Biases in the big data and what we can do about them as a community

Cynthia Chen (UW) and Ryan Wang (Northeastern)

UW workshop on big data, AI and transportation planning applications

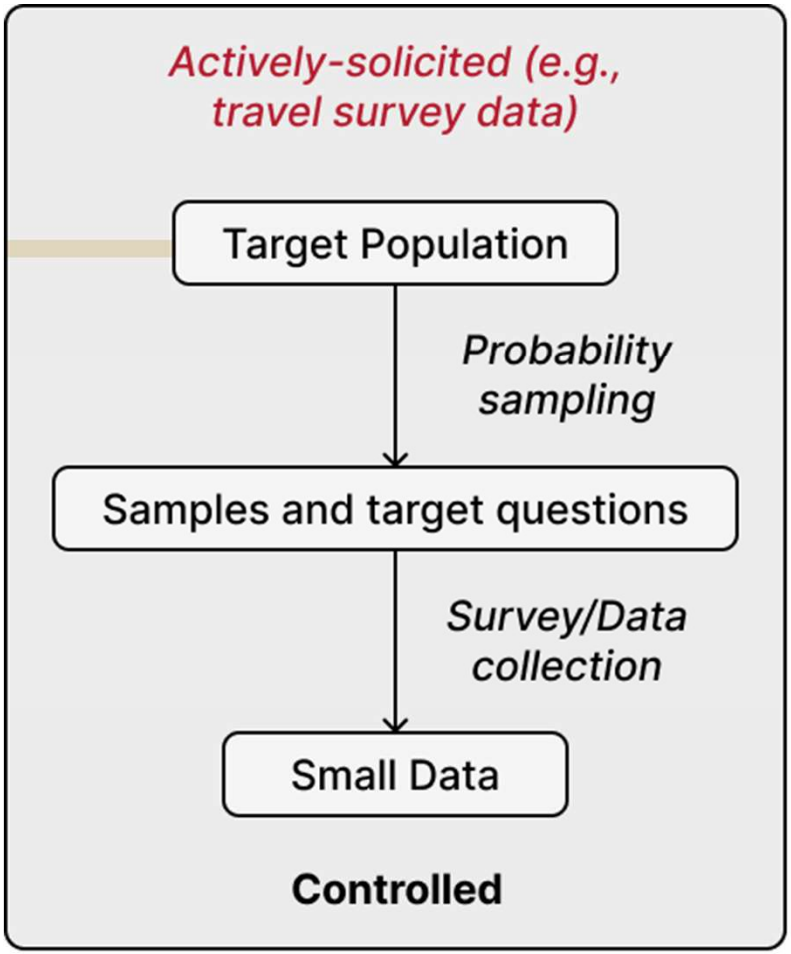
CIVIL & ENVIRONMENTAL ENGINEERING
UNIVERSITY of WASHINGTON

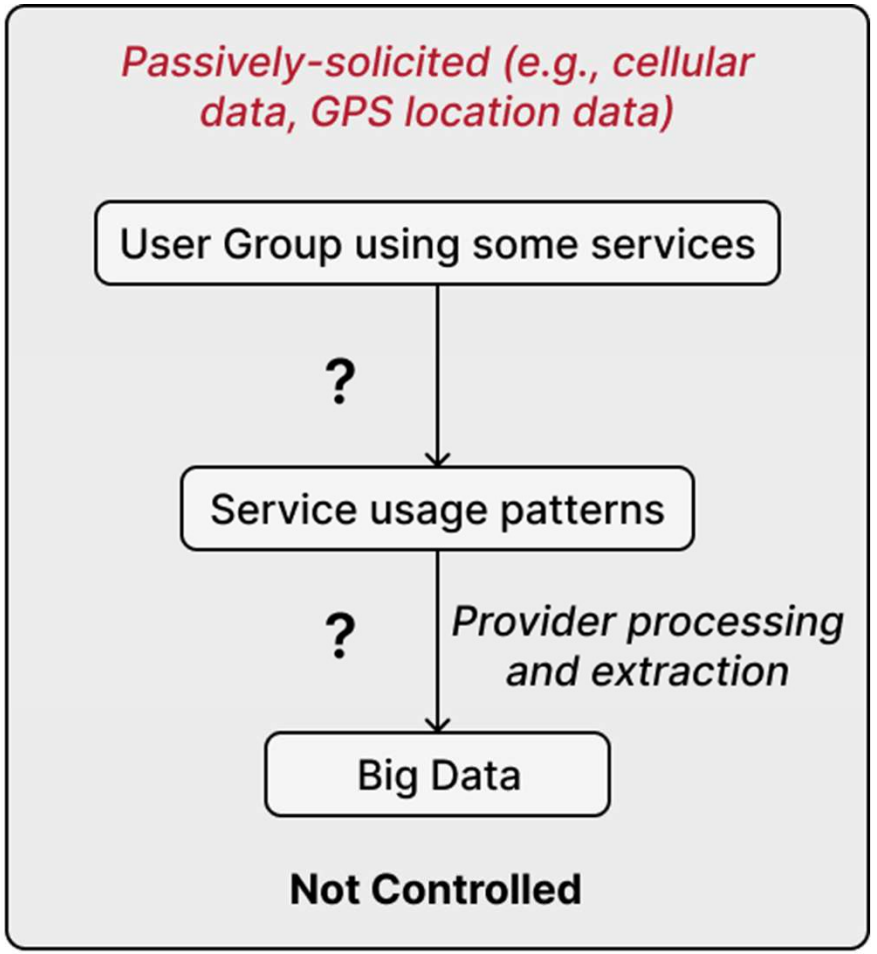
Outline

- Motivation
- Quantifying biases in Location-Based-Service (LBS) data
  - Data stability, sparsity, pre- and post-processing, algorithms, and socio-demographic and built environment factors
- Where opportunities lie in transportation planning?
- Areas that we can work together
  - Scientific aspects
  - Practical aspects

Funding acknowledgments

> CIS program (now ISP program);
> NSF AI Institute
> Amazon Middle Mile Transportation (Tim Jacobs)





Active Data Generation

Passive Data Generation

Table 1. An example HTS (left) and LBS dataset (right)

Actively-generated HTS data						
Date	**Departure Time**	**Arrival Time**	**Departure Census Tract[1]**	**Arrival Census Tract[1]**	**Trip Purpose**	**Main Travel Mode**
2020-01-02	7:00 AM	7:45 AM	53033006100	53033005300	Work	Bus
2020-01-02	12:00 PM	12:20 PM	53033005300	53033005200	Lunch	Walk
2020-01-02	12:45 PM	1:10 PM	53033005200	53033005300	Work	Walk
2020-01-02	4:30 PM	4:50 PM	53033005300	53033005100	Recreation	Bus
2020-01-02	5:50 PM	6:40 PM	53033005100	53033006100	Home	Bus

[1] HTS collects lat and long information for every trip origin and destination, which are then converted to census tracts.

Passively-generated LBS data			
Time	**Latitude**	**Longitude**	**Location Accuracy (m)**
2020-01-02 7:13:30	42.824731	-71.115226	65
2020-01-02 7:29:11	42.837882	-71.057814	20
2020-01-02 11:11:31	42.851232	-70.913241	12
2020-01-02 13:10:22	42.858141	-70.913241	5
2020-01-02 18:32:38	42.823326	-71.112916	10

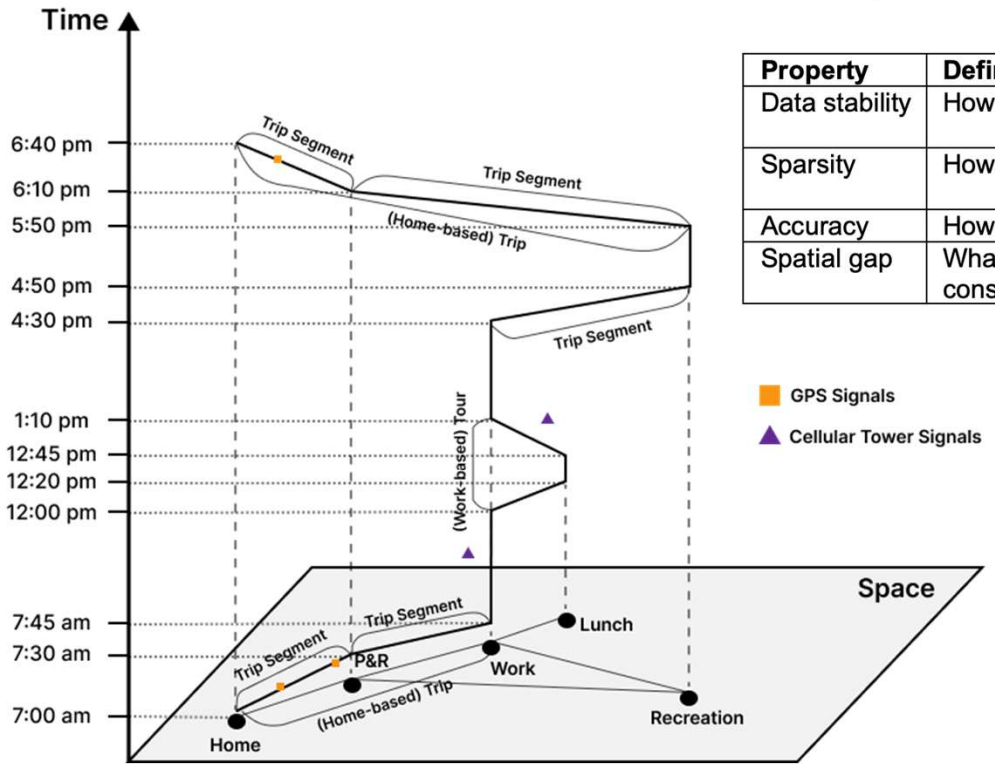


Table 3. Properties of LBS Data

Property	Definition	Metric
Data stability	How stable is the LBS data over time?	Number of devices per time period Number of records per person per day
Sparsity	How temporally sparse is the LBS data?	Intra-day occupancy Inter-day occupancy
Accuracy	How spatially accurate is the LBS data?	Locational accuracy in meters
Spatial gap	What is the spatial gap (in km) between consecutive observations?	Jumping distance in kilometers

Outline

> Motivation

> Quantifying biases in Location-Based-Service (LBS) data

- Data stability, sparsity, pre- and post-processing, algorithms, and socio-demographic and built environment factors

> Where opportunities lie in transportation planning?

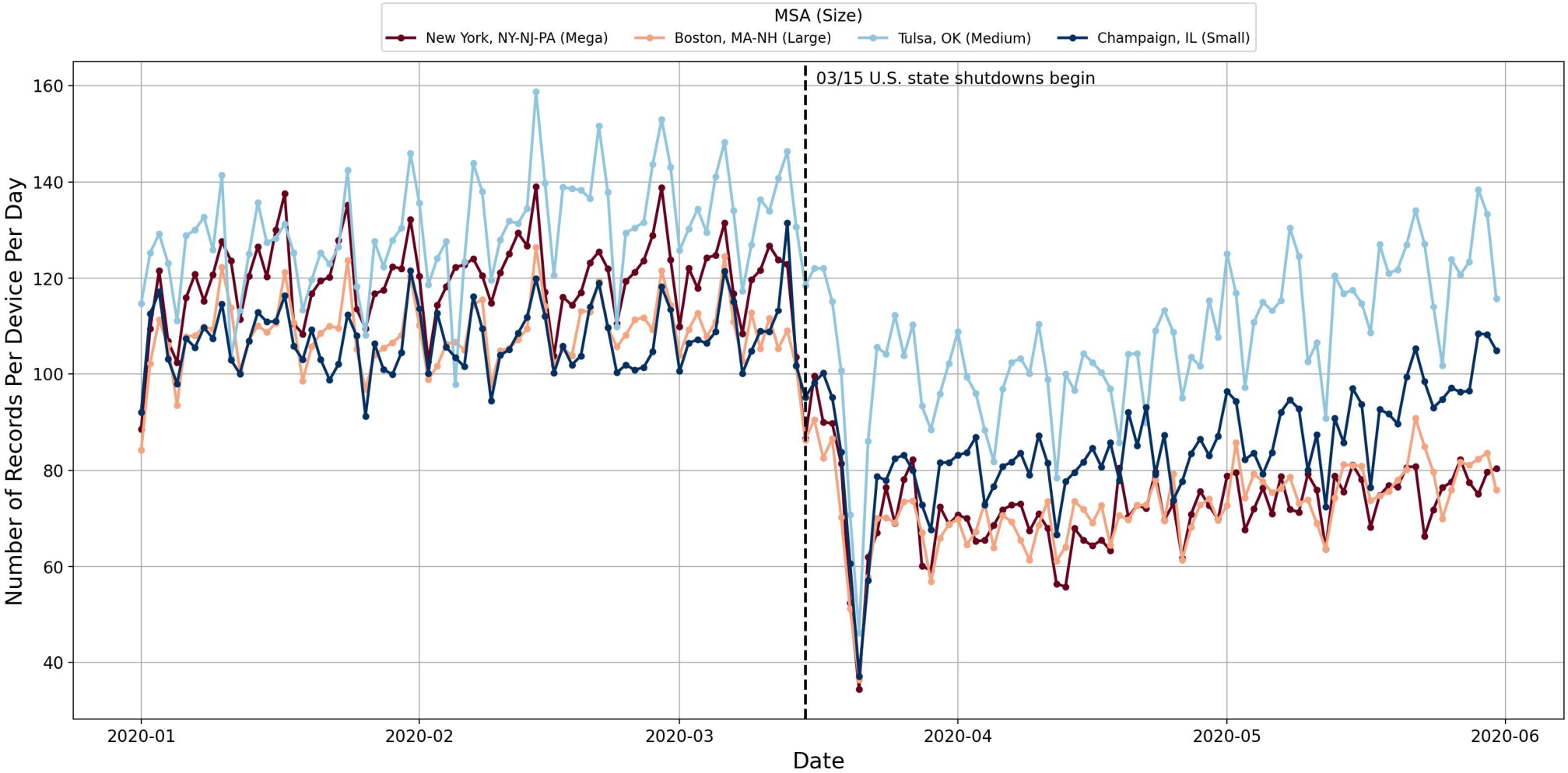
> Areas that we can work together

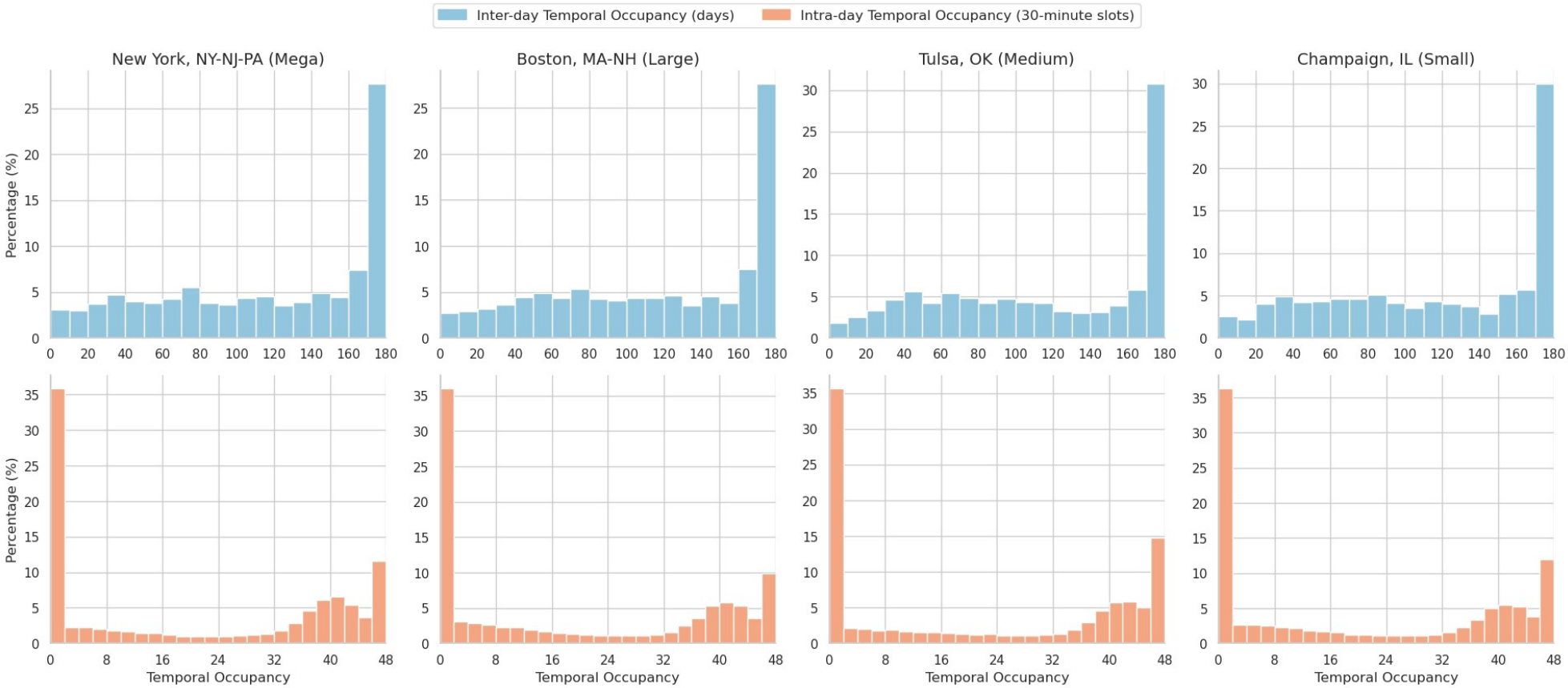
- Scientific aspects
- Practical aspects

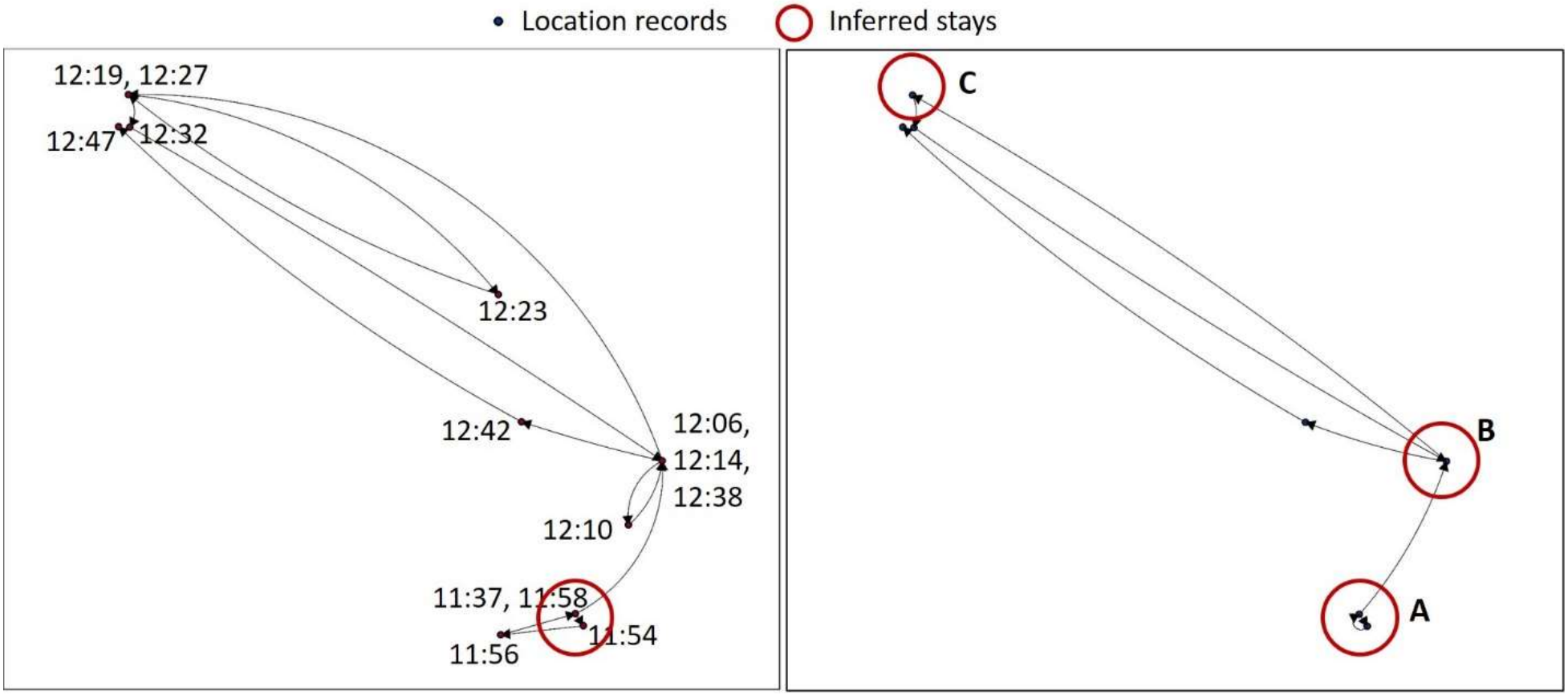
A sample of 11 metro regions in U.S.

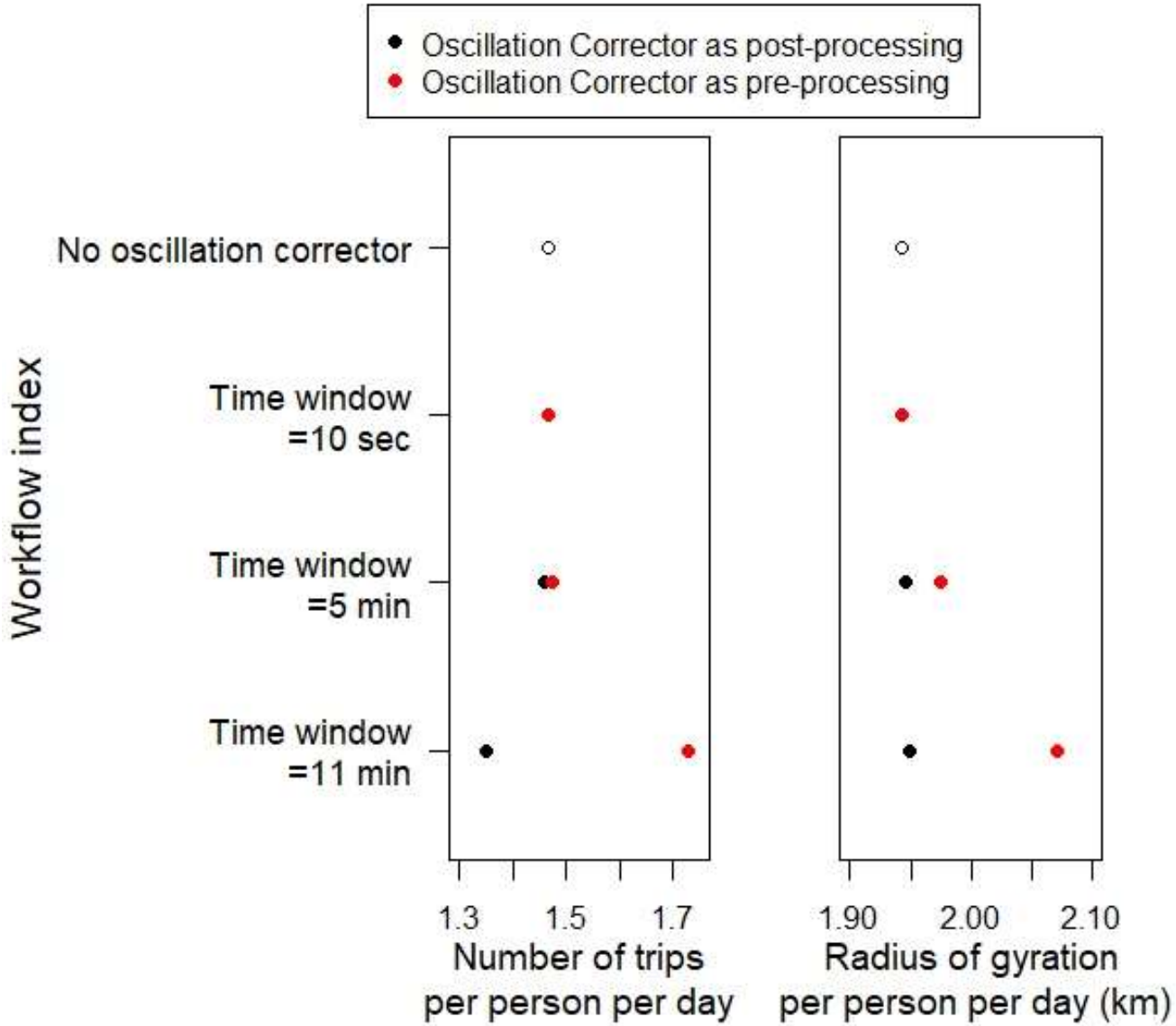
MSA Type	MSA Name	Population	Sampling rate[1] (%)	Study sample size[2]	Population density (pp/sq. m^2)
Mega	*New York-Newark-Jersey City, NY-NJ-PA*	*22,432,947*	*4.8*	*15,992*	*2,934.9*
	Los Angeles-Long Beach-Anaheim, CA	12,872,322	4.2	15,000	2,652.9
Large	*Boston-Cambridge-Newton, MA-NH*	*4,900,550*	*6.1*	*10,000*	*1,405.7*
	Seattle-Tacoma-Bellevue, WA	4,034,248	4.6	10,000	687.3
	Baltimore-Columbia-Towson, MD	2,844,510	10.1	10,000	1,090
Medium	*Tulsa, OK*	*1,034,123*	*10.6*	*10,000*	*164.8*
	Fresno, CA	1,171,617	7.5	10,000	170.4
	Tyler, TX	233,479	11.7	10,000	262.5
Small	*Champaign-Urbana, IL*	*236,514*	*7.9*	*10,000*	*155.6*
	Sebring-Avon Park, FL	105,618	10.0	10,000	103.8
	Cheyenne, WY	100,984	8.1	9,008	37.5

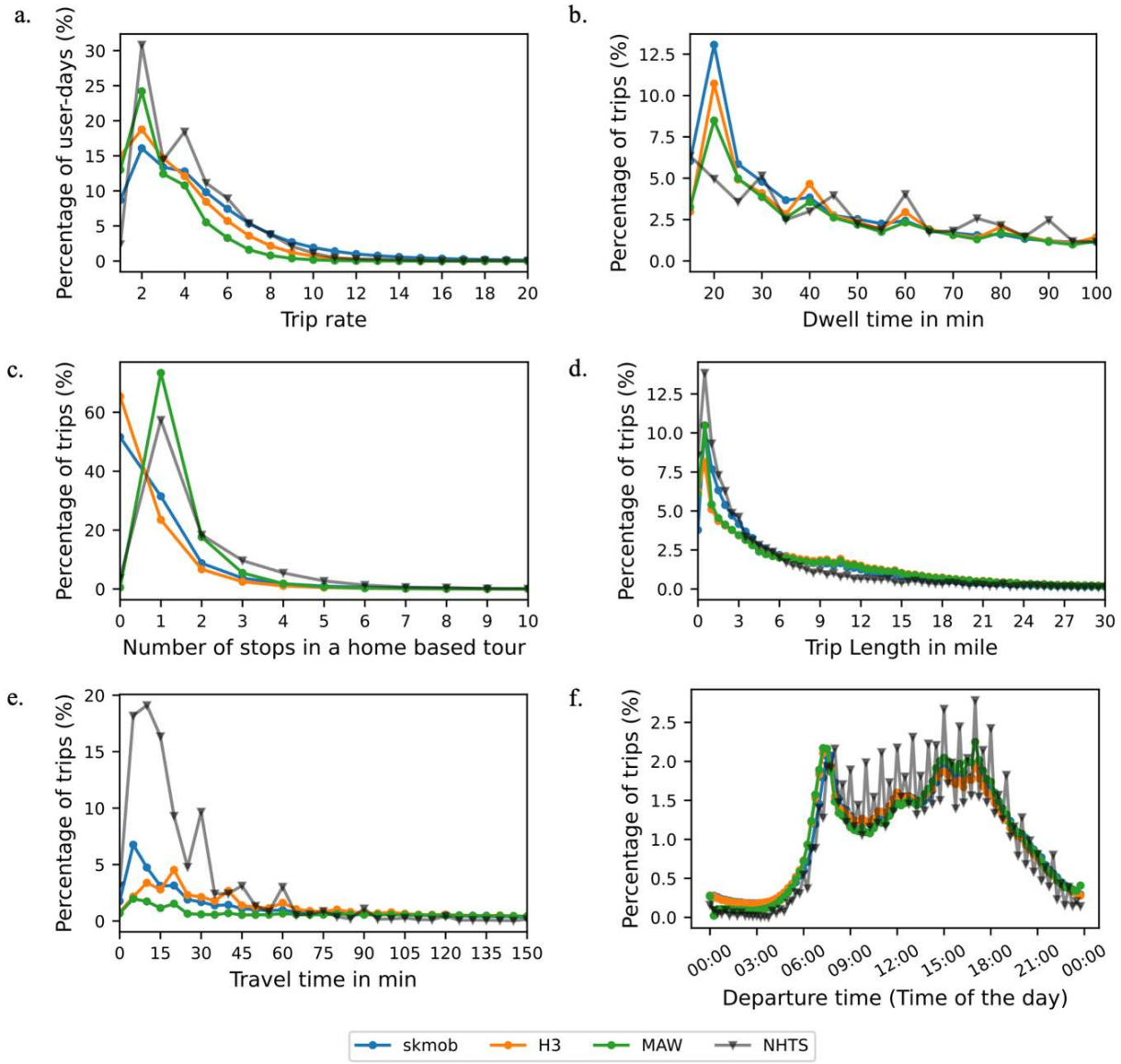
[1] Sampling rate is the ratio of number of unique devices with an inferred home location located in the area divided by the area's population. [2] Study sample size is the size of the data used for the analysis in this paper.

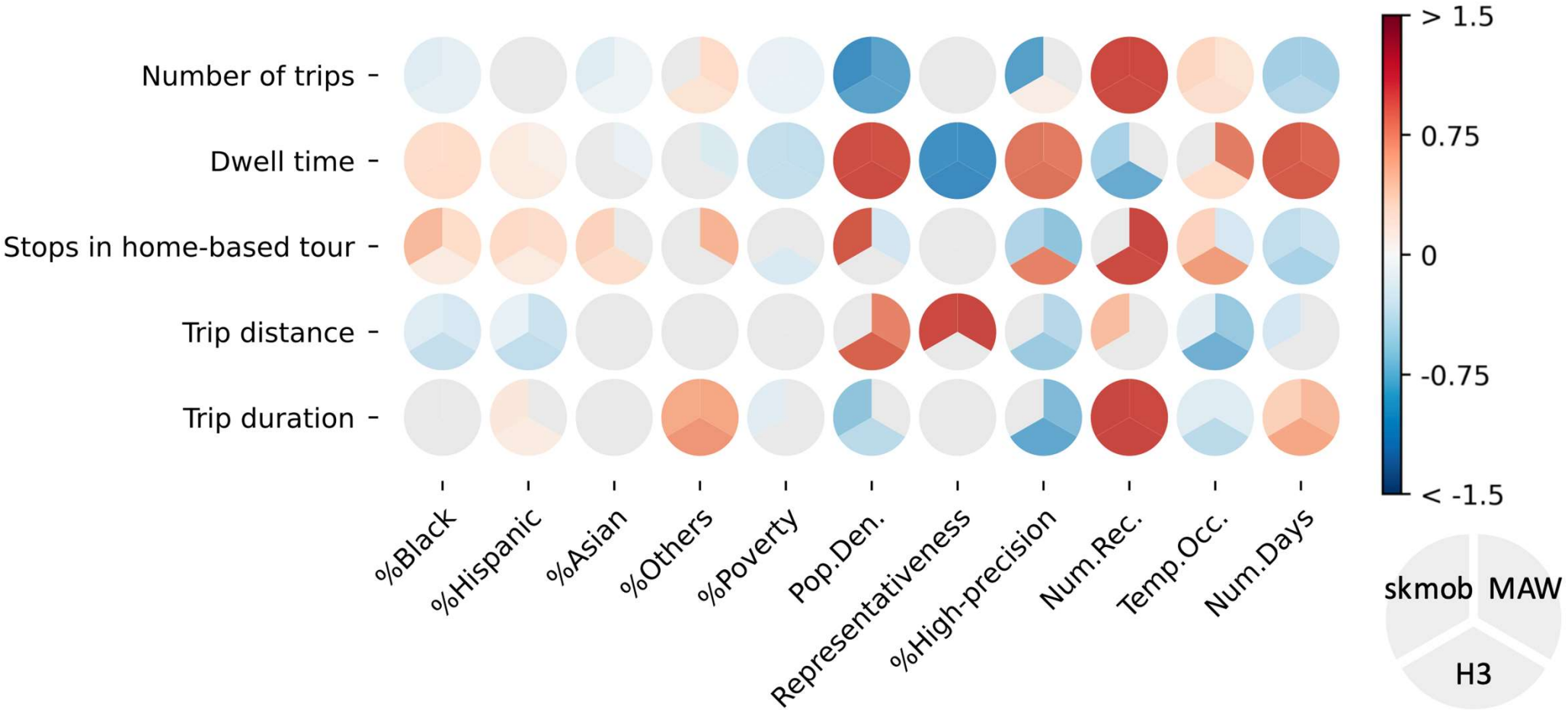

MSA (Size)
New York, NY-NJ-PA (Mega)
Boston, MA-NH (Large)
Tulsa, OK (Medium)
Champaign, IL (Small)
03/15 U.S. state shutdowns begin
Number of Records Per Device Per Day
160
140
120
100
80
60
40
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
Date


Inter-day Temporal Occupancy (days)
Intra-day Temporal Occupancy (30-minute slots)
New York, NY-NJ-PA (Mega)
Boston, MA-NH (Large)
Tulsa, OK (Medium)
Champaign, IL (Small)
Percentage (%)
Temporal Occupancy


Location records
Inferred stays
12:19, 12:27
12:47
12:32
12:23
12:42
12:06,
12:14,
12:38
12:10
11:37, 11:58
11:54
11:56
C
B
A

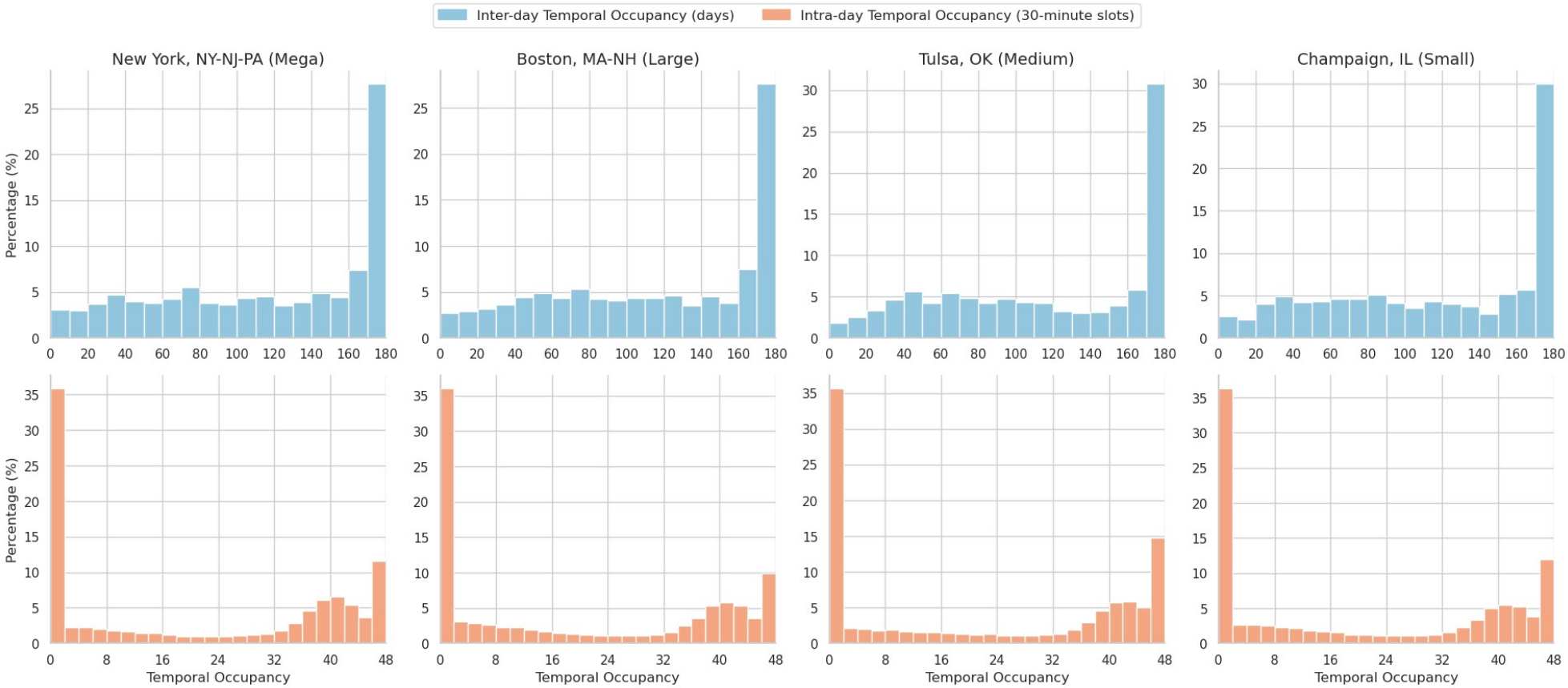

Oscillation Corrector as post-processing
Oscillation Corrector as pre-processing
Workflow index
No oscillation corrector
Time window =10 sec
Time window =5 min
Time window =11 min
1.3
1.5
1.7
Number of trips per person per day
1.90
2.00
2.10
Radius of gyration per person per day (km)


a.
Percentage of user-days (%)
Trip rate
b.
Percentage of trips (%)
Dwell time in min
c.
Percentage of trips (%)
Number of stops in a home based tour
d.
Percentage of trips (%)
Trip Length in mile
e.
Percentage of trips (%)
Travel time in min
f.
Percentage of trips (%)
Departure time (Time of the day)
skmob
H3
MAW
NHTS


Number of trips
Dwell time
Stops in home-based tour
Trip distance
Trip duration
%Black
%Hispanic
%Asian
%Others
%Poverty
Pop.Den.
Representativeness
%High-precision
Num.Rec.
Temp.Occ.
Num.Days
> 1.5
0.75
0
-0.75
< -1.5
skmob
MAW
H3

Outline

- Motivation
- Quantifying biases in Location-Based-Service (LBS) data
  - Data stability, sparsity, pre- and post-processing, algorithms, and socio-demographic and built environment factors
- Where opportunities lie in transportation planning?
- Areas where we can work together
  - Scientific aspects
  - Practical aspects


Inter-day Temporal Occupancy (days)
Intra-day Temporal Occupancy (30-minute slots)
New York, NY-NJ-PA (Mega)
Boston, MA-NH (Large)
Tulsa, OK (Medium)
Champaign, IL (Small)
Percentage (%)
Temporal Occupancy

Outline

- Motivation
- Quantifying biases in Location-Based-Service (LBS) data
  - Data stability, sparsity, pre- and post-processing, algorithms, and socio-demographic and built environment factors
- Where opportunities lie in transportation planning?
- Areas that we can work together
  - Scientific aspects
  - Practical aspects

Where do opportunities lie?

- Aiding National Household Travel Survey (NHTS) data
- Analysis of travel patterns by population segments
- Real time, targeted policy formulation and evaluations
- Resilience and adaptation analysis
- LBS as the foundational data for micro-simulation of activity-travel patterns

Outline

- Motivation
- Quantifying biases in Location-Based-Service (LBS) data
  - Data stability, sparsity, pre- and post-processing, algorithms, and socio-demographic and built environment factors
- Where opportunities lie in transportation planning?
- Areas that we can work together
  - Scientific aspects
  - Practical aspects

Scientific aspects

- Data pre-processing
- Data fusion
- Change point detection
- Imputing demographics
- Targeted policy formulation and evaluation

Practical aspects

- Journals and funding sponsors shall require reporting of the characteristics of the data used in the study;
- Sensitivity analyses shall be conducted with respect to how trips and modes are detected;
- We shall compile a list of open-source data sources as benchmarks;
- Better, a central registry shall be established for all studies using big data for planning related analysis;