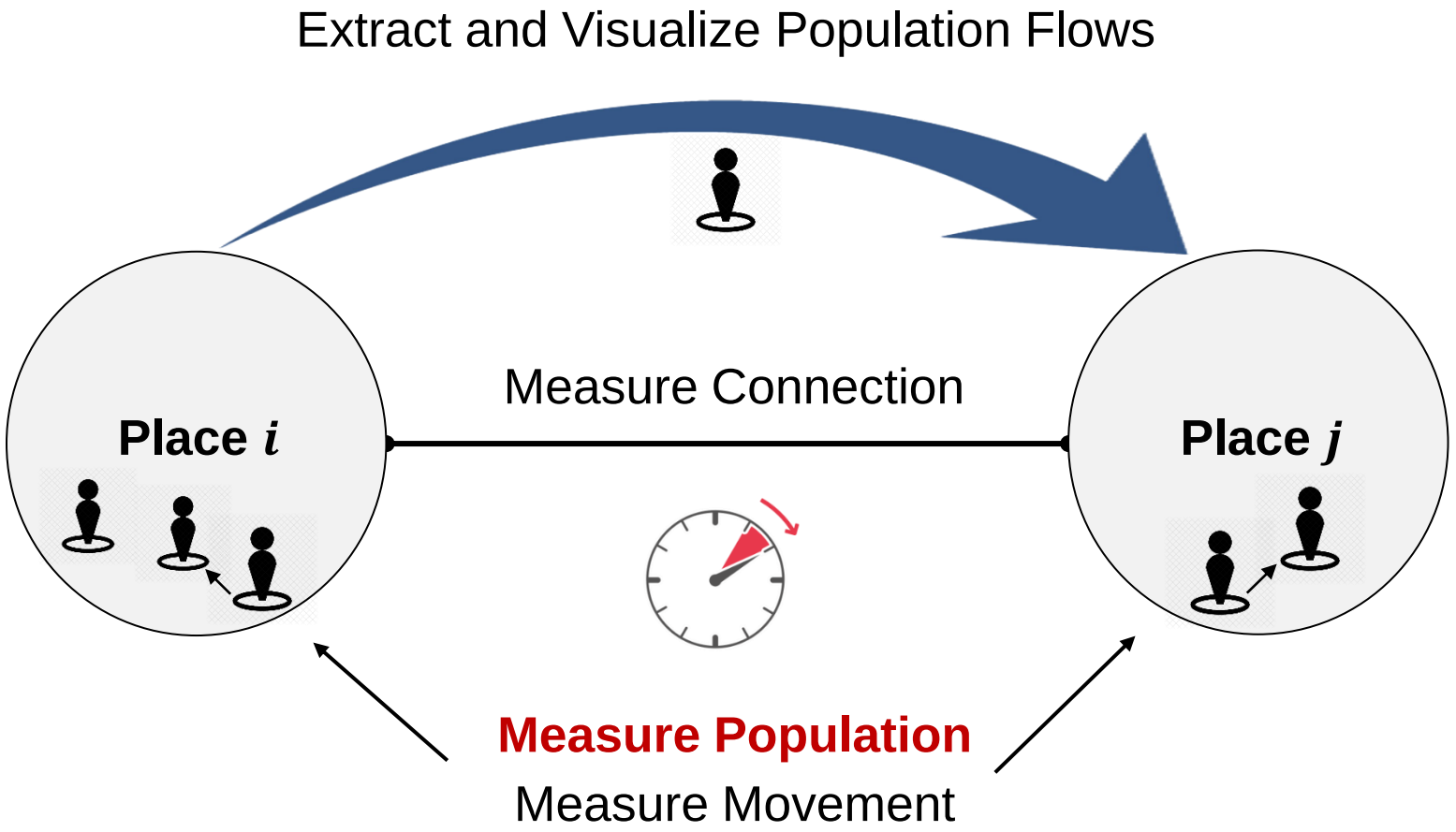
Interactive web portal for population flow visualization



DEMO

http://gis.cas.sc.edu/GeoAnalytics/COVID19.html

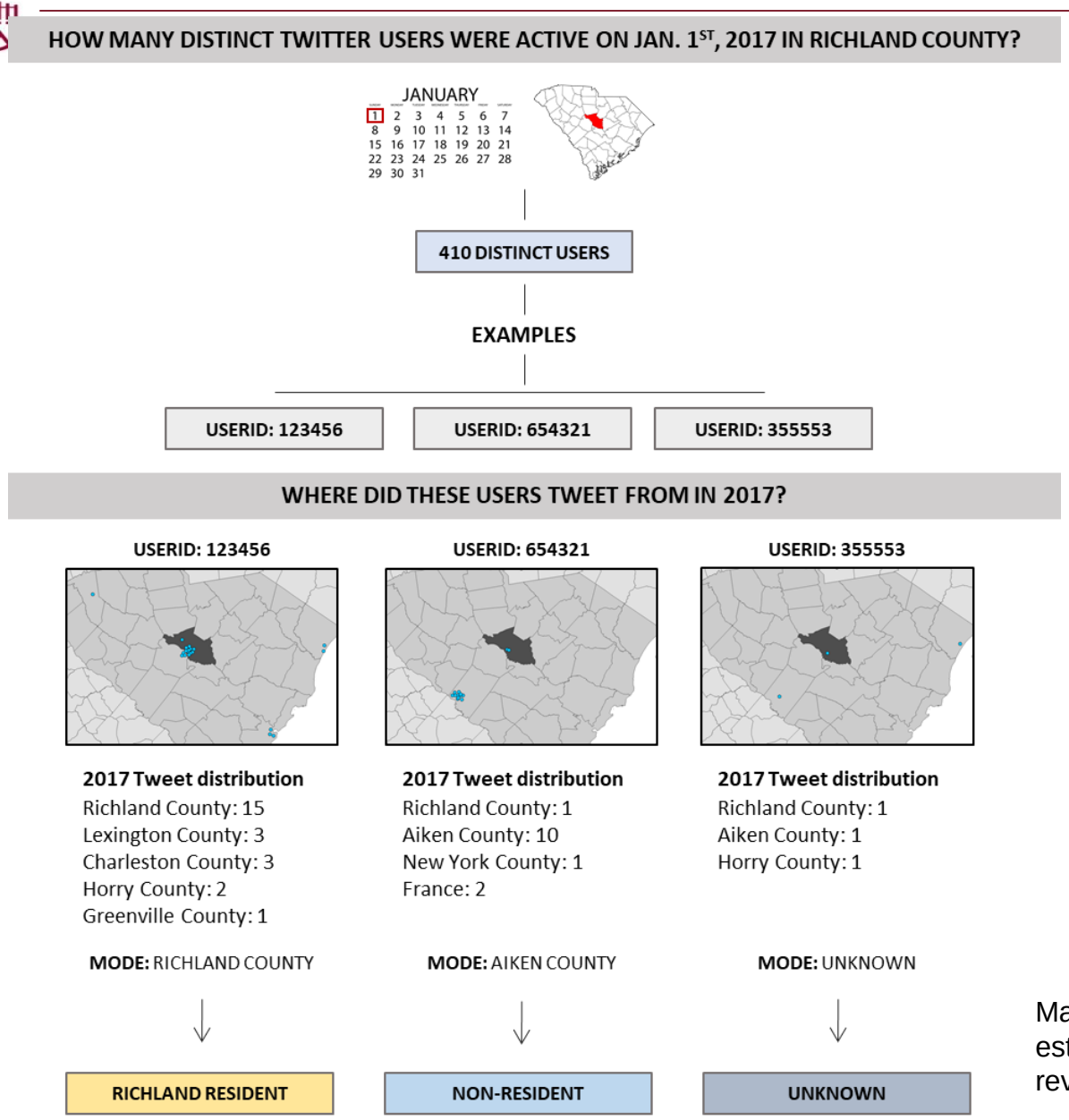
What human mobility information can we extract from social media data for COVID-19?



Place can be: **community, city/county, state, country**

Time period can be: **hour, day, week, month, year**

Measure Population: visitors and residents



We developed a method to estimate the daily number of **visitors (VN)** and **residents (RN)** at a specific geographic scale (e.g., county level) using geotagged tweets.

We then defined two indices (for :
Visitor Index (**VI**)
Resident Index (**RI**)

Daily z-score of each county based on one-year period (365 days).

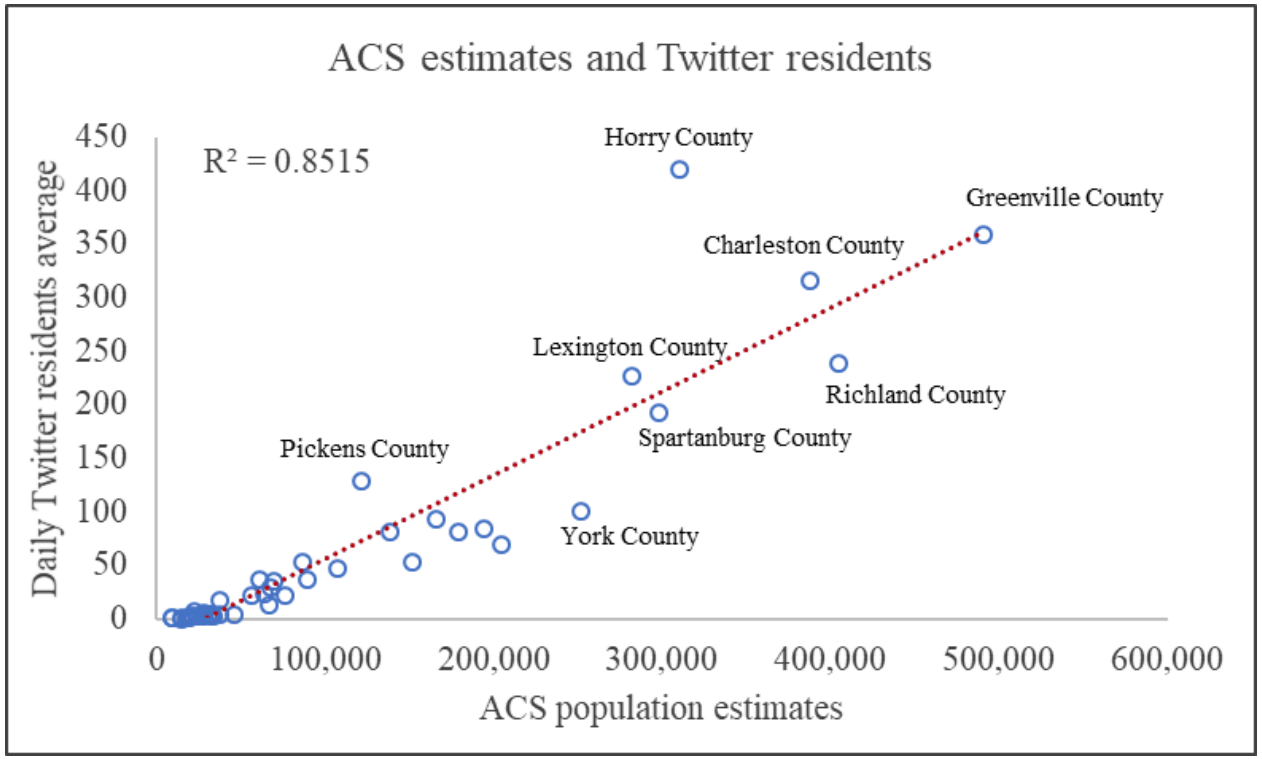
$$z = \frac{x - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation

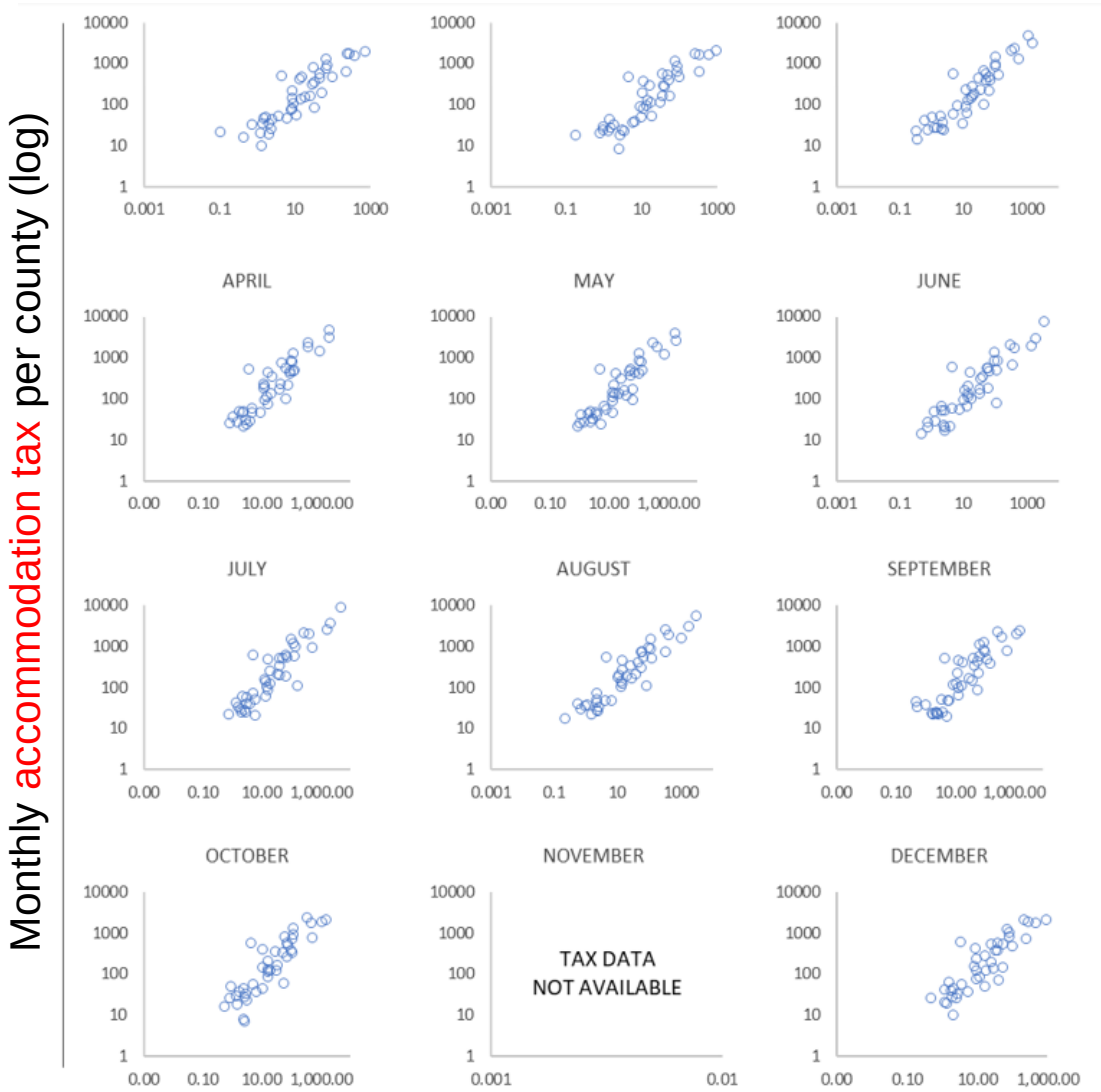
Martin Y., Li Z., Ge Y., Towards real-time population estimates: introducing Twitter daily estimates of residents and non-residents at the county level, *Applied Geography* (in review)

Measure Population: visitors and residents

Validation our approach with South Carolina data



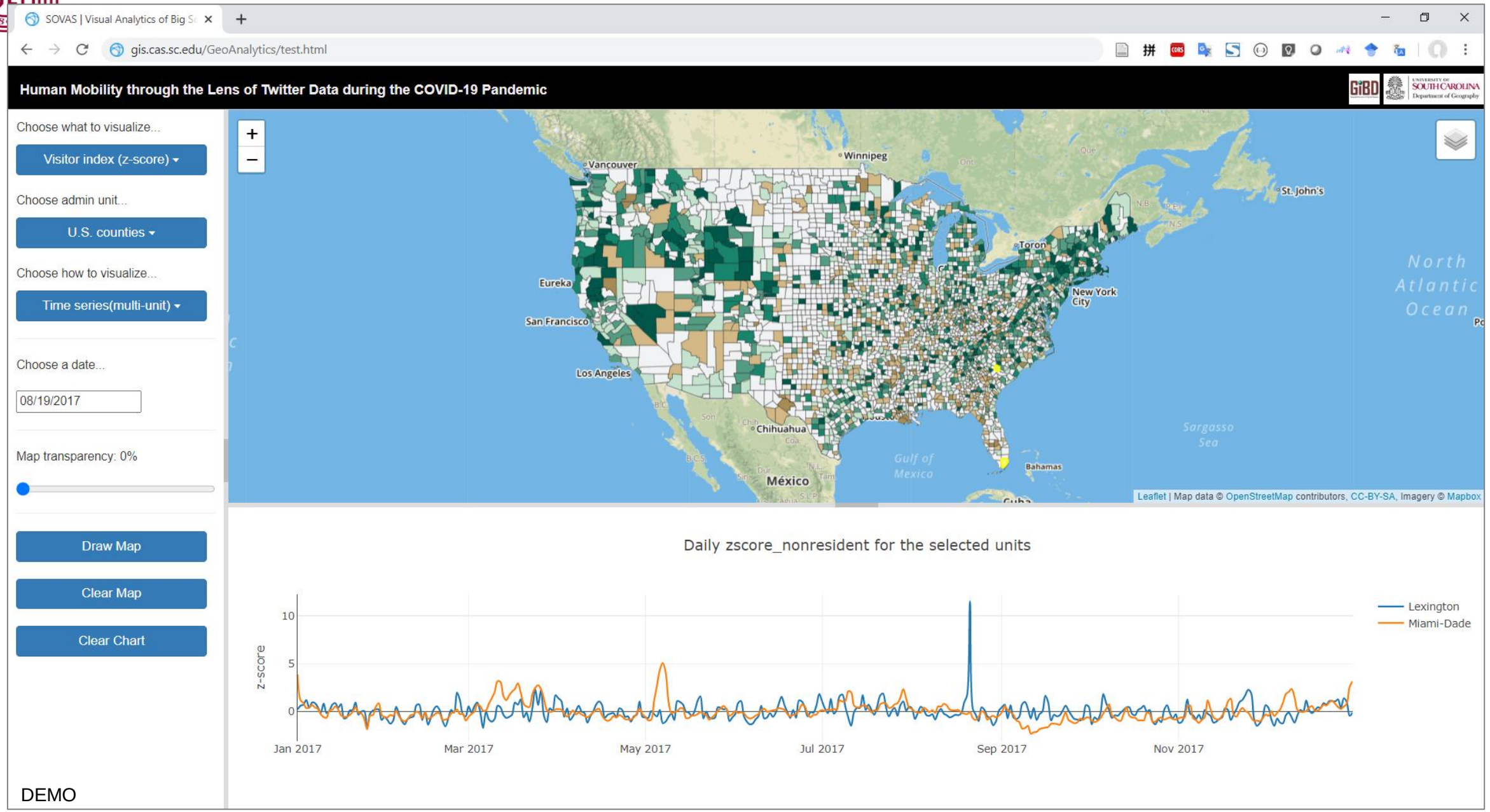
Scatter plot of 5-year ACS county population estimates and the daily average of total active Twitter residents per county (46 counties in total).



Monthly non-resident Twitter users (visitors) per county (log)

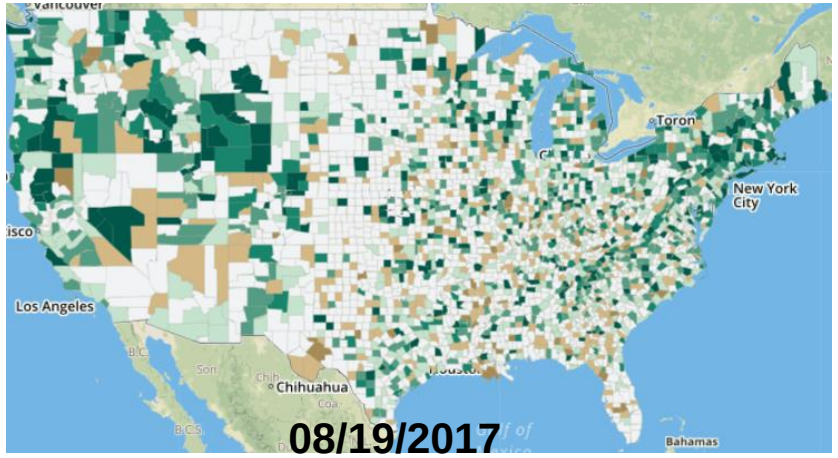
Martin Y., Li Z., Ge Y., Towards real-time population estimates: introducing Twitter daily estimates of residents and non-residents at the county level, *Applied Geography* (in review)

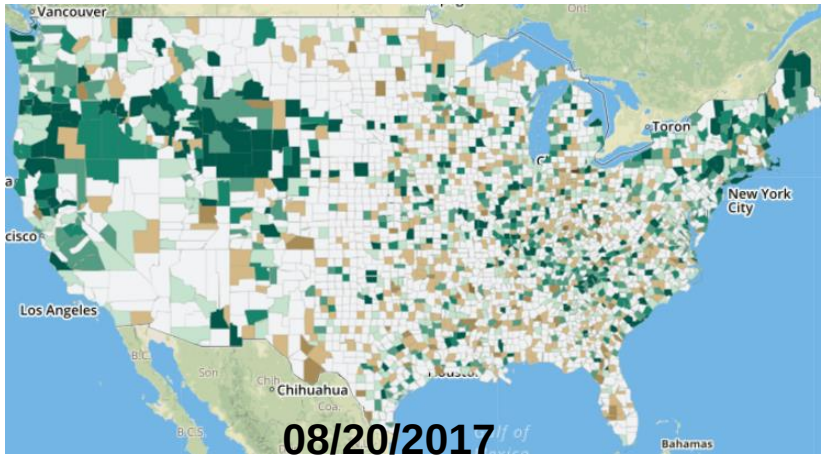
Measure Population: explore the spatiotemporal patterns

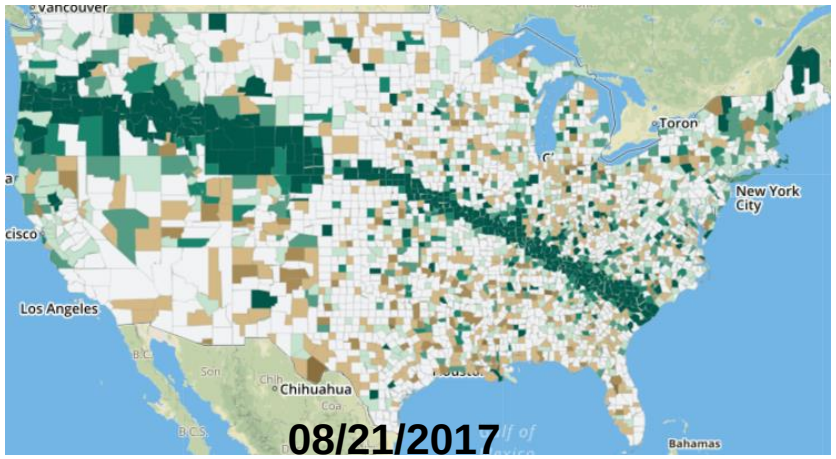


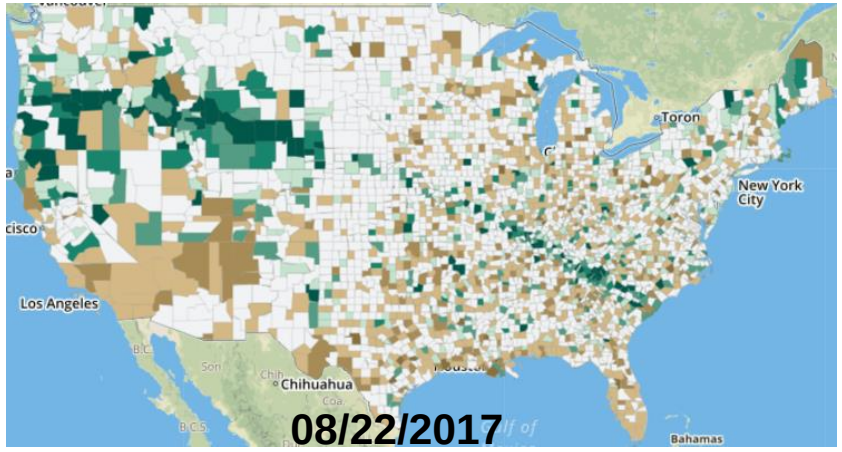
DEMO

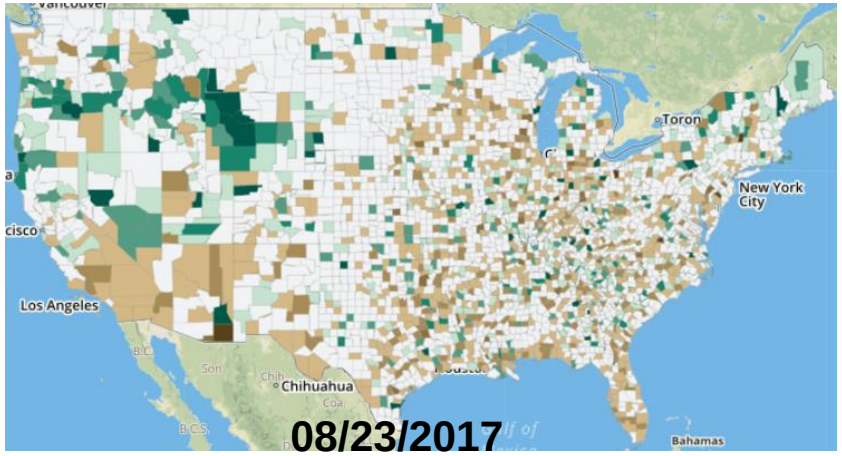
Measure Population: explore the spatiotemporal patterns

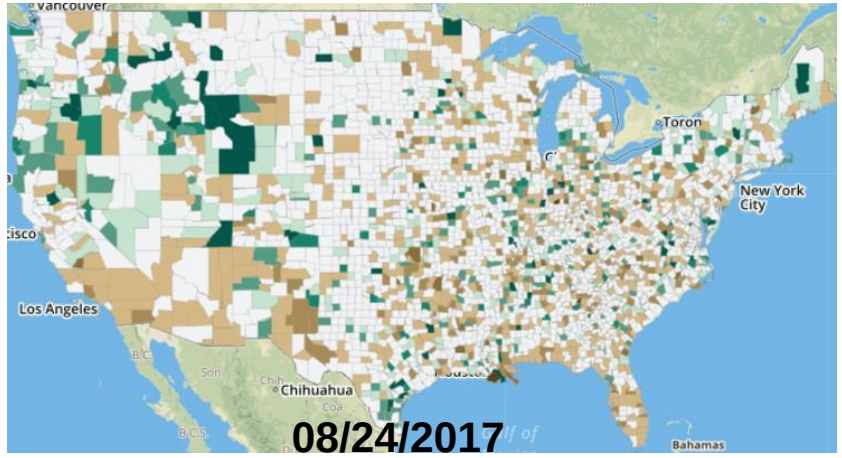


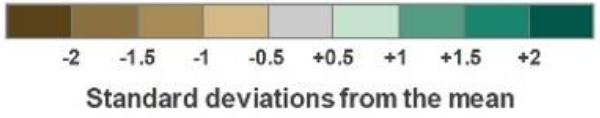






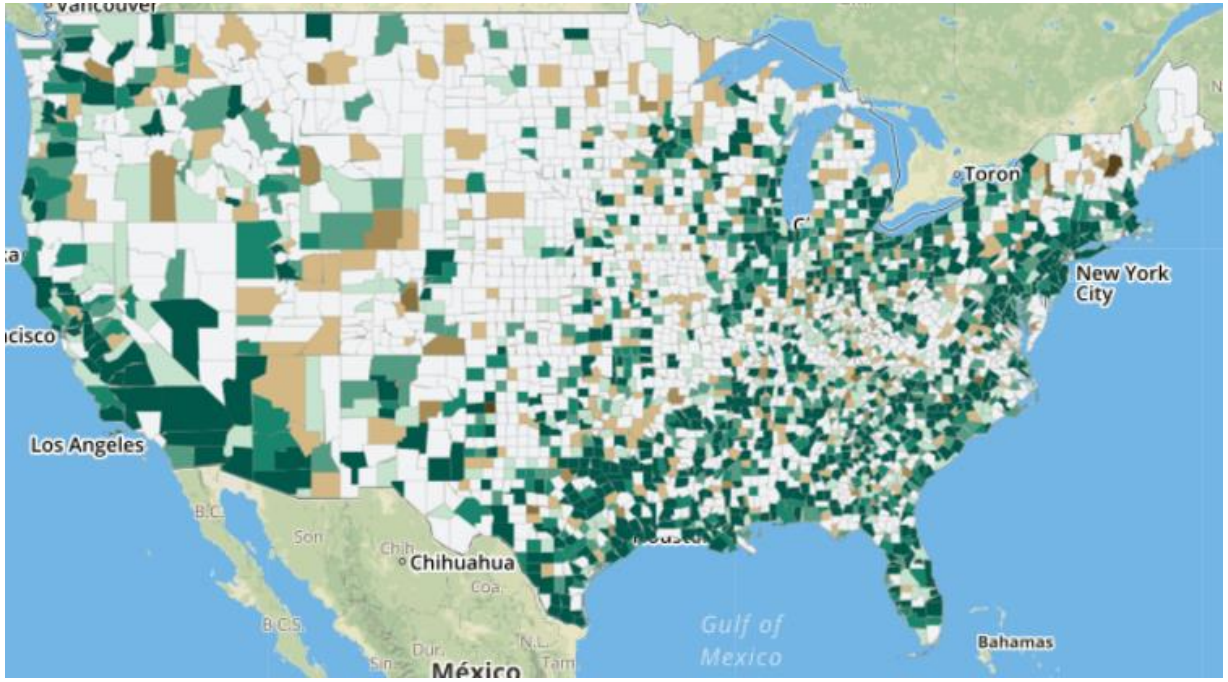




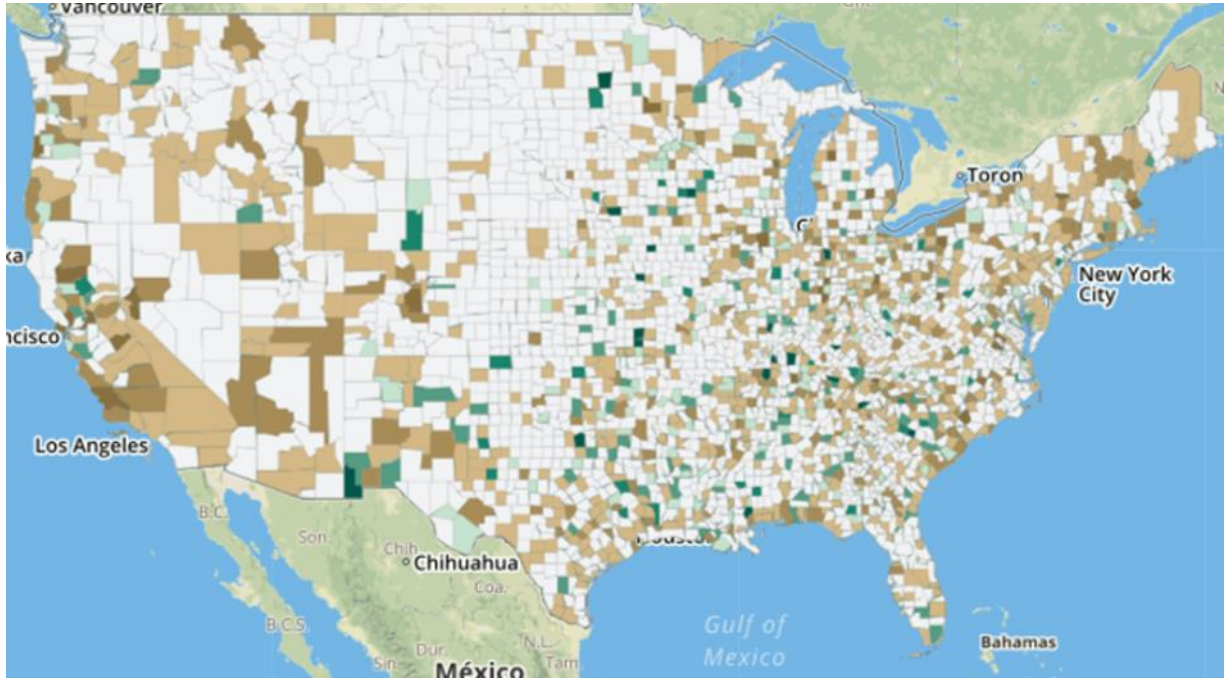


County level visitor changes before, during, and after the 2017 total solar eclipse.

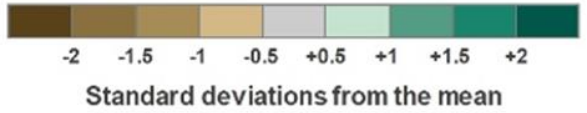
Measure Population: explore the spatiotemporal patterns



Thanksgiving Day, 11/28/2019

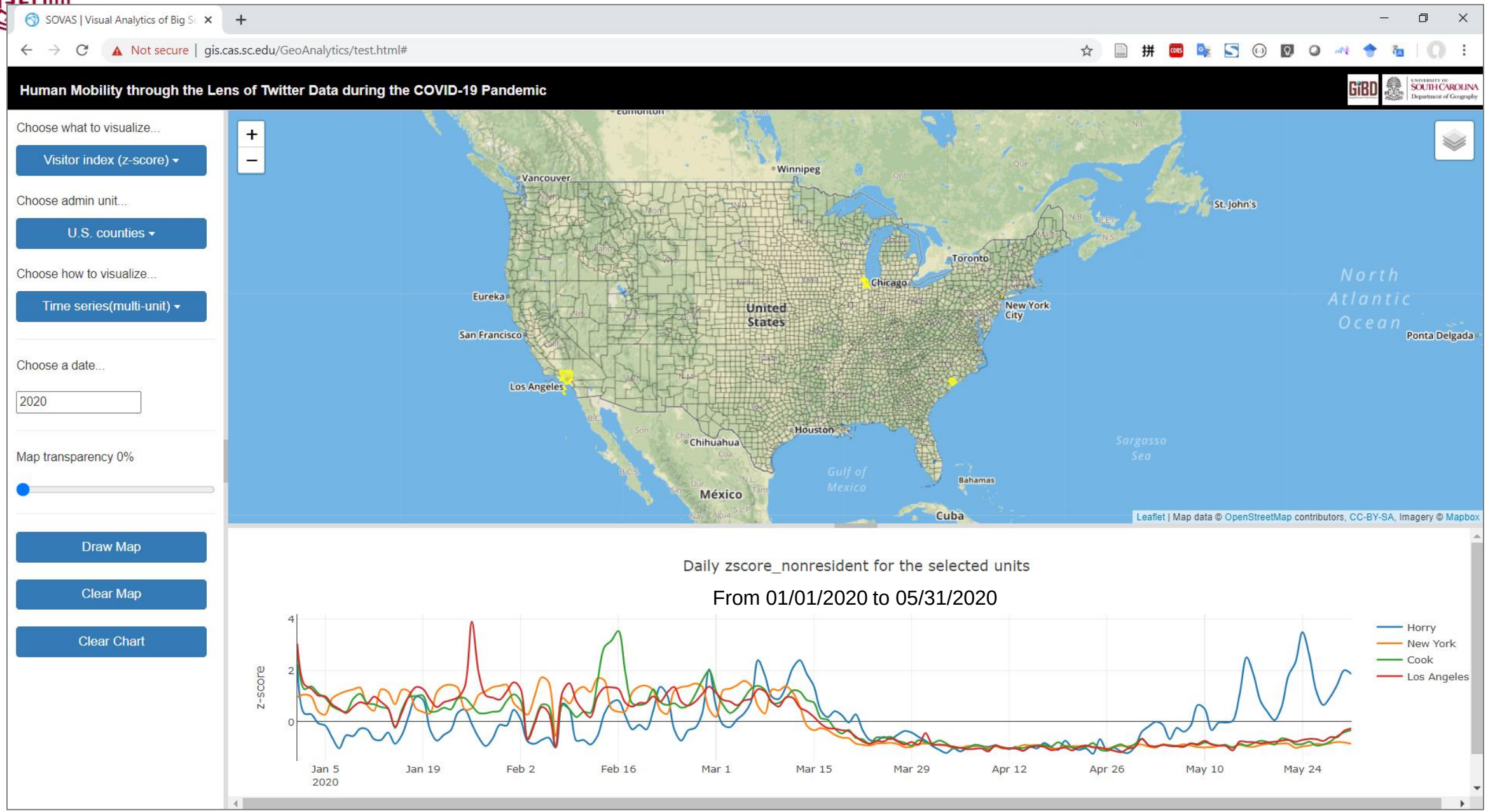


A week after Thanksgiving, 12/05/2019

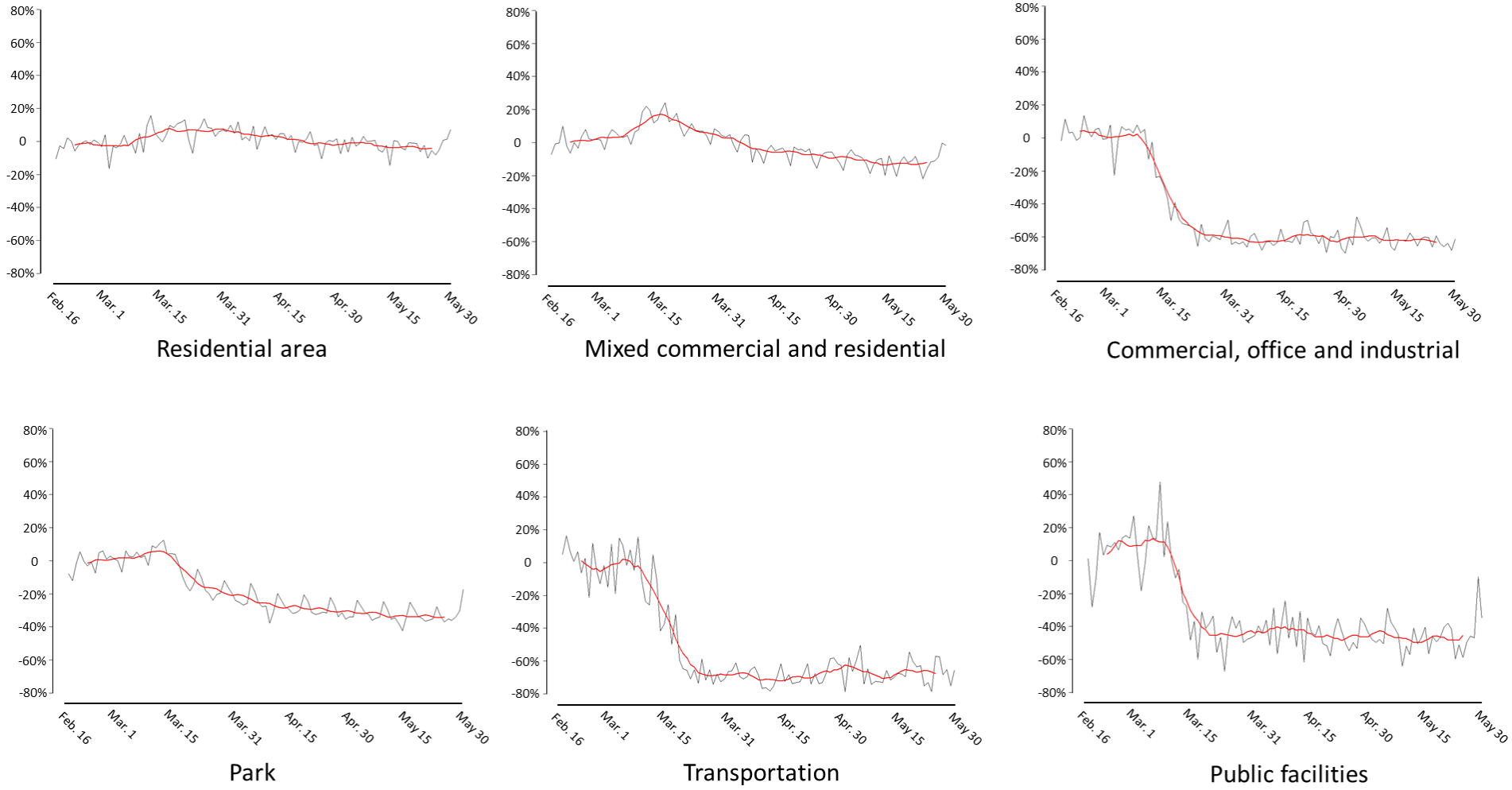


Results indicate that VI index at the county level can accurately reflect regular (seasonal tourism, such as holidays) and non-regular events (such as the 2017 total solar eclipse).

Measure Population: explore the spatiotemporal patterns



Measure Population: explore the spatiotemporal patterns



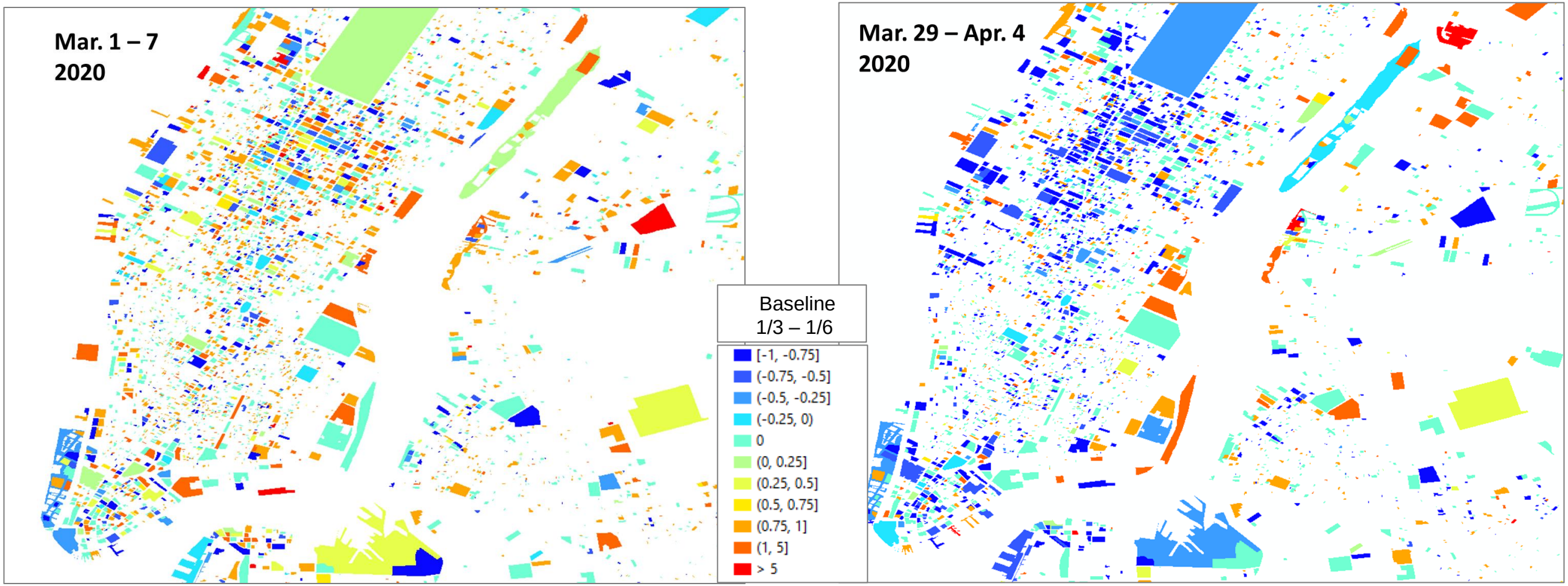
NYC number of Twitter user changes for different land use types from 02/16 to 05/31, 2020

Baseline
Jan. 3 – Feb. 6

Jiang Y., Li Z., Huang X., (2020) Spatiotemporal patterns of human mobility during COVID-19 in New York City using Twitter data, (in preparation)

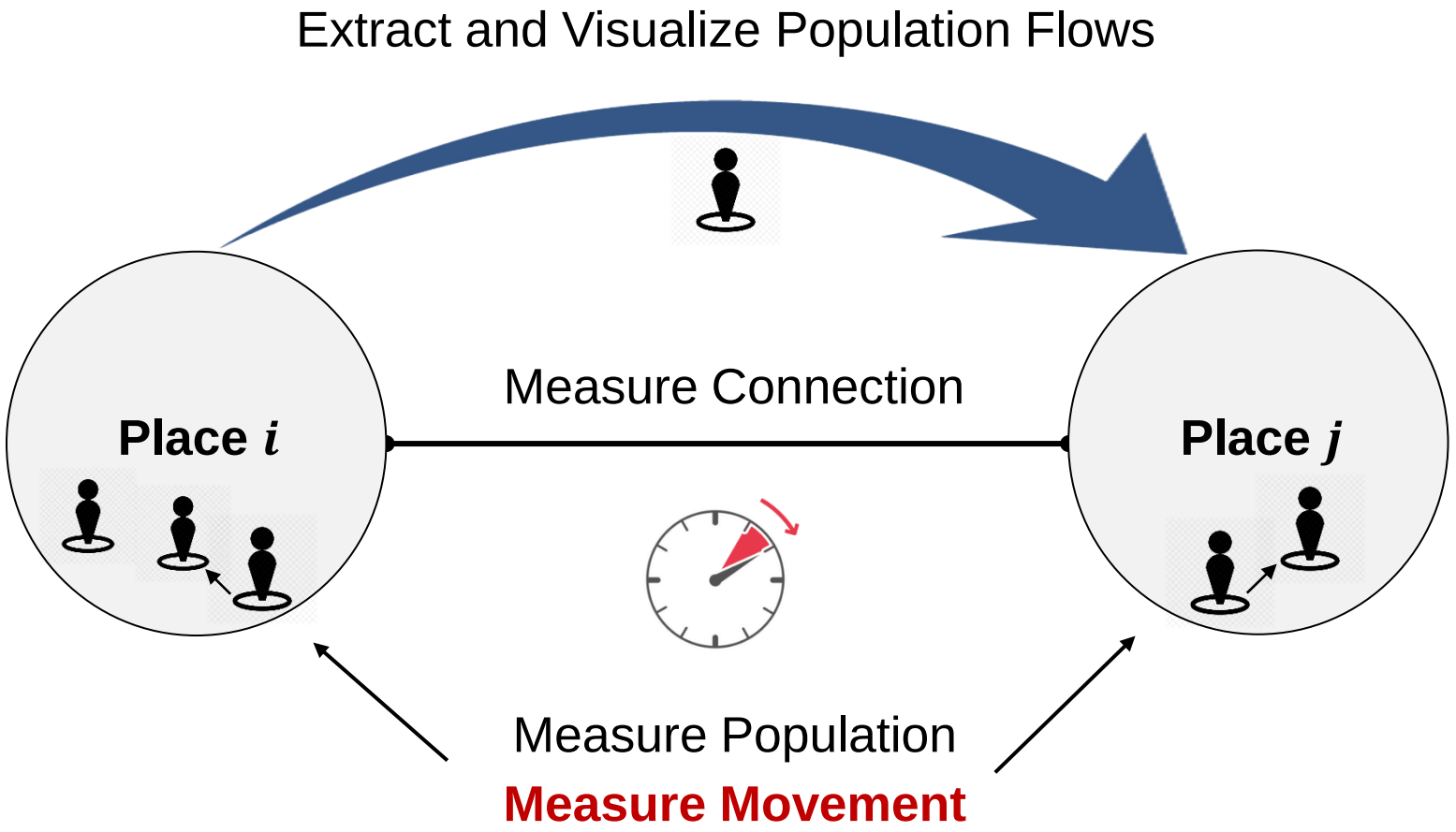
Measure Population: explore the spatiotemporal patterns

Population changes captured with twitter data at the tax lot level of NYC (PLUTO data)



Jiang Y., Li Z., Huang X., (2020) Spatiotemporal patterns of human mobility during COVID-19 in New York City using Twitter data, (in preparation)

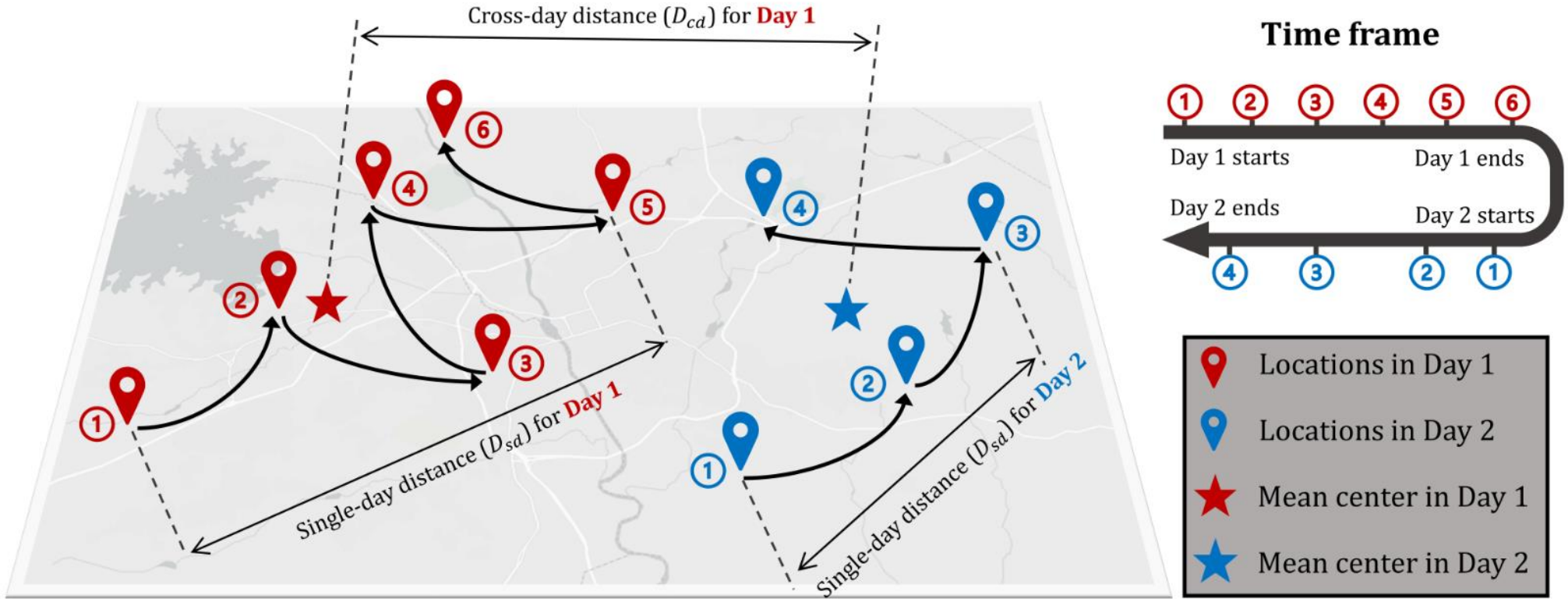
What human mobility information can we extract from social media data for COVID-19?



Place can be: **community, city/county, state, country**

Time period can be: **hour, day, week, month, year**

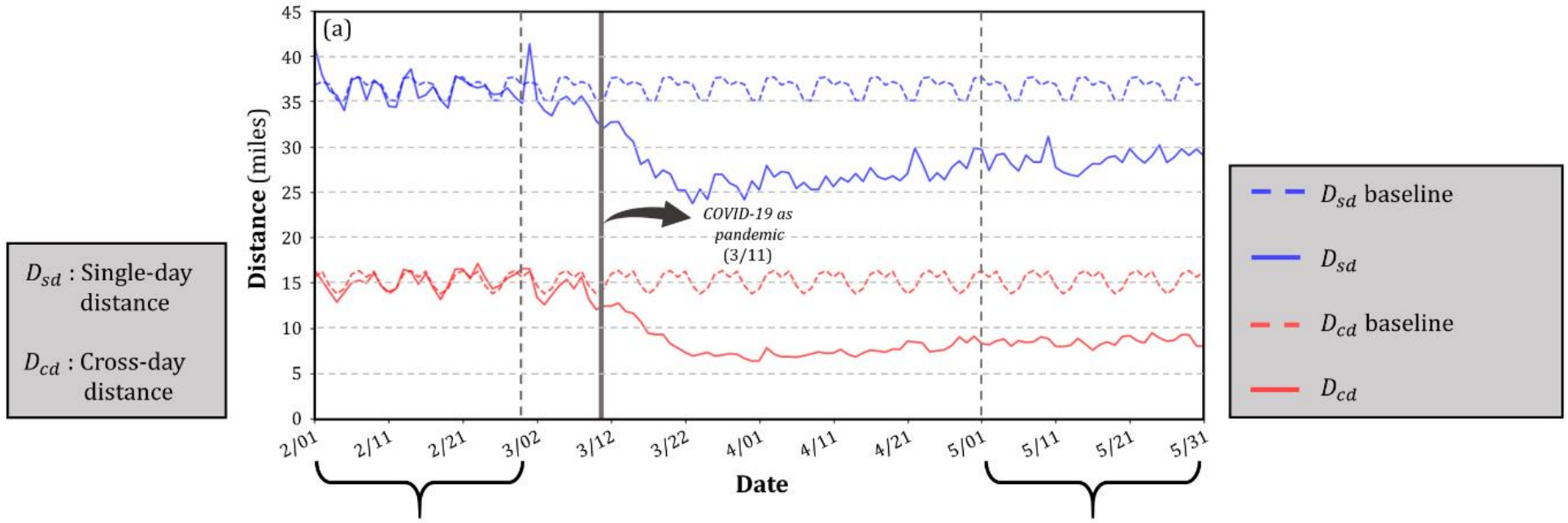
Measure Movement: daily travel distance (or area)



We developed two travel distance measurements: single day distance and cross-day distance.

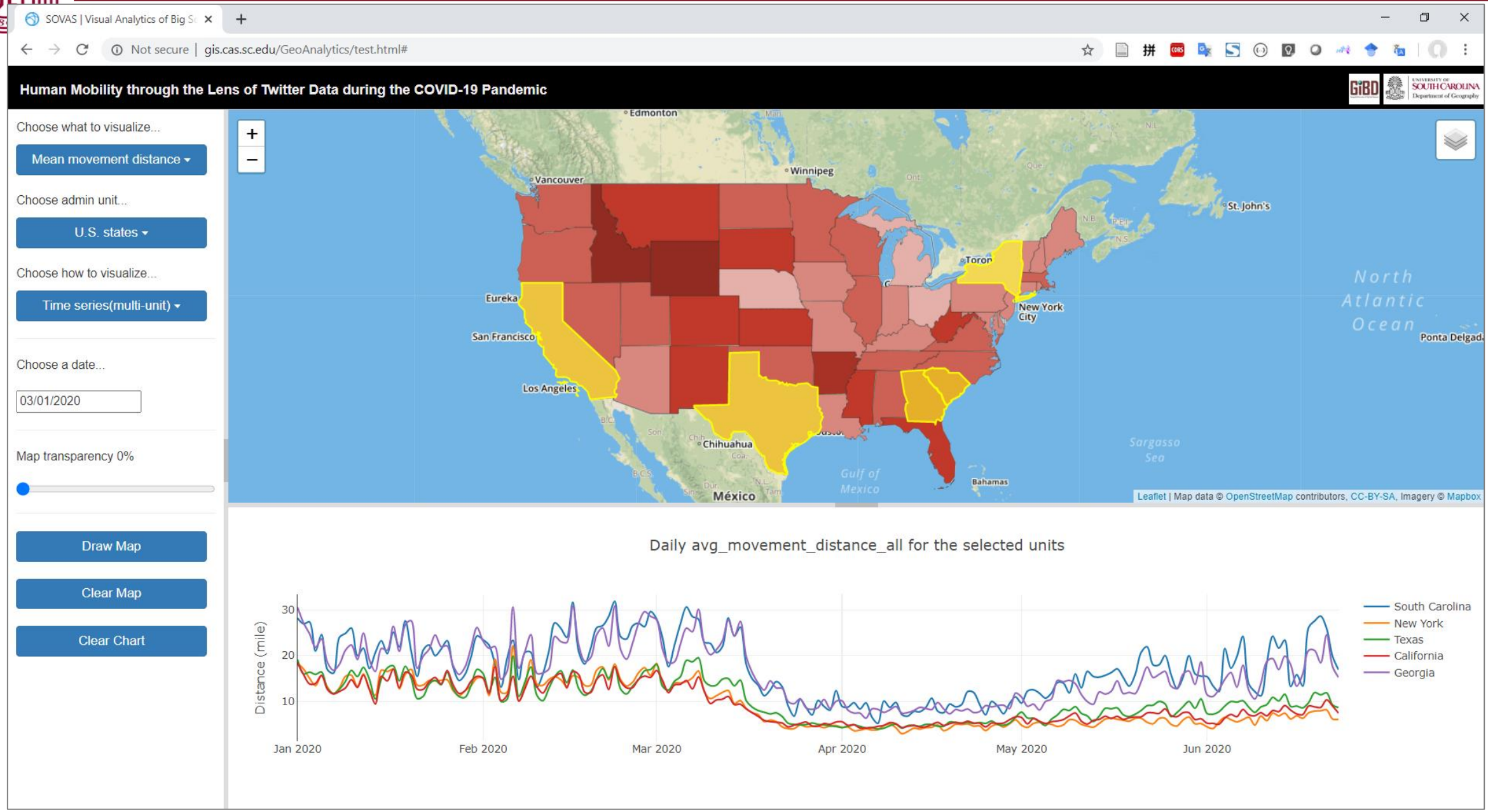
Huang X., Li Z., Jiang Y., Li X., Porter D., (2020) Twitter, human mobility, and COVID-19, *PlosONE* (in review), preprint at: http://shorturl.at/ehp03

Measure Movement: changes of world travel distance



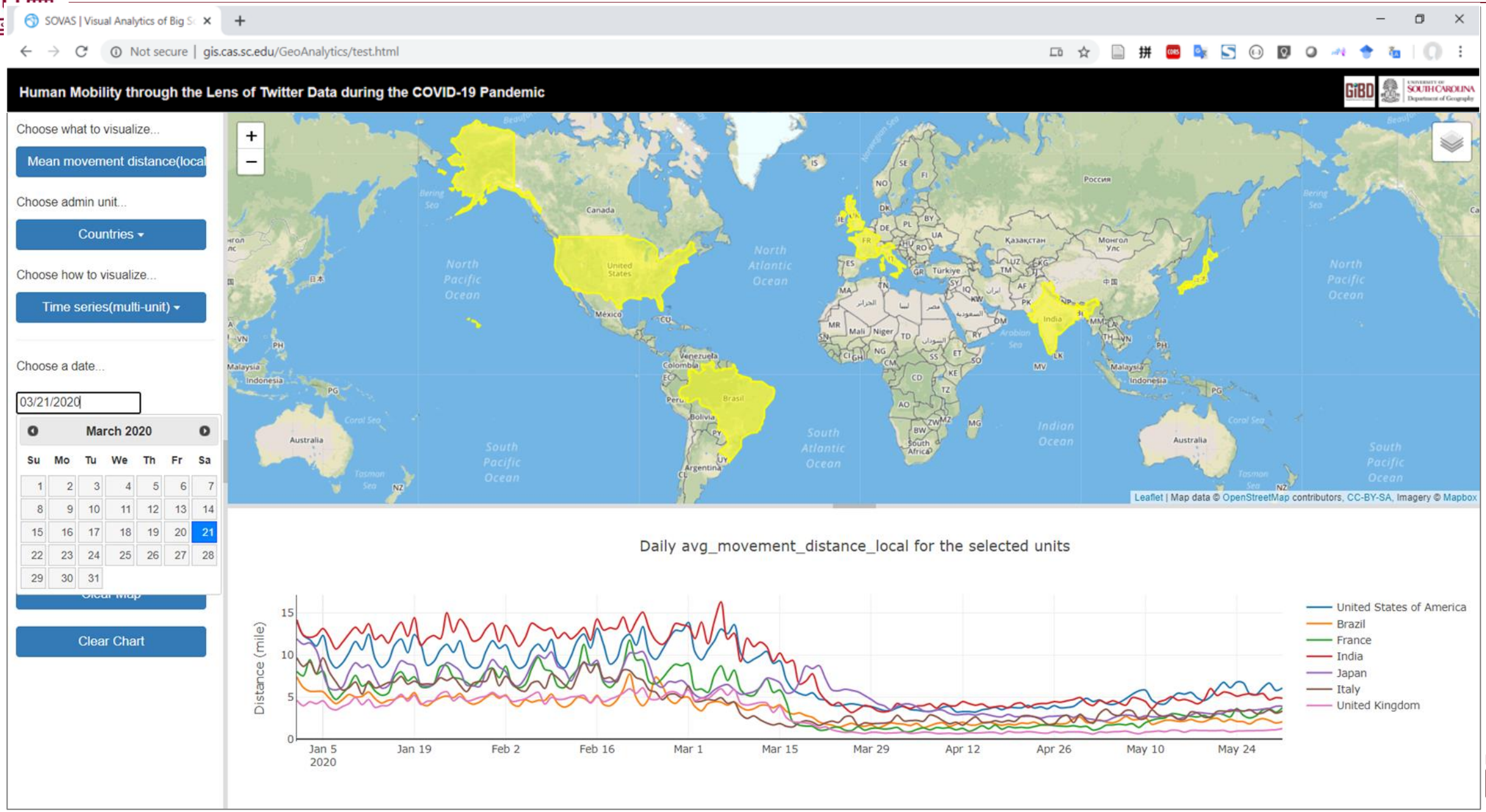
Huang X., Li Z., Jiang Y., Li X., Porter D., (2020) Twitter, human mobility, and COVID-19, *PlosONE* (in review), preprint at: http://shorturl.at/ehp03

Measure Movement: explore the spatiotemporal patterns

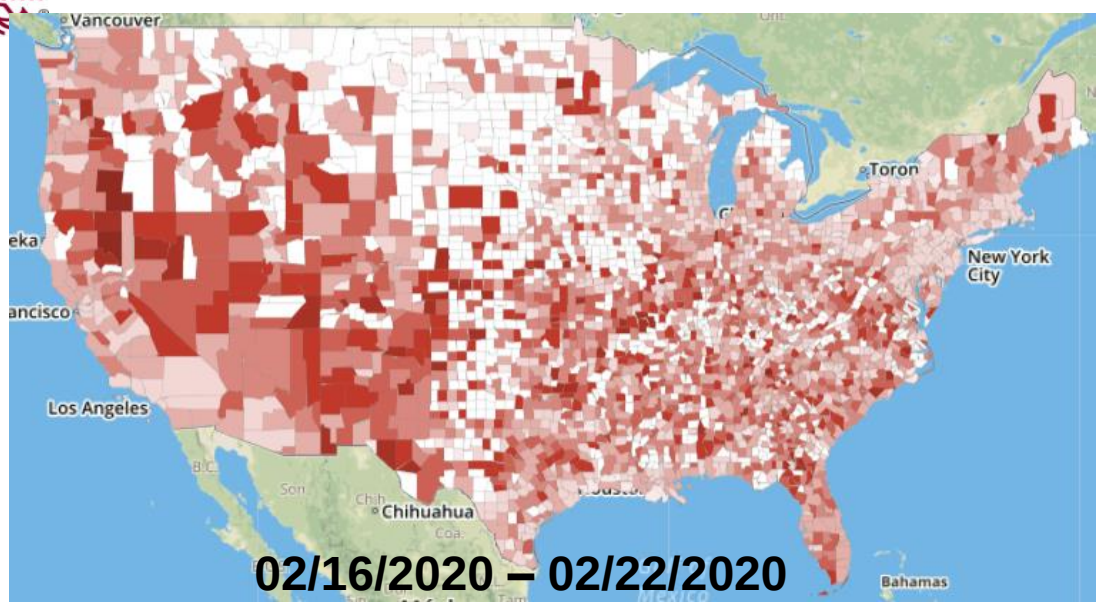


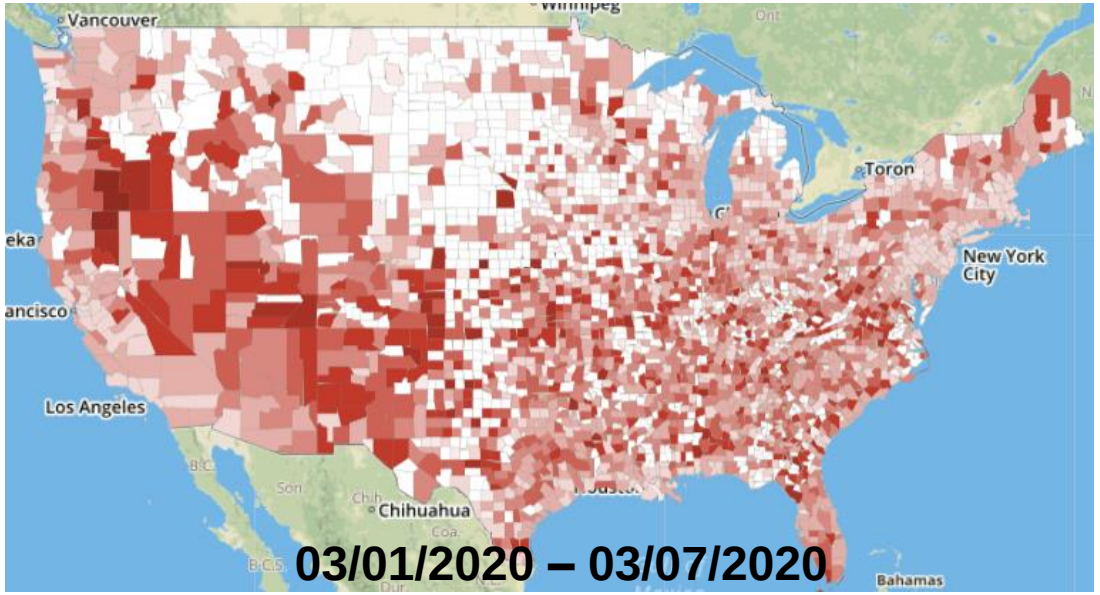
DEMO

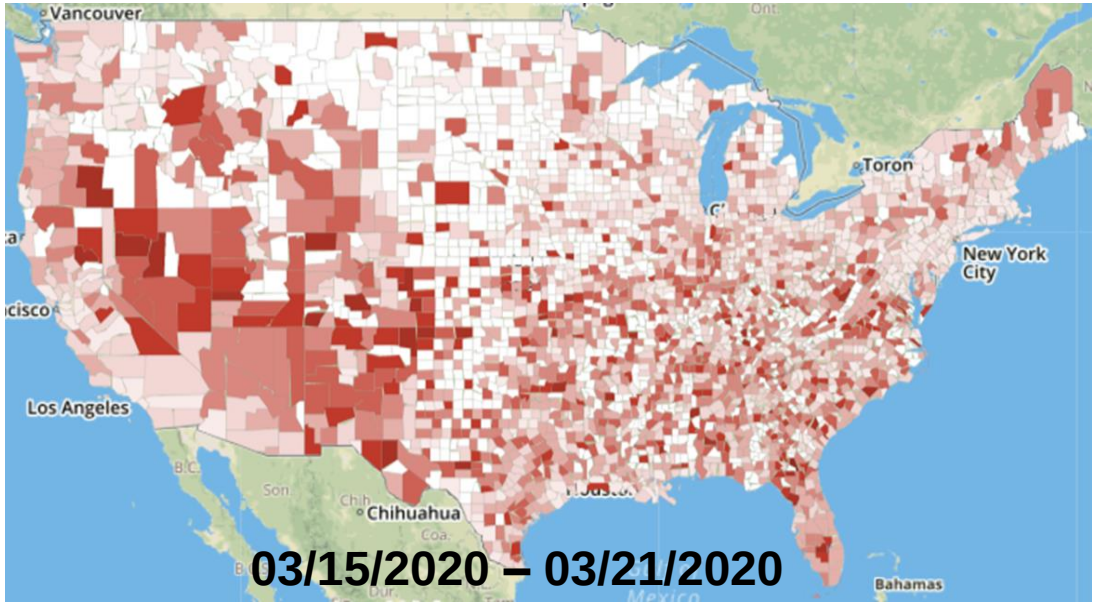
Measure Movement: explore the spatiotemporal patterns

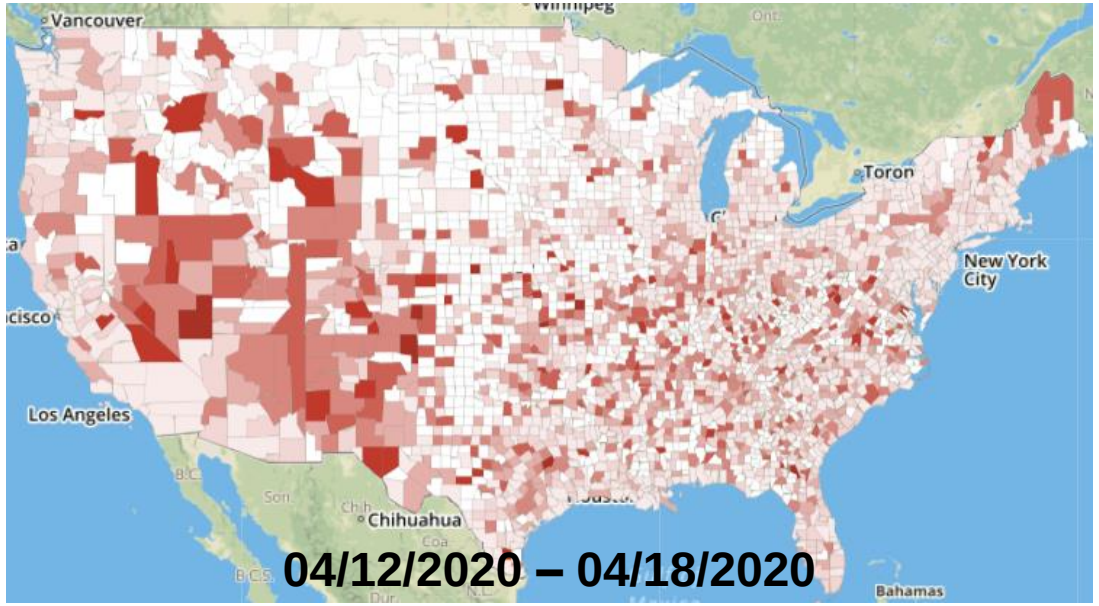


Measure Movement: explore the spatiotemporal patterns

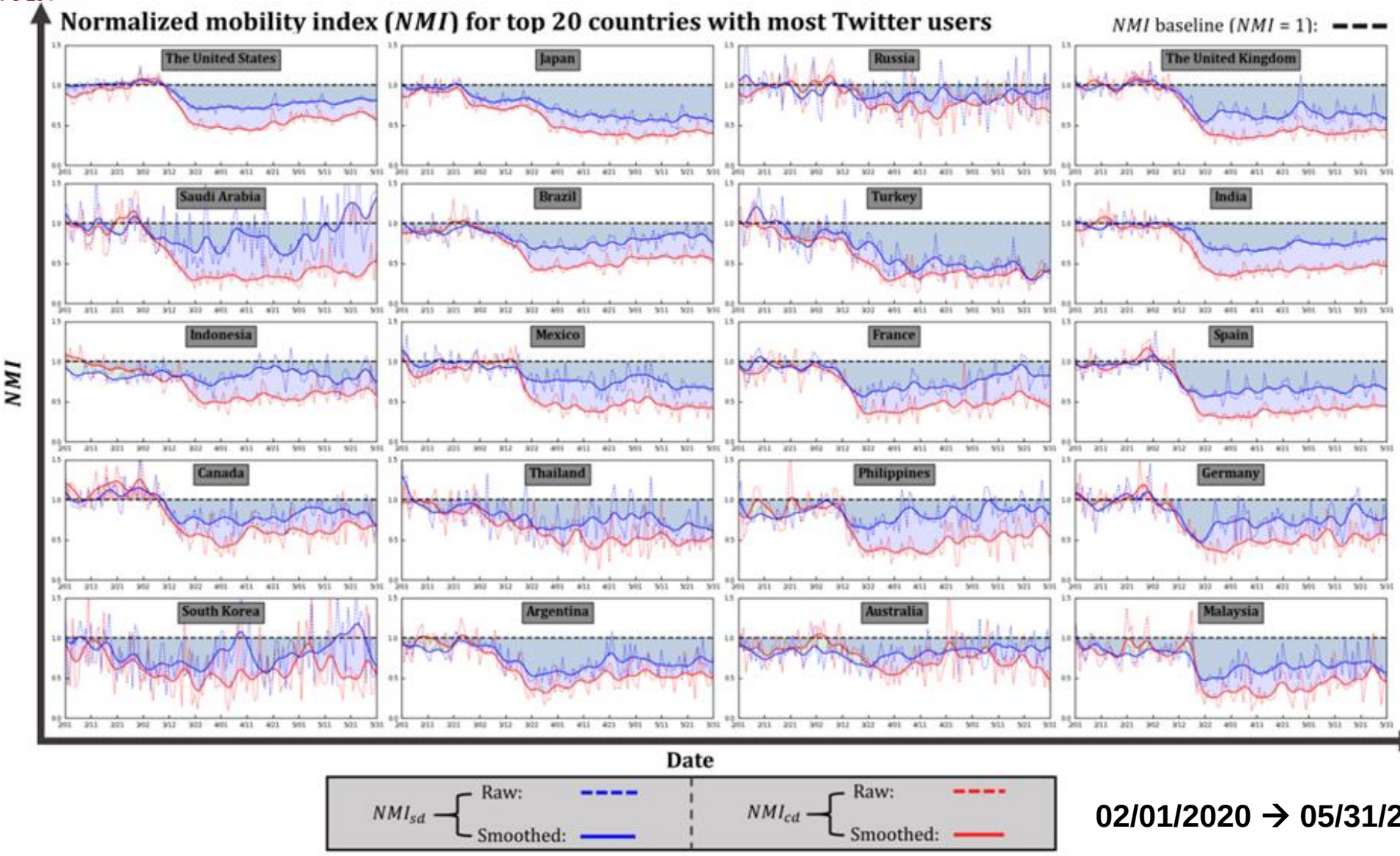








Measure Movement: normalized mobility index (NMI)



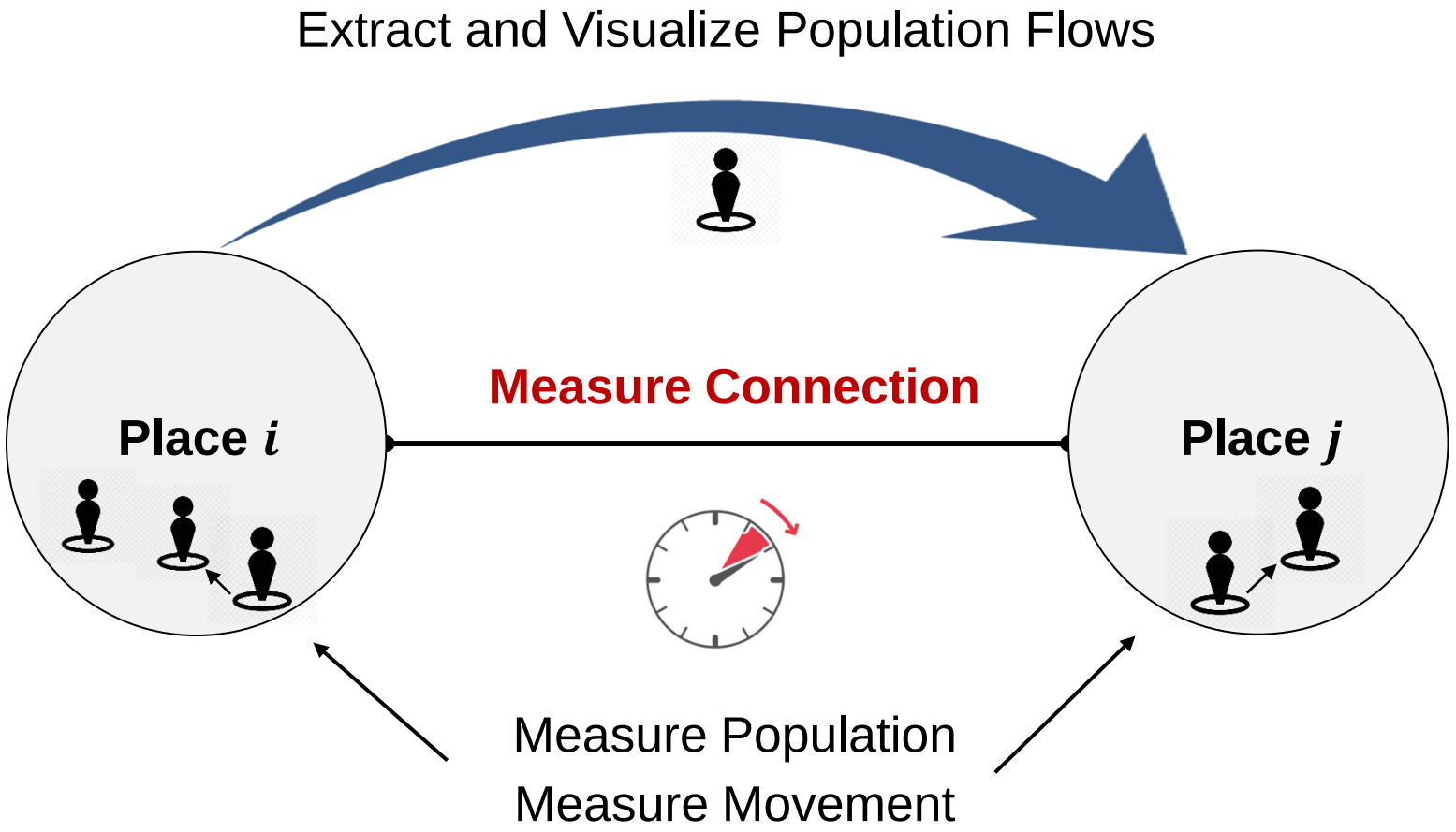
Mobility baselines are set for each corresponding day of a week to generate **NMI**.

NMI can be generated daily at different geographic scales.

02/01/2020 → 05/31/2020

Huang X., Li Z., Jiang Y., Li X., Porter D., (2020) Twitter, human mobility, and COVID-19, *PlosONE* (in review), preprint at: http://shorturl.at/ehp03

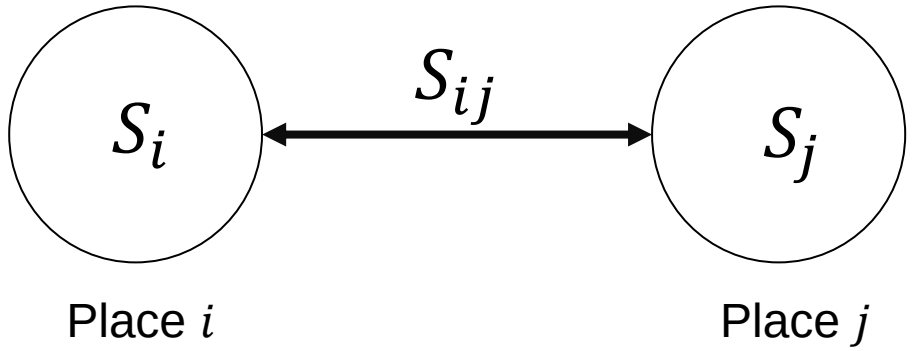
What human mobility information can we extract from social media data for COVID-19?



Place can be: **community, city/county, state, country**

Time period can be: **hour, day, week, month, year**

Measure Connection: place connectedness based on shared users



S_i Number of twitter users in place i

S_j Number of twitter users in place j

S_{ij} Number of shared twitter users between place i and j

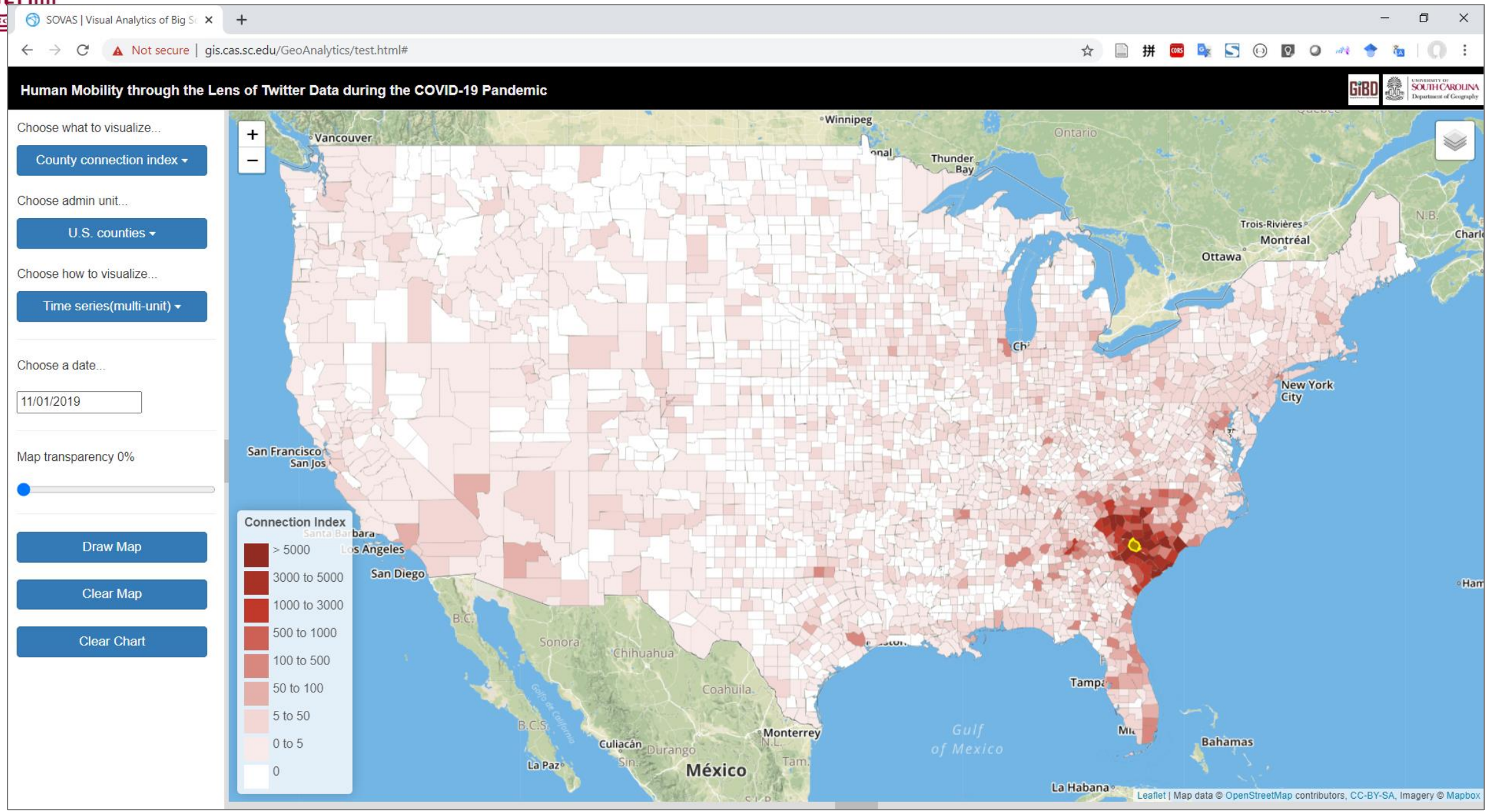
We define the **Connectedness Index (CI)** between two places (e.g., county) for a specified time period (t) as the normalized number of shared users based on the users from each area (**control for population**).

$$CI(t) = \frac{S_{ij}^2(t)}{S_i(t) \cdot S_j(t)}$$

Temporally, the CI can be computed as daily, weekly, monthly, yearly.
Spatially, the CI can be computed in varying spatial scales, e.g., county, state, country.

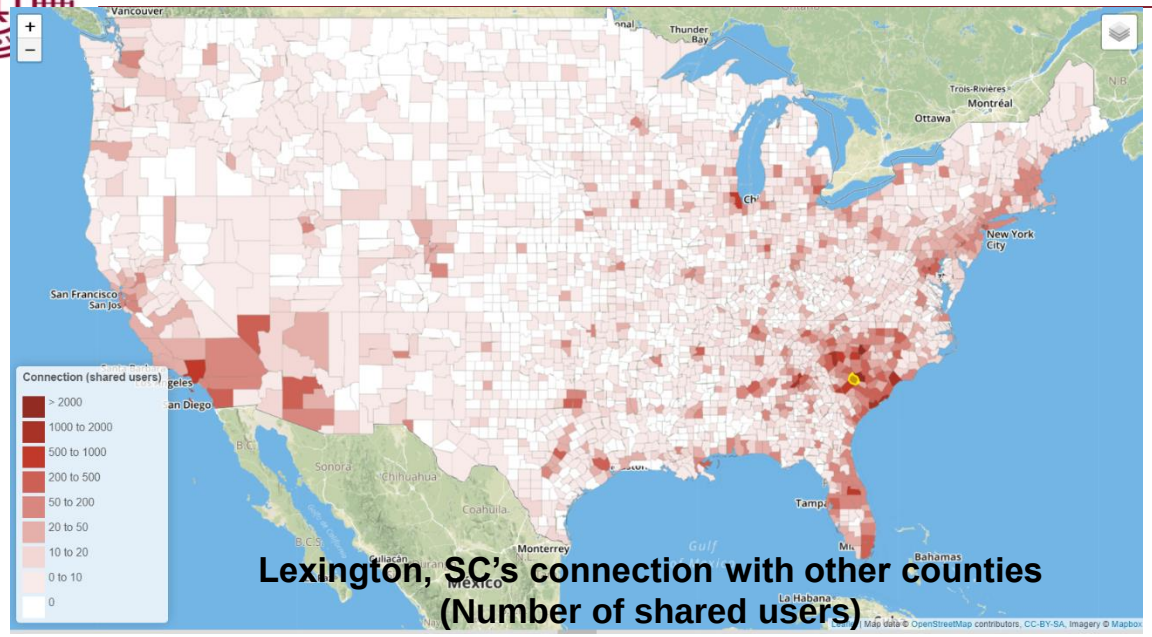
Li, Z., et al., (2020) Measuring Connectedness between Places using Big Social Media Data, (in preparation)

Measure Connection: explore the spatial patterns

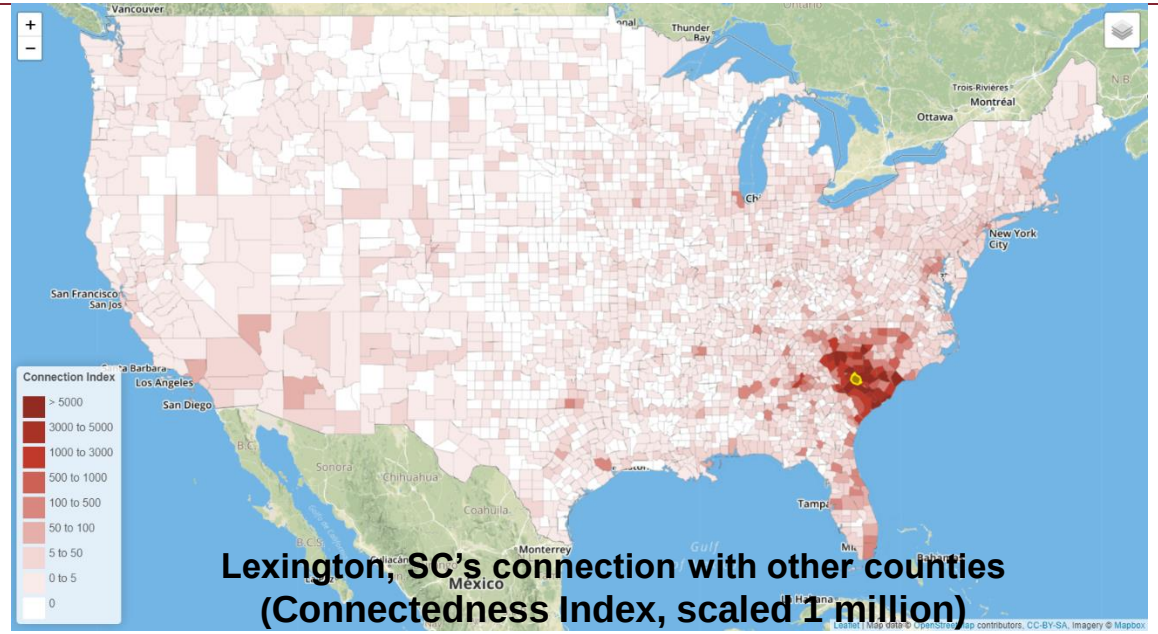


DEMO

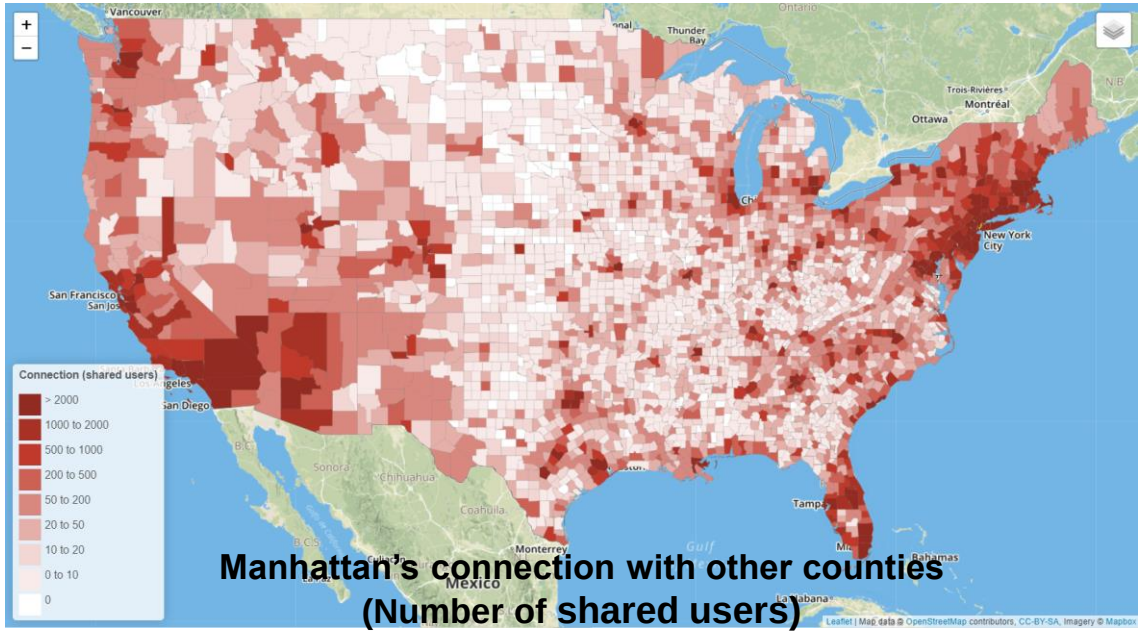
Measure Connection: explore the spatial patterns



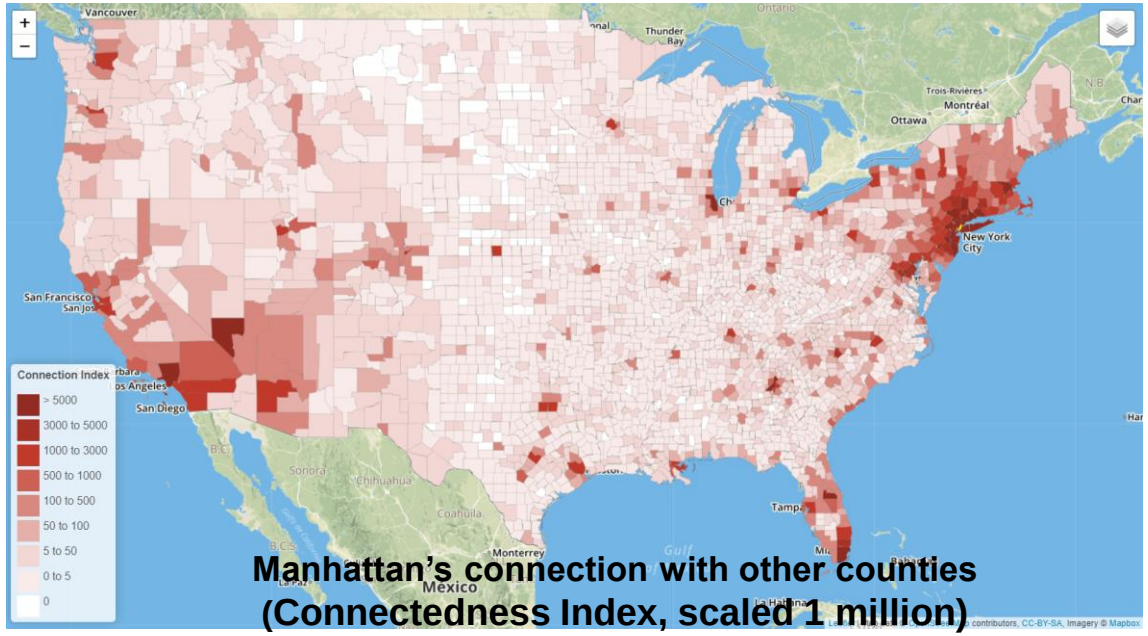
Lexington, SC's connection with other counties (Number of shared users)



Lexington, SC's connection with other counties (Connectedness Index, scaled 1 million)

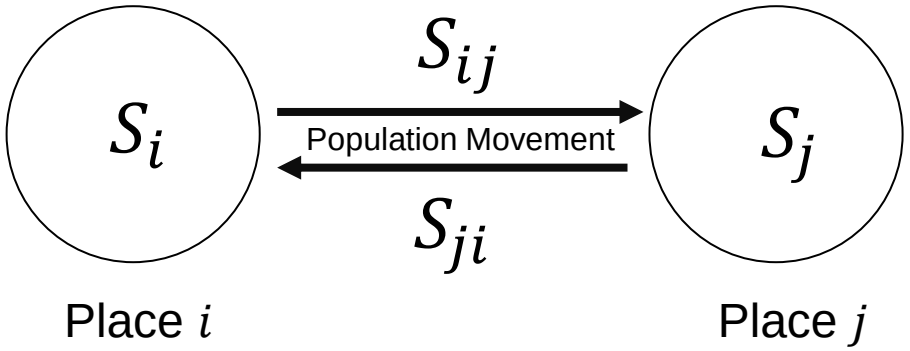


Manhattan's connection with other counties (Number of shared users)



Manhattan's connection with other counties (Connectedness Index, scaled 1 million)

Measure Connection: place connectedness based on movement



With direction

S_{ij} : number of people from place *i* to place *j*

S_{ji} : number of people from place *j* to place *i*

The CM can be computed as daily, weekly, monthly, yearly with varying spatial scales.

$$CI_{ij} = \frac{S_{ij}^2}{S_i \cdot S_j}$$

$$CI_{ji} = \frac{S_{ji}^2}{S_i \cdot S_j}$$

Connectedness Matrix (CM)

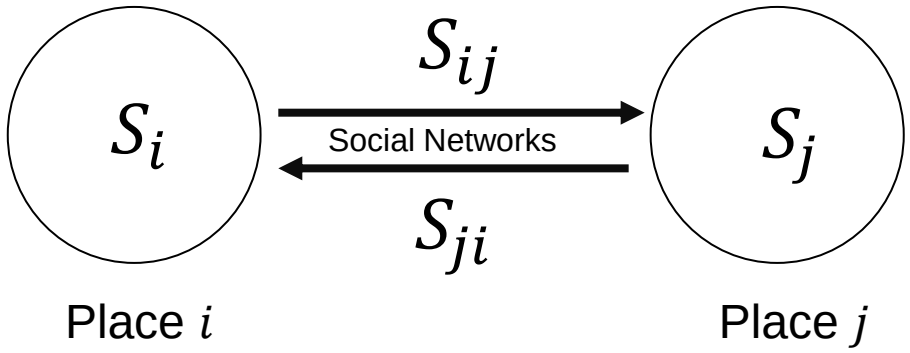
$$\boldsymbol{CM(t)} = \begin{bmatrix} CI_{11} & \cdots & CI_{n1} \\ CI_{12} & \ddots & CI_{n2} \\ \vdots & & \vdots \\ CI_{1n} & \cdots & CI_{nn} \end{bmatrix}$$

input to

predictive model

Li, Z., et al., (2020) Measuring Connectedness between Places using Big Social Media Data, (in preparation)

Measure Connection: place connectedness based on social network



We are also developing Connectedness Matrix (CM) based on the information flows on social networks

The information flow (with direction) will be derived from retweeting, mentioning, and following.

Keywords will be further used to extract the flow of specific information, e.g., COVID-19 topics.

$$CI_{ij} = \frac{S_{ij}^2}{S_i \cdot S_j}$$

$$CI_{ji} = \frac{S_{ji}^2}{S_i \cdot S_j}$$

Connectedness Matrix (CM)

$$\boldsymbol{CM(t)} = \begin{bmatrix} CI_{11} & \cdots & CI_{n1} \\ CI_{12} & \ddots & CI_{n2} \\ \vdots & & \vdots \\ CI_{1n} & \cdots & CI_{nn} \end{bmatrix}$$

input to

Virtual Space

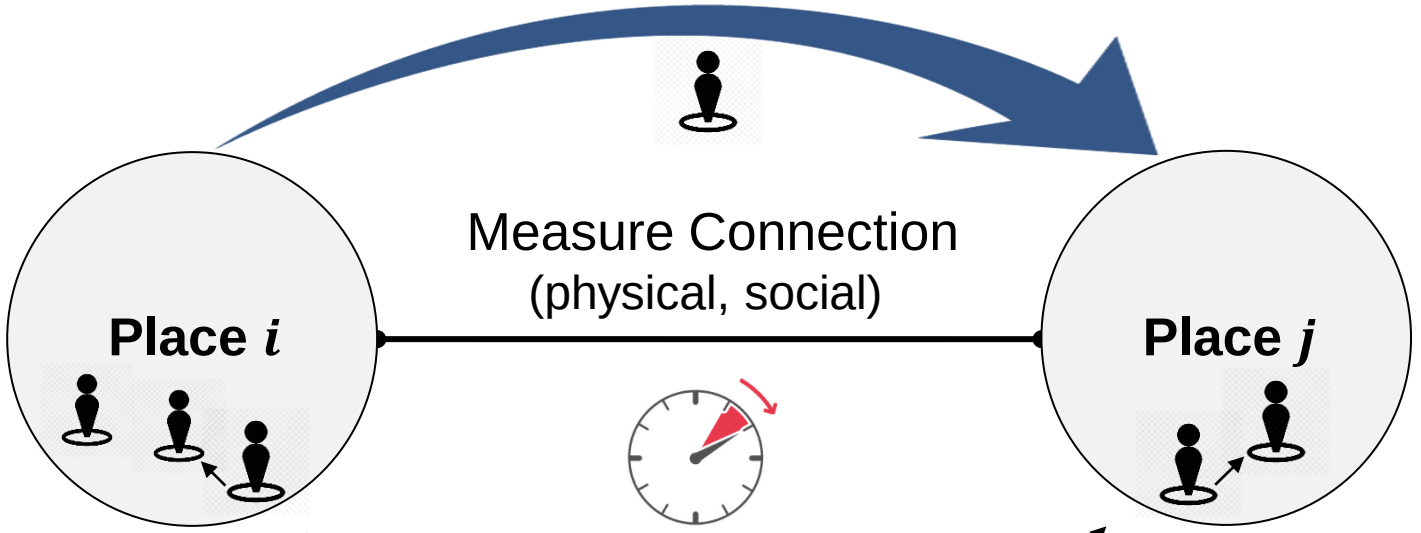
S_i

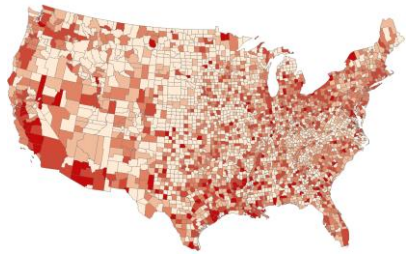
predictive model

Li, Z., et al., (2020) Measuring Connectedness between Places using Big Social Media Data, (in preparation)

Summary

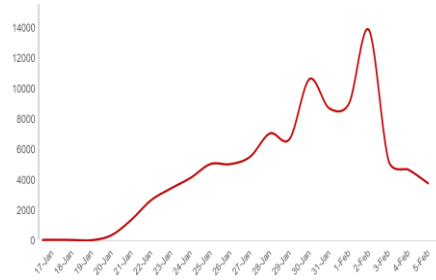
Extract and Visualize Population Flows





Risk maps

Predictive Modeling (with other datasets)

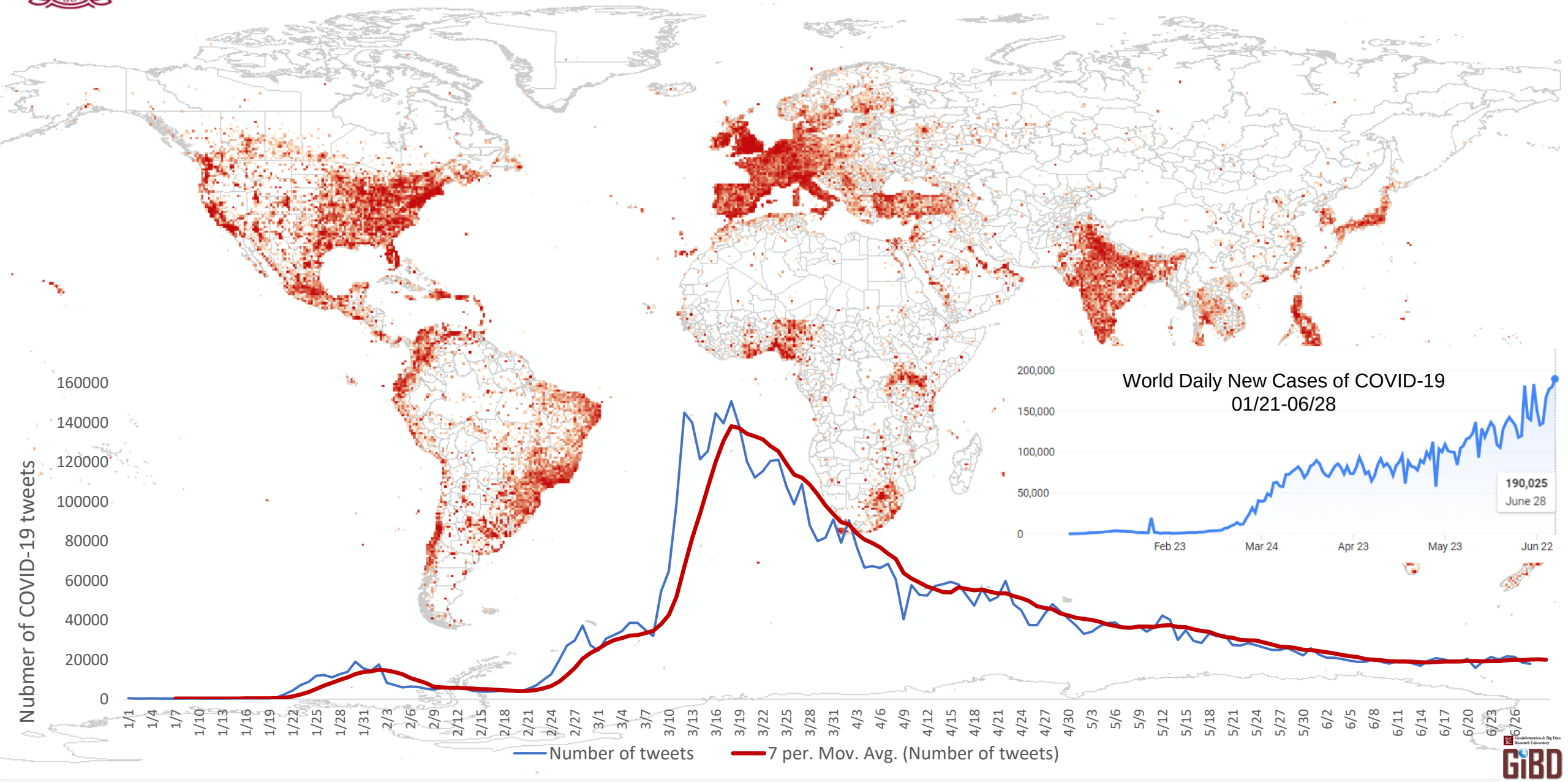


Daily cases

What's next?

- Integrate human mobility index and connectedness index/matrix into predictive models for the spread of COVID-19
  - Collaborating with our public health colleagues and Dr. Xinyue Ye's group
- Compare and combine twitter data with other mobility data sources
  - Mobile phone data
  - Google/Apple/Facebook mobility data (Baidu/Tencent data?)
- Examine the spatial disparity of the mobility changes and its relationship with the spatial disparity of the infected cases
  - Associate the mobility changes with socioeconomic and demographic factors at the different geographic scales
- New NIH grant → social media component
  - Develop an array of health-related indices (e.g., mental health, sentiment, HIV risk behaviors, awareness etc.) using text mining, AI at different spatiotemporal scales.

Spatiotemporal Distribution of Global COVID-19 Tweets from 01/01 - 06/28, 2020



Opportunities

Fully funded Ph.D. and Master student positions

http://gis.cas.sc.edu/gibd/opportunities/

Geoinformation and Big Data Research Laboratory @ USC

Making sense of Geospatial Big Data

Fully Funded Ph.D. Student Position: Geospatial Big Data Analytics for Health (updated on 04/27/2020)

The Geoinformation and Big Data Research Laboratory (GIBD) at the Department of Geography, University of South Carolina (USC) is looking for a highly motivated Ph.D. student with a great passion for geospatial big data research, **starting in Fall 2020 or Spring 2021**. GIBD is a collaborative effort of a group of faculty and students, conducting interdisciplinary research on geospatial big data analytics, spatiotemporal analysis/modeling, high-performance computing and CyberGIS within the area of data and computational intensive GIScience. By synthesizing advanced computing technologies, geospatial methods and spatiotemporal principles, GIBD aims to advance knowledge discovery and decision making to support domain applications including disaster management, human mobilities, public health, and climate change.

Health is intrinsically linked to geospatial context—where and how people interact with natural, built, social, economic and cultural environments directly influences human health experience, decision, outcome, policy-making and planning. Analyzing big health data with location information (e.g., electronic health records and social media data) offers an invaluable opportunity to improve the quality and efficiency of healthcare. The Ph.D. student will be expected to work with an interdisciplinary team to conduct cutting-edge research on geospatial big data analytics (e.g., analyzing massive social media data and heathcare records) by leveraging and/or developing advanced spatiotemporal analysis methods and computing algorithms/tools in the context of health science.

This is a fully funded position. The student will be supervised by **Dr. Zhenlong Li** and will also join the Big Data Health Science Center at USC, a campus-wide interdisciplinary enterprise that conducts cutting-edge health research and discovery, offers professional development and academic training, and provides service to the community and industry.

The lab also has a funded Master student position starting in Fall 2020 or Spring 2021.

If you are interested or have any questions, please don't hesitate to contact Dr. Li at zhenlong@sc.edu.

What's New

Check out our new paper "Twitter, human mobility, and COVID-19"

UofSC Big Data Health Science Center investigators receive $1.25 million NIH grant to develop data-driven strategies in fighting COVID-19

Recent Publications

Using geotagged tweets to track population movements to and from Puerto Rico after Hurricane Maria, Population and Environment

Spatiotemporal Event Detection: A Review, International Journal of Digital Earth

Read more...

UNIVERSITY OF SOUTH CAROLINA
Department of Geography

References

- Ning H., Li Z., Hodgson M., Wang C., (2020), Prototyping a Social Media Flooding Photo Screening System Based on Deep Learning, *ISPRS International Journal of Geo-Information, 9(2), 104;* https://doi.org/10.3390/ijgi9020104
- Martín Y., Cutter S.L. Li Z, Emrich C., Mitchell, J.T. (2020) Using geotagged tweets to track population movements to and from Puerto Rico after Hurricane Maria. *Population and Environment ,* https://doi.org/10.1007/s11111-020-00338-6
- Hu L., Li Z., Ye X., (2020) Delineating and Modelling Activity Space Using Geotagged Social Media Data, *Cartography and Geographic Information Science*, https://doi.org/10.1080/15230406.2019.1705187
- Li Z., Huang Q., Emrich C., (2019) Introduction to Social Sensing and Big Data Computing for Disaster Management, *International Journal of Digital Earth*, 12(11), 10.1080/17538947.2019.1670951
- Huang X., Li Z., Wang C., Ning H., (2019), Identifying disaster related social media for rapid response: a visual-textual fused CNN architecture, *International Journal of Digital Earth, https://doi.org/10.1080/17538947.2019.1633425*
- Martín Y., Cutter S., Li Z.,(2019) Bridging social media and survey data for the evacuation assessment of hurricanes, *Natural Hazard Review, 21(2), https://doi.org/0.1061/(ASCE)NH.1527-6996.0000354*
- Jiang Y., Li Z., Cutter S.,(2019) Social network, activity space, sentiment, and evacuation: what can social media tell us? *Annals of the American Association of Geographers,* DOI: 10.1080/24694452.2019.1592660
- Dahal, Kumar, Li Z.,(2019) Spatiotemporal Topic Modeling and Sentiment Analysis of Global Climate Change Tweets, *Social Network Analysis and Mining*, https://doi.org/10.1007/s13278-019-0568-8
- Huang X., Wang C. Li Z. Ning H.,(2018) A visual-textual fused approach to automated tagging of flood-related tweets during a flood event, *International Journal of Digital Earth, 10.1080/17538947.2018.1523956*
- Hu F., Li Z., Yang C., Jiang Y. (2018) A graph-based approach to detect the tourist movement pattern using social media data, *Cartography and Geographic Information Science, 10.1080/15230406.2018.1496036*
- Huang X., Wang C., Li Z. (2018) A Flooding Probability Reconstruction Approach by Enhancing Near Real-Time Imagery with Real-Time Gauges and Tweets, *IEEE Transactions on Geoscience and Remote Sensing*, 10.1109/TGRS.2018.2835306
- Wang C., Li Z., Huang X. (2018) Geospatial assessment of flooding dynamics and risks of the October'15 South Carolina Flood, Congaree River Watershed, *Southeastern Geographer,* 58(2), 164-180
- Jiang Y., Li Z., Ye X. (2018) Understanding Demographic and Socioeconomic Bias of Geotagged Twitter Users at the County Level, *Cartography and Geographic Information Science,* 15230406.2018.1434834
- Huang X., Wang C., Li Z., (2018) A Near Real-time Flood Mapping Approach by Integrating Post-event with Satellite Imagery and Flood-related Tweets, *Annals of GIS*
- Martín, Y., Li, Z., & Cutter, S. L. (2017). Leveraging Twitter to gauge evacuation compliance: spatiotemporal analysis of Hurricane Matthew. PLoS one, 12(7), e0181701.
- Li Z., Wang C., Emrich C., Guo D.(2017) A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods, *Cartography and Geographic Information Science,* doi: 10.1080/15230406.2016.1271356

Acknowledgements: Funding

- **2020-2021, RAPID: Monitoring the Spatial Spread of COVID-19 through the Lens of Human Movement using Big Social Media Data, National Science Foundation (NSF)**
- 2020-2022, A Preliminary Study of using Social Media to Monitor the Spatial Propagation of COVID-19 and Quantify the Effectiveness of the Control Measures, USC COVID-19 Internal Funding Initiative
- 2017-2018, Enhancing Situational Awareness by Mining Big Social Media Data in Near-real Time for Disaster Management: A CyberGIS Approach, USC OVPR, ASPIRE

UNIVERSITY OF SOUTH CAROLINA

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

Thank you !

Questions/Comments?

Zhenlong Li
Geoinformation and Big Data Research Laboratory (GIBD)
Big Data Health Science Center
Hazards and Vulnerability Research Institute
Department of Geography, University of South Carolina

zhenlong@sc.edu

For more information about our research, please visit our lab website:
http://gis.cas.sc.edu/gibd
http://gis.cas.sc.edu/gibd/covid-19

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