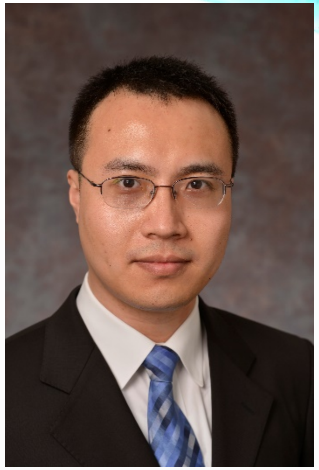
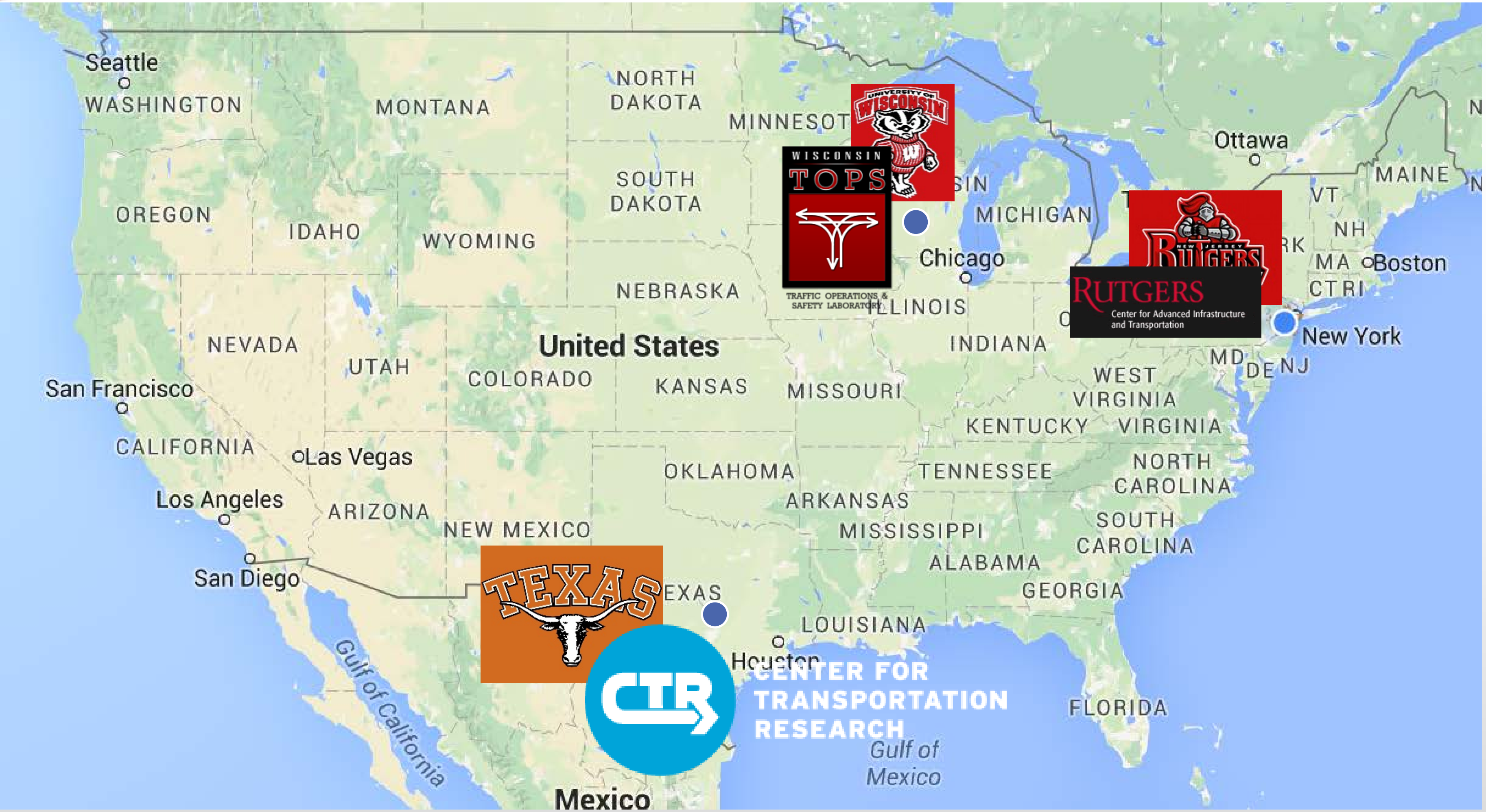
LOCATION-BASED SOCIAL NETWORK (**LBSN**) DATA: EMERGING **BIG DATA** SOURCES FOR TRAVEL DEMAND AND ACTIVITY MODELING

2/11/2016, Seminar at Northwestern University, Evanston, IL,

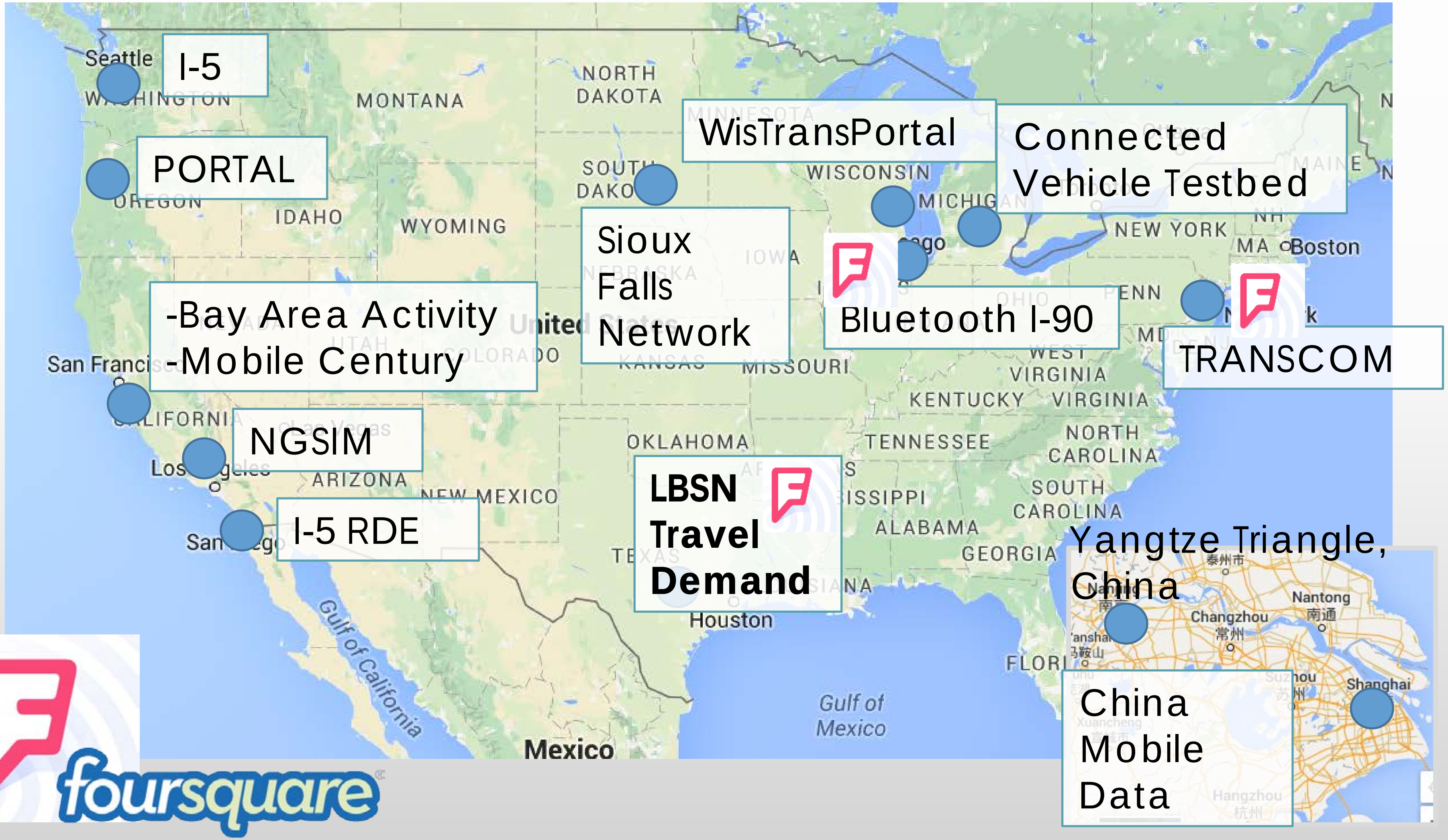
Peter J. Jin, Ph.D., Department of Civil and Environmental Engineering,
Rutgers, The State University of New Jersey



- **Peter J. Jin**, Ph.D., Assistant Professor, CEE, Rutgers University
- Education:
  - Ph.D.: CEE, University of Wisconsin-Madison, 2009, Advisor: Prof. Bin Ran
  - M.S.: CEE, University of Wisconsin-Madison, 2007, Advisor: Prof. Bin Ran
  - B.S.: Automation, Tsinghua University, China, 2005
- Employment
  - Assistant Professor, Rutgers, The State University of New Jersey, 2014-now
  - Postdoctoral Fellow: The University of Texas at Austin, 2011-2013, Advisor: Dr. C. Michael Walton.
  - Research Associate: University of Wisconsin-Madison, 2010-2011, Advisor: Prof. Bin Ran
- Research Area:
  - Transportation Big Data Analytics
  - Traffic Operations (Active Traffic and Demand Management, Mobile Sensor Data)
  - Connected Vehicles, Autonomous Vehicles, Ridesharing
  - Unmanned Aerial Vehicles
- Publications: 31 Journal, 47 Conference Papers


UNIVERSITY OF WISCONSIN
WISCONSIN TOPS
TRAFFIC OPERATIONS & SAFETY LABORATORY
RUTGERS
RUTGERS
Center for Advanced Infrastructure and Transportation
TEXAS
CTR
CENTER FOR TRANSPORTATION RESEARCH
Seattle
WASHINGTON
MONTANA
NORTH DAKOTA
MINNESOTA
SOUTH DAKOTA
OREGON
IDAHO
WYOMING
NEBRASKA
MICHIGAN
Chicago
ILLINOIS
INDIANA
Ottawa
MAINE
VT
NH
MA
Boston
CT
RI
New York
NEVADA
UTAH
United States
COLORADO
KANSAS
MISSOURI
San Francisco
CALIFORNIA
Las Vegas
MD
DE
NJ
WEST VIRGINIA
VIRGINIA
KENTUCKY
OKLAHOMA
TENNESSEE
NORTH CAROLINA
Los Angeles
ARIZONA
NEW MEXICO
ARKANSAS
MISSISSIPPI
SOUTH CAROLINA
San Diego
ALABAMA
GEORGIA
LOUISIANA
Houston
Gulf of California
FLORIDA
Gulf of Mexico
Mexico

TRAFFIC OPERATIONS/PLANNING DATASETS



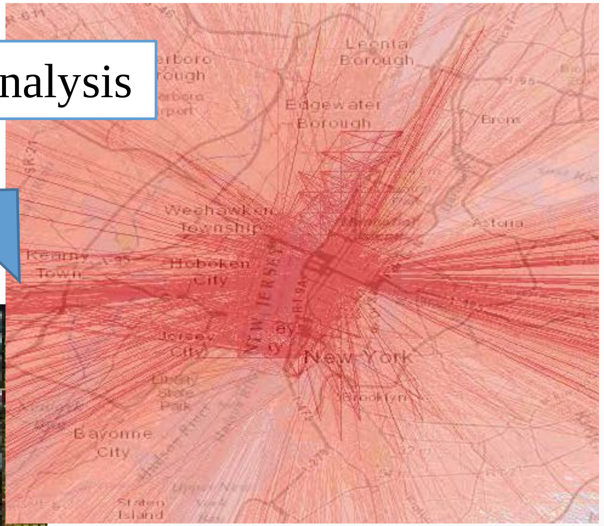
TRANSPORTATION BIG DATA?

- **Volume**: 24*4 Operations, Historical ITS Data
- **Variety**: Sensor, Probe, Infrastructure, Survey, Secondary
- **Veracity:** Agency versus Crowdsourcing (WAZE) data
- **Velocity:** 10Hz DSRC, I-Pass/EZ-Pass => 5-10 Year NHTS Data
- **Value:** Public Sector (Congestion mitigation) versus Consumer Market (1.1 Billion WAZE)

BIG DATA DECISION-MAKING

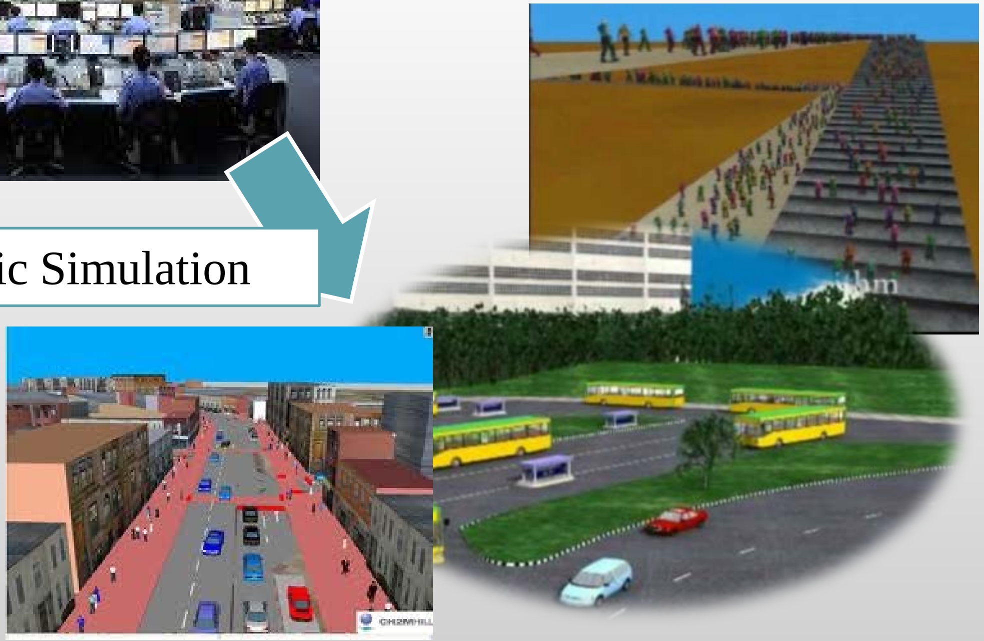
Traveler Information

Travel Demand Analysis

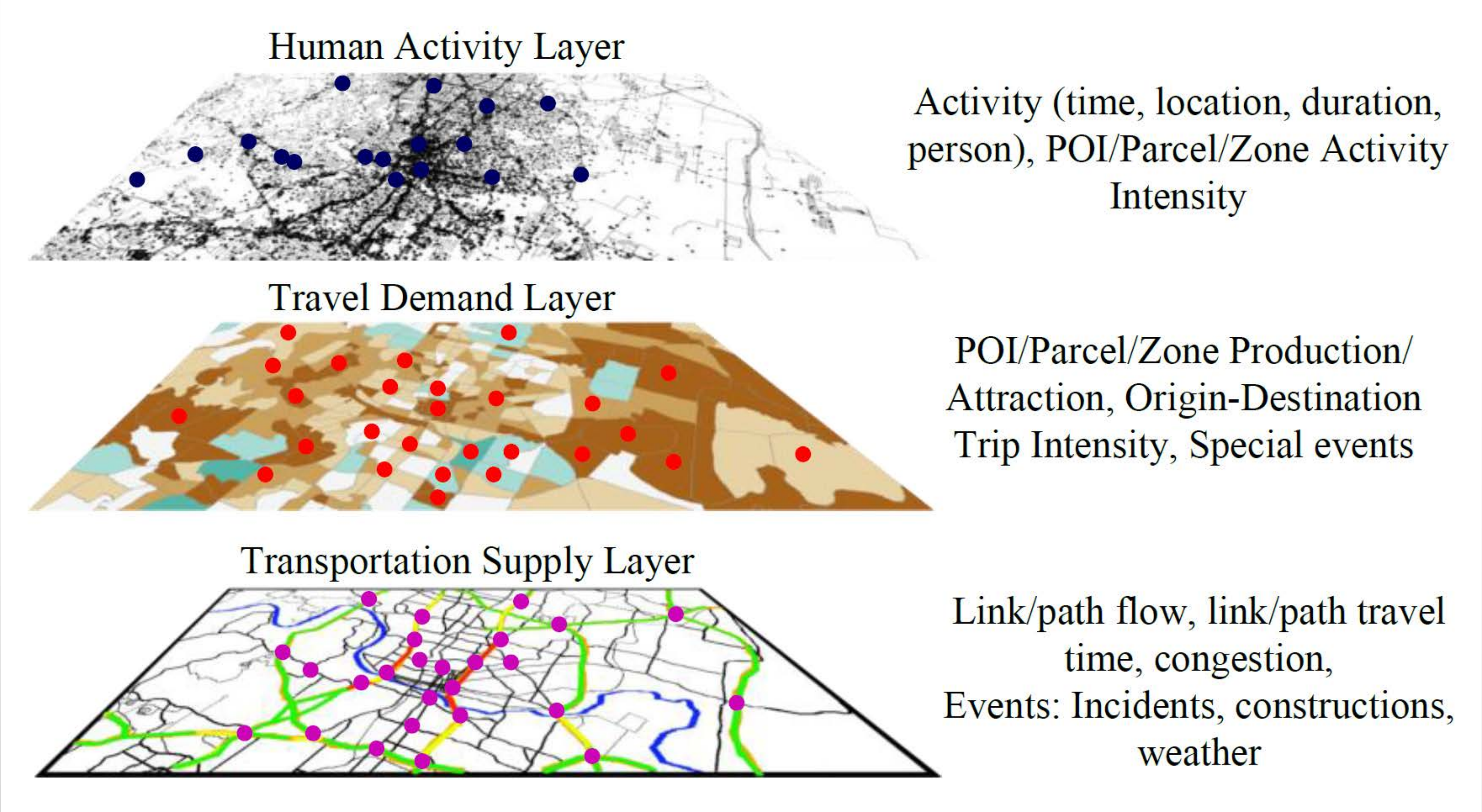


Emergency Response/Planning

Traffic Simulation



MULTI-LAYER FRAMEWORK



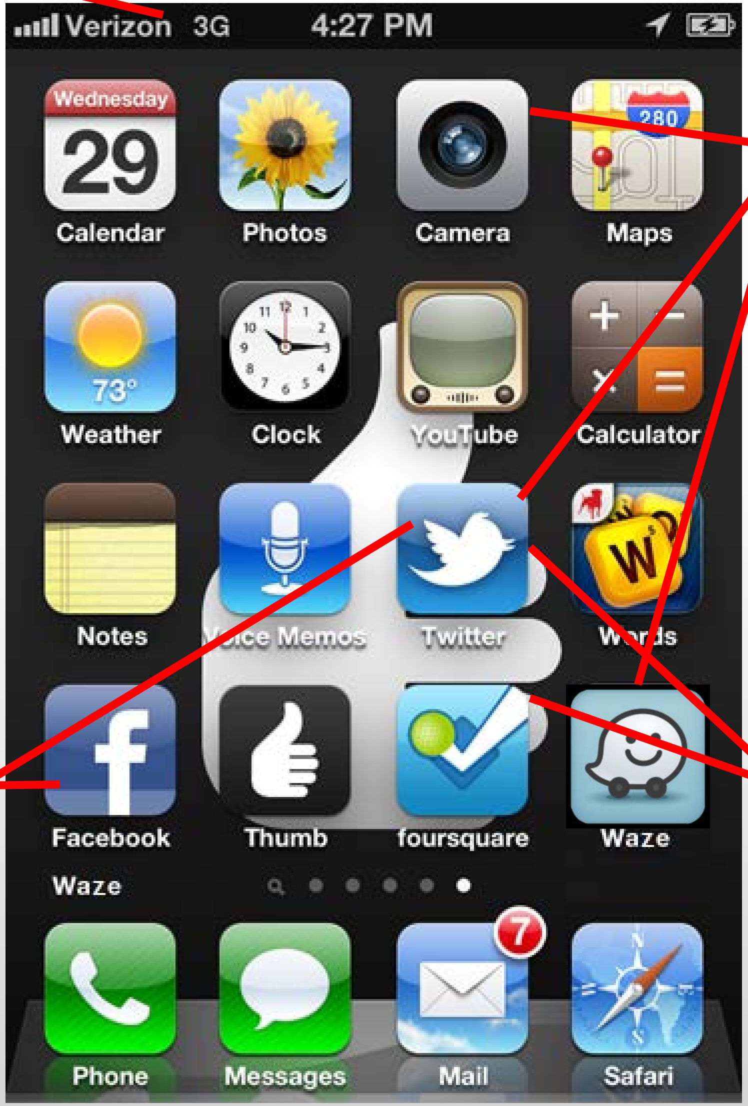
EMERGING SECONDARY DATA SOURCE

Information and Concerns	Survey	GPS	License Plate	Blue-tooth	Smart Phone	Cell Phone	Social Media	VS-LBSN
Origin-Destination	Y	Y	Y	Y	Y	Y	Y	Y
Mode choice	Y	M	Y-auto	Y-auto	Y	Y	Y	M
Trip Purpose	Y	M	M	N	Y	M	Y	Y
Routes	Y	Y	Y	Y	Y	Y	N	M
Trip Frequency	Y	Y	Y	Y	Y	Y	Y	M
Trip Chain	Y	Y	M	M	Y	M	Y	N
Traveler Characteristics	Y	M	M	N	M	M	Y	M
Passive Data Collection	N	Y	M	Y	M	Y	M	Y
Major Privacy Concern	M	M	Y	N	M	M	Y	N
Respondent Burden	High	Medium	No	No	No/M	No	N	N
Sampling Bias	M	M	N	Y	M	M	Y	Y
Sufficient Sample Size	M	M	Y	Y	M	Y	Y	Y
Trip information confirmation	Y	M	M	M	M	M	Y	Y
Spatial resolution	Low	Low	Low	Low	High	High	High	High
Temporal resolution	Low	High	High	High	High	High	High	High

M: Maybe (implies information may be indirectly estimated).
Source (except VS-LBSN): NCHRP Report 735 (Schiffer, 2012). Y-auto: Using Automobile mode.

BIG DATA ON A SMARTPHONE

<u>Cellphone Location</u>
Location data Mandated by E-911. Can provide user locations, travel times, travel routes, etc.



<u>Crowdsourcing</u>: <u>"WAZE"</u>:
User contributed information: incidents, congestion, transit delay, facility performance, cyber attacks etc.

<u>Social Network</u>
"Twitter, Facebook"
Rich content data.
Social updates may include user activity, social events, user interaction, user satisfaction/complaints, "tagged" user locations

<u>Location-based Social Networking</u>
"Foursquare, Twitter, FB"
Geo-tagged social network messages: ***checkins***.
Announcing the arrivals at points of interests (e.g. office, restaurants, bars, coffee shops, transit terminals, transit lines. Provide confirmed trip time/ destination/ purpose info.

E911 WIRELESS LOCATION TECHNOLOGIES

Technology	Network	Handset/Network	Location Accuracy	E-911 Compliance
Cell ID	All Networks	Network	100m-3km	Phase 1
Cell ID + TA	GSM	Network	500 m	Phase 1
Cell ID + RTT	UMTS	Network	16-450 m	Phase 1
AFLT	CDMA	Network	200-400m	Phase 1
EFLT	CDMA	Network	250-300 m	Phase 2
TDOA, AOA, TOA	All Network	Network	100 - 200 m	Phase 2
U-TDOA	All Network	Network	50m	Phase 2
E-OTD	GSM	Network	50m	Phase 2
AGPS, GPS, GPS Hybrids	All Networks	Handset and Network Hybrid	5 - 30m	Phase 2
Wi-Fi AP	All Networks	Handset	indoor: 3-10 m/ outdoor 20-50 m	Phase 2, NG E911
Bluetooth	All Networks	Handset	3-10m	NG E911

http://www.ipcgps.com/uploads/docs/Intro_to_Location_Technologies-1.pdf

CELLULAR PROBE DATA PROVIDERS

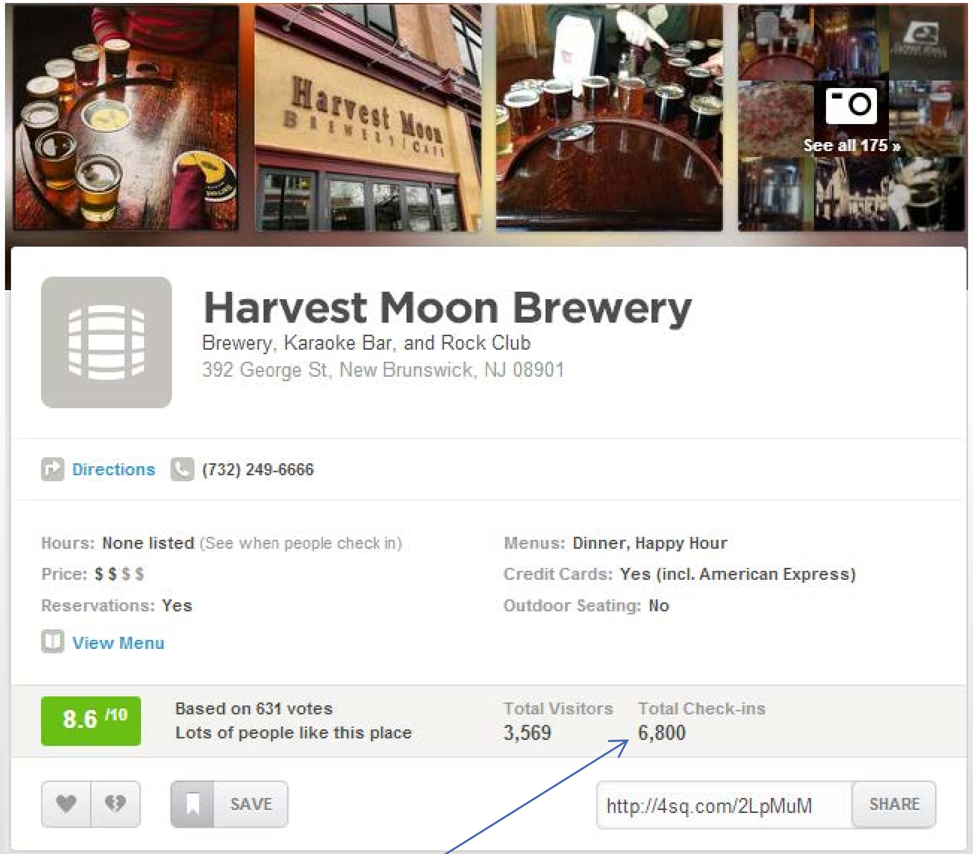
Industry Name	Country	Carrier Partnership	Operation Time	Coverage	Handset /Network
ITIS*	U.K.	Vodafone (U.K.), O2(U.K.), Telefónica (Spain)	1997	United Kingdom, Mainland Europe, the United States, Israel, and internationally	Network
Globis	Canada	Bell Mobility (Canada)	1998-2013	Canada and United States	Handset (A-GPS)
IntelliOne**	USA	U.S. Wireless, Rogers Wireless (Canada)	1999	Canada and United States	Network
Applied Generics ***	U.K.	Vodafone (Netherlands) AT&T (USA)	1999	The Netherlands, United States, Canada and Mexico	Network
AirSage	USA	Sprint (USA), Verizon (USA)	2000	USA	Network/ Handset
CellInt	Israel	Cellcom (Israel)	2005	US, Europe and Middle East	Network
MeiHui	China	China Mobile, China Unicom, China Telecom (China)	2004	Shanghai, et al	Network
Nokia	Finland	AT&T（USA)，T-mobile (USA)	2008	San Francisco and the Bay Area	Handset

* ITIS was acquired by INRIX in 2011.
** IMS (Intelligent Mechatronic Systems) has acquired IntelliOne in 2011
*** TomTom acquired Applied Generics in 2006.

CHALLENGES WITH CELLPHONE LOCATION DATA

- **Benefits**: Large and real-time spatial-temporal coverage, Route tracking, Large penetration rate
- **Accuracy**: Positioning error
- **Context**: Unconfirmed origin-destinations
- **Availability**: Need strong partnership with wireless carriers
- **Privacy**: User Consent, Snowden events

FOURSQUARE



Real-time Arrival Counts

Check-in Information

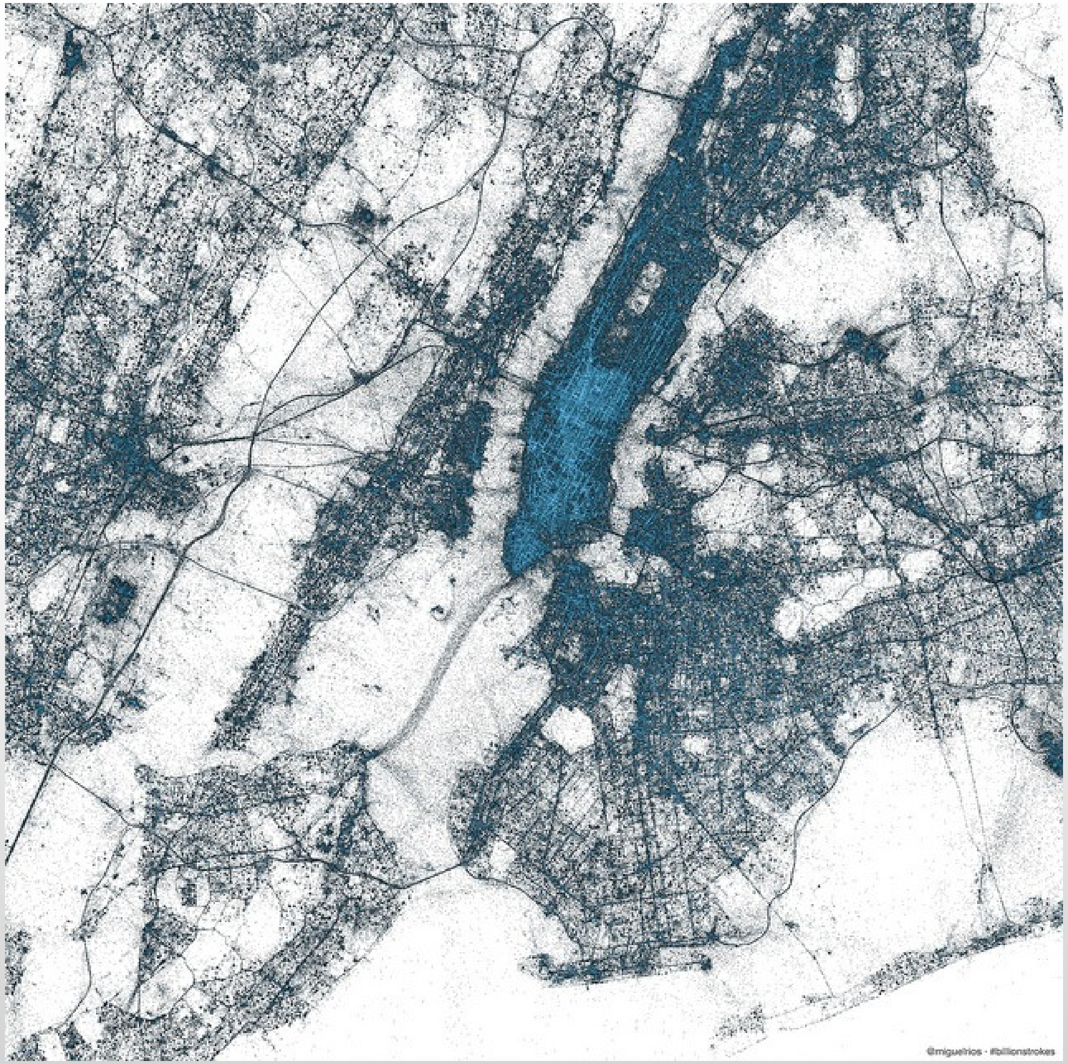
Field
id
type
timeZoneOffset
createdAt
private
shout
user

venue
location
source
event
photos
comments
likes
overlaps
score

Confirmed Trip Purpose through content!

Venue-side data public and no privacy issue | Two-hour frequency

GEO-TAGGED TWITTER DATA



Top Geo-tagging Sources on Twitter: Foursquare, Instagram, etc.

FOURSQUARE PULSE AND ACTIVITIES

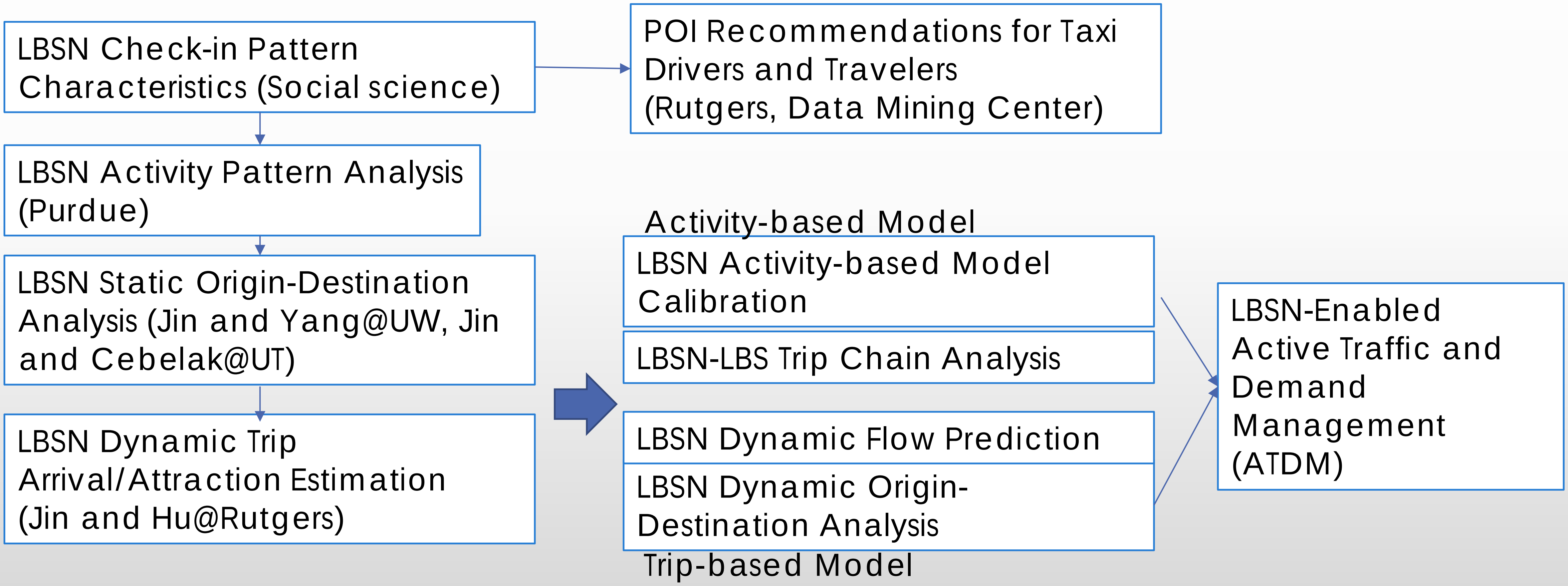
Spatial-Temporal Pattern of Urban Travel Activities

Travel Mode Information (Ferry, Transit, Tunel, etc.)

WHAT CAN WE DO WITH THE CHECK-IN DATA?

- **Travel demand information:**
  - Confirmed destinations, Accurate positioning
  - Real-time check-in patterns at venues
  - Inferring Origin-Destination Information
  - Integration with location data
- **Limitations**:
  - Activity sampling bias
  - Population sample bias
  - Lack of tracking: Only a fraction of open-data (Foursquare-twitter Bridge) for tracking and tracking is incomplete

LBSN RESEARCH ROADMAP



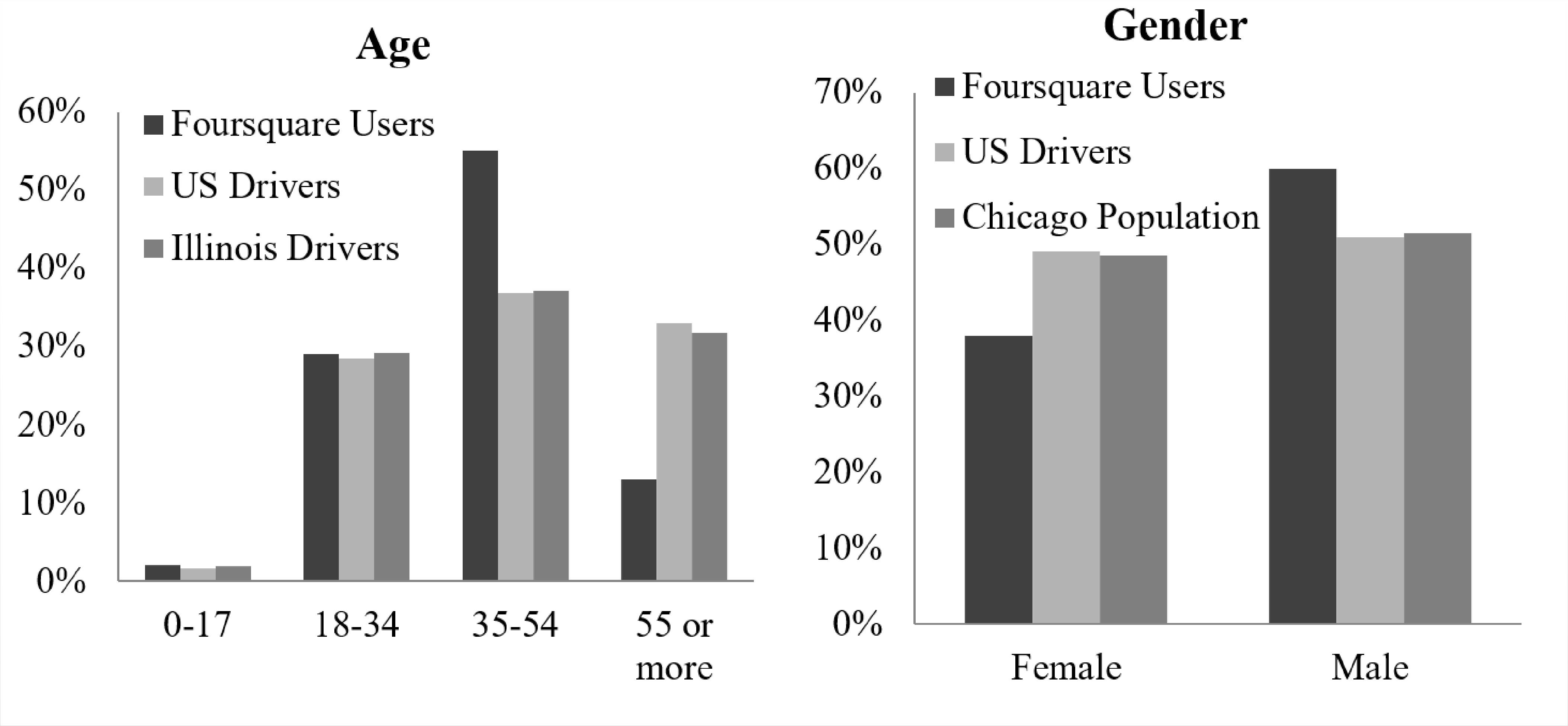
RESEARCH DATASETS

- LBSN Check-in through Foursquare Venue API
  - Bi-hourly check-in snapshots at over 5000 venues in Chicago and Austin, One month.
  - GNIP Twitter Foursquare and Geo-tagged data: Austin, Chicago, and NYC (pending)
- LBSN Firehose Data
  - Real-time global check-in feeds, One-year.

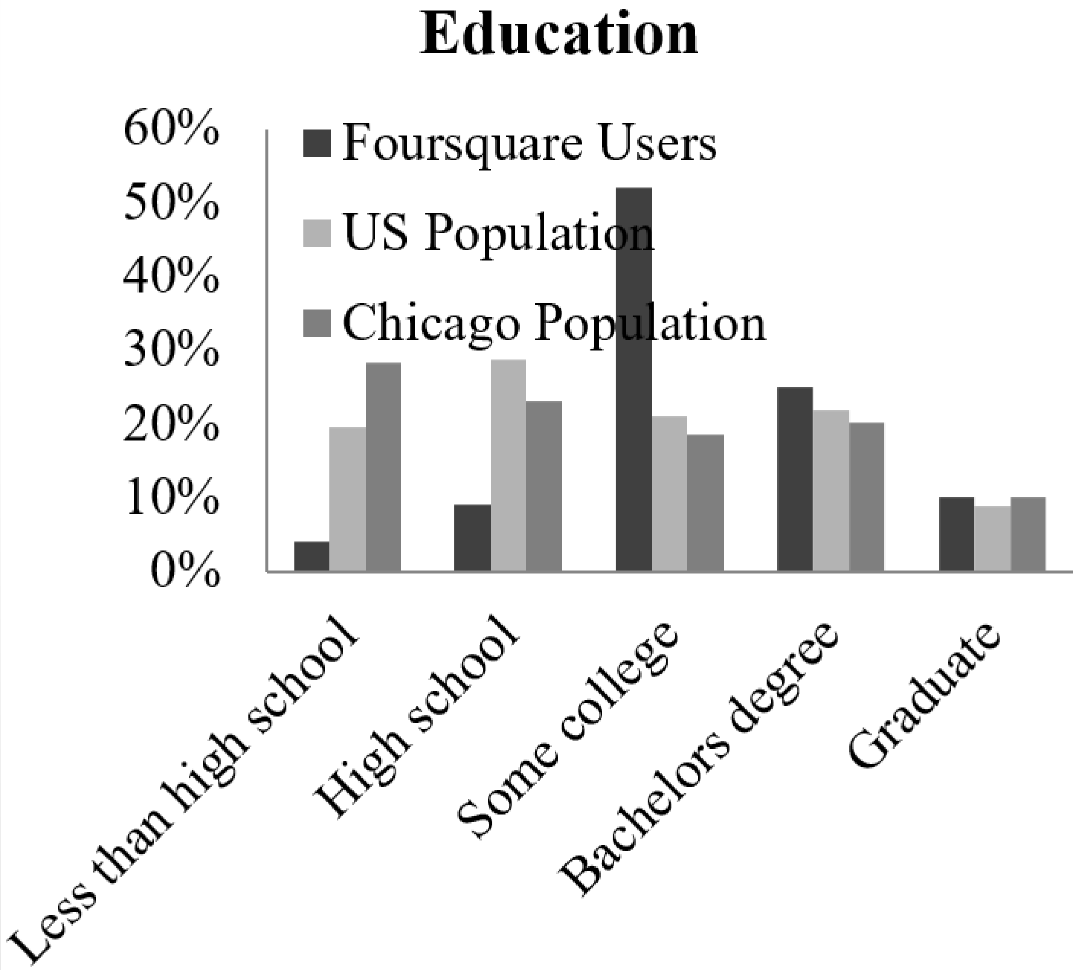
PUBLICATIONS

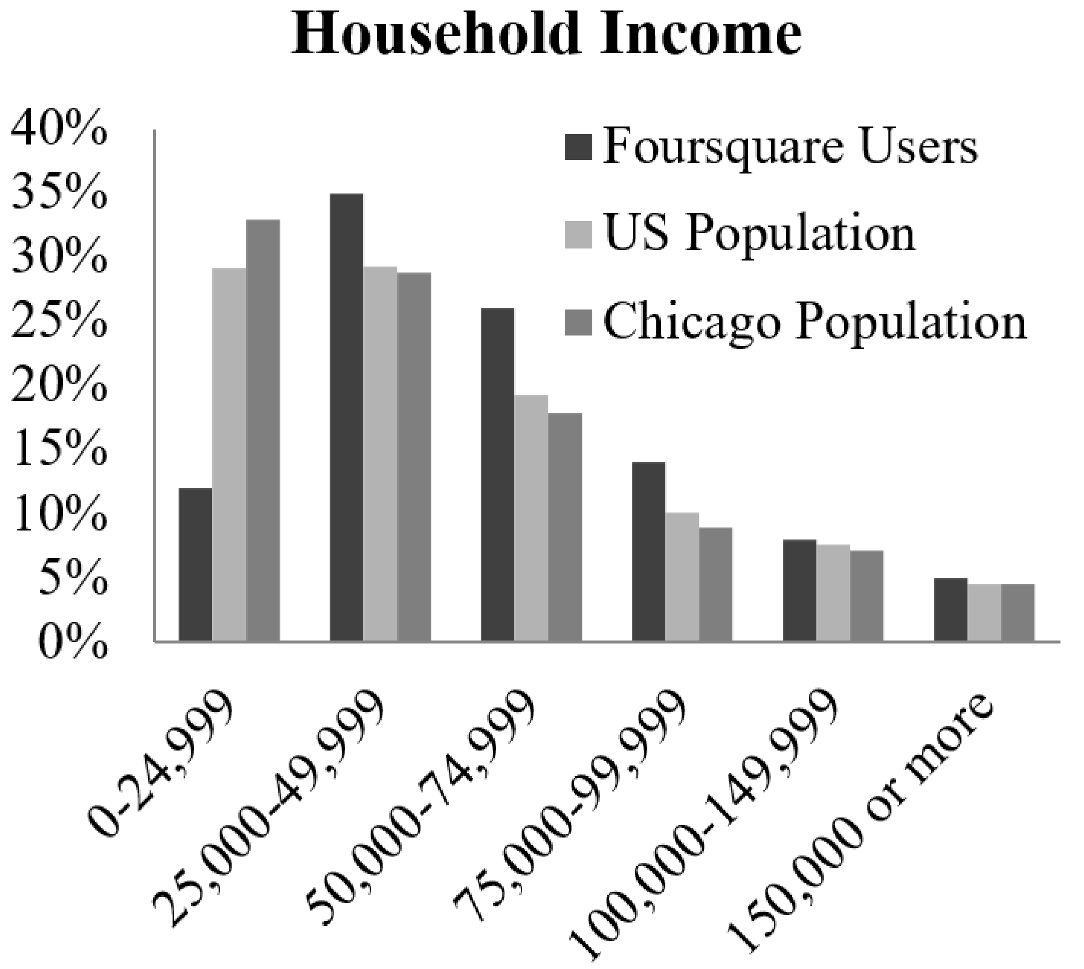
- Journal Publications
  - **F. Yang**, **J. Jin**, Y. Cheng, and B. Ran, *Origin-Destination Estimation for Non-Commuting Trips Using Location-based Social Networking Data*, International Journal of Sustainable Transportation, 9(8), 551-564, 2015
  - **P. J. Jin**, M. Cebelak, F. Yang, J. Zhang, C. M. Walton, and B. Ran, *Location-Based Social Networking Data: Exploration into Use of Doubly-Constrained Gravity Model for Origin-Destination Estimation*, Transportation Research Record, 2430(8), 72-82, 2014
  - **M. Cebelak**, **P. J. Jin**, and C. M. Walton, Transportation Planning Through Peer-to-Peer Modeling, 16-4531, TRB 95th Annual Meeting, January 2016.
  - **W. Hu**, and **P. J. Jin**, Adaptive Hawkes Process Formulation for Estimating Urban Trip Attraction with Location-Based Social Networking Data, 16-4766, TRB 95th Annual Meeting, Washington D.C., January 2016.
- Book chapter:
  - F. Yang, J. Jin, M. Cebelak, C.M.Walton, B. Ran, The Application of Venue-Side Location Based Social Networking (VS-LBSN) Data in Dynamic Origin-Destination Estimation, “Mobile Technologies for Activity-Travel Data Collection and Analysis”, Editor: Rasouli & Harry Timmermans, IGI Global.
- Working Paper:
  - W. Hu, P. J. Jin, The Anti-Sensing Model for Urban Travel demand Estimation with Location-based Social Network (LBSN) Data, ISTTT/Trans. Res. Part C

DEMOGRAPHICS OF FOURSQUARE



DEMOGRAPHICS OF FOURSQUARE



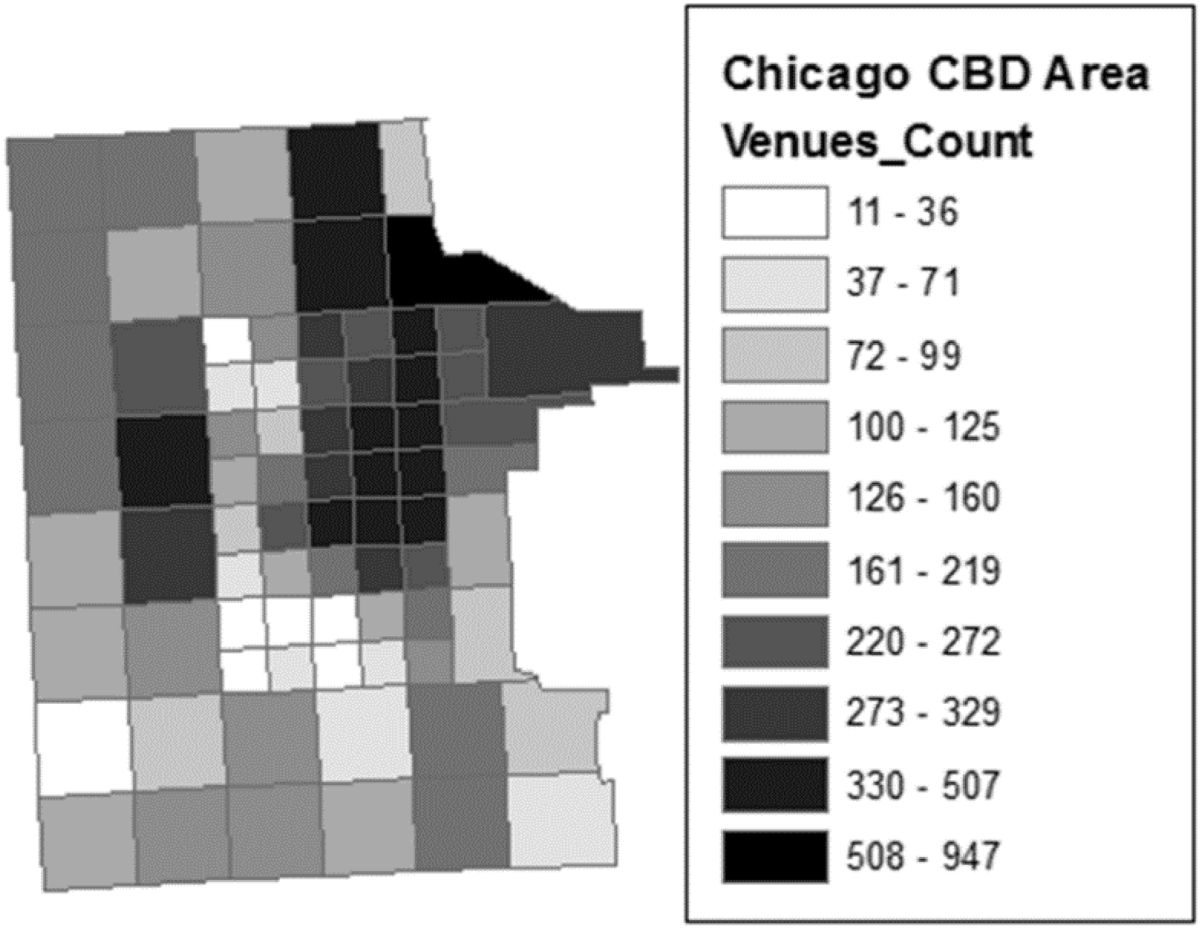


STATIC ORIGIN-DESTINATION ANALYSIS

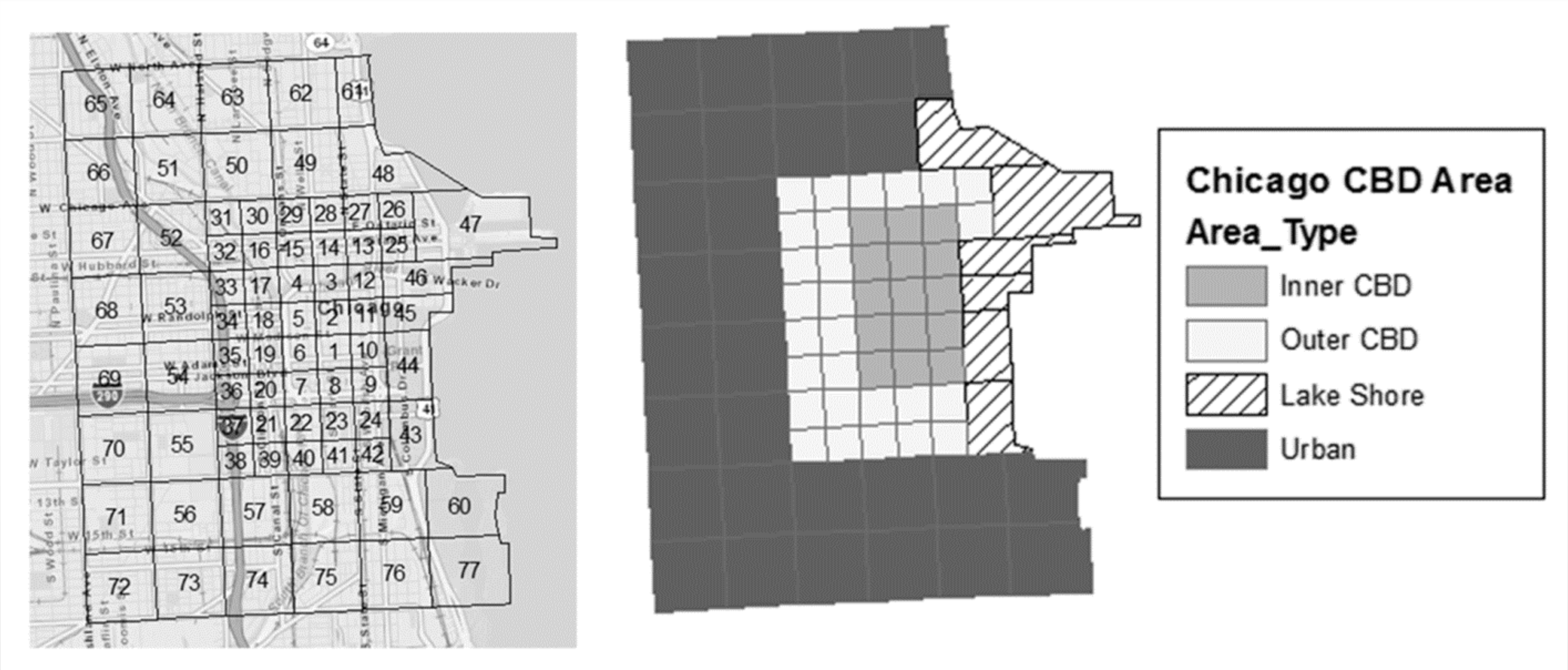
- Motivations: Start with the most observed LBSN venue categories for static travel demand analysis
- Methodologies: Clustering-based Sampling + Singly-Constrained Gravity model
- LBSN Data:
  - Bi-hourly Check-in Counts in Chicago Area, 16021 venues, June 19, 2011 and July 9, 2011
  - Bi-hourly Check-in Counts in Austin Area,
- Reference Data:
  - 2010 CMAP (Chicago Metropolitan Agency for Planning) OD Matrices

FOURSQUARE VENUE DISTRIBUTION

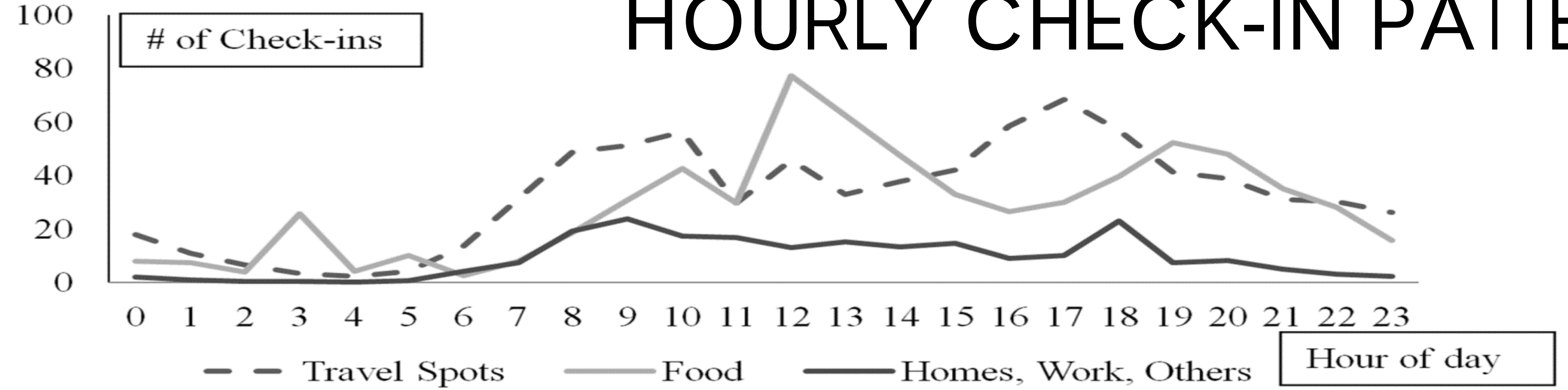
Chicago, IL



CHICAGO AREA TAZ



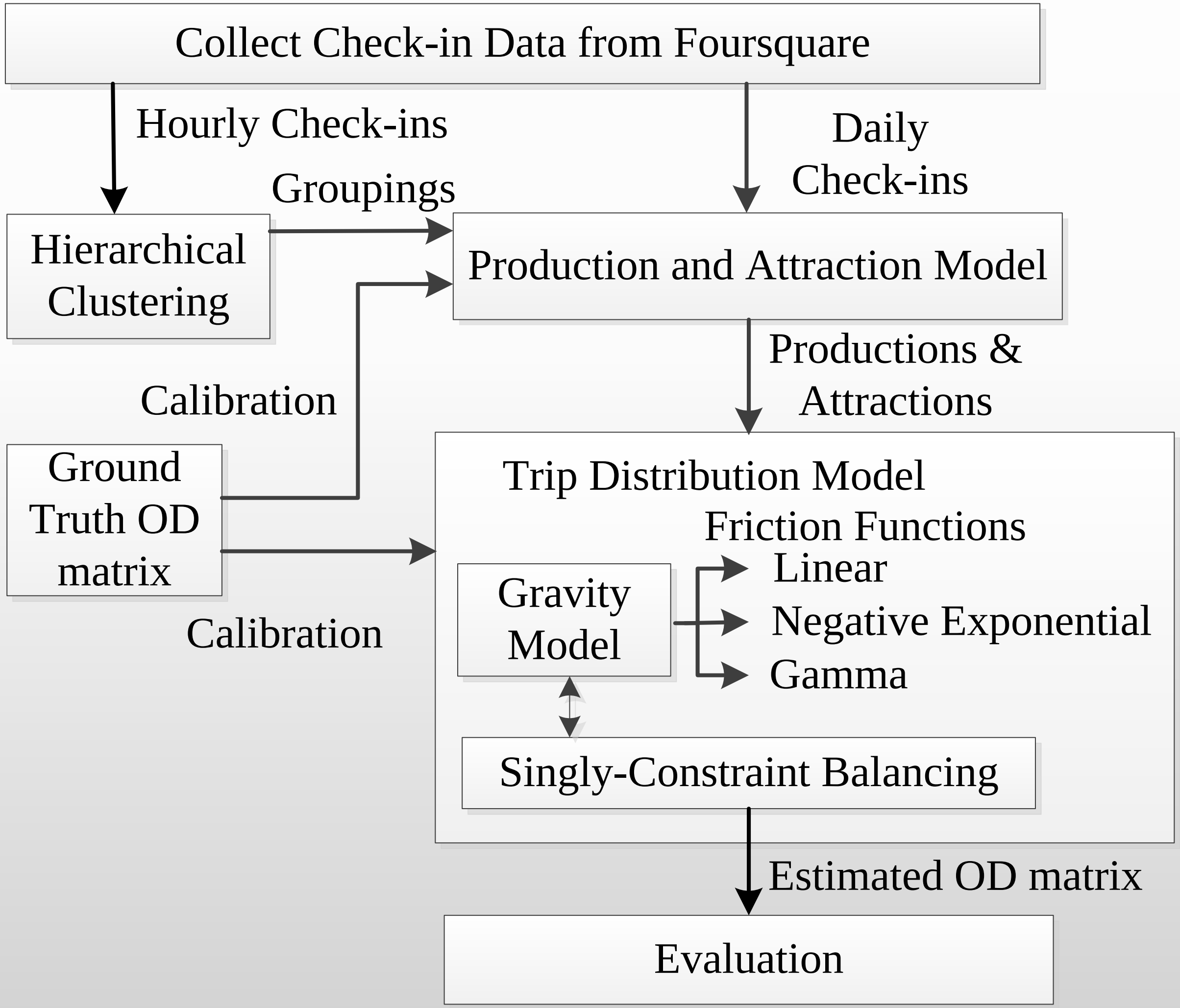
HOURLY CHECK-IN PATTERN



a

b

MODELING FRAMEWORK



Sample Estimation on Zonal Production and Attraction

Prod. And Attr. Calibration

Trip Distribution

OD Calibration

SAMPLING MODEL

- Production and Attraction Estimation
- $P_i = \sum_{k=1}^{K} p_k x_{ik} + p_0 ,\ \ i = 1,2, \dots , N$
- $A_j = \sum_{k=1}^{K} a_k x_{jk} + a_0 ,\ \ j = 1,2, \dots , N$

P_i: Trip production at origin zone *i*
A_j: Trip attraction at destination zone *j*
x_{ik}: Check-ins for venue type *k* in origin zone *i*
x_{jk}: Check-ins for venue type *k* in destination zone *j*
p_k, a_k: Coefficients for estimating the trip production/attraction contribution according to total check-ins for venue type *k*
N: The total number of TAZs
K: The total number of venue types
$p_{0,}$ a_0: The constant terms

RESIDUAL TERM

- Trip Conservation:

$$\sum_{i=1}^{N} \boldsymbol{P_i} = \sum_{j=1}^{N} \boldsymbol{A_j}$$

$$\sum_{i=1}^{N} \left(\sum_{k=1}^{K} \boldsymbol{p_k x_{ik} + p_0} \right) = \sum_{j=1}^{N} \left(\sum_{k=1}^{K} \boldsymbol{a_k x_{jk} + a_0} \right)$$

- Therefore,

$$a_0 = \frac{1}{N} \left[\sum_{i=1}^{N} \left(\sum_{k=1}^{K} p_k x_{ik} + p_0 \right) - \sum_{j=1}^{N} \left(\sum_{k=1}^{K} a_k x_{jk} \right) \right]$$

TRIP DISTRIBUTION

$$P_i = \sum_n p_n x_{in}\ , i = 1, 2, \ldots, 77$$

$$A_j = \sum_n a_n x_{jn}\ , j = 1, 2, \ldots, 77$$

$$\hat{T}_{ij} = P_i \frac{A_j F_{ij}}{\sum_j A_j F_{ij}}$$

Where

x_{in}: Check-ins for venue type n in origin zone i

x_{in}: Check-ins for venue type n in destination zone j

p_n: The fraction of non-commuting check-ins for venue type n in trip production.

a_n: The fraction of non-commuting check-ins for venue type n in trip attraction.

$\hat{T}_{ij}$: Trips made between origin zone i and destination zone j.

P_i: Production from zone i

A_j: Attraction of zone j

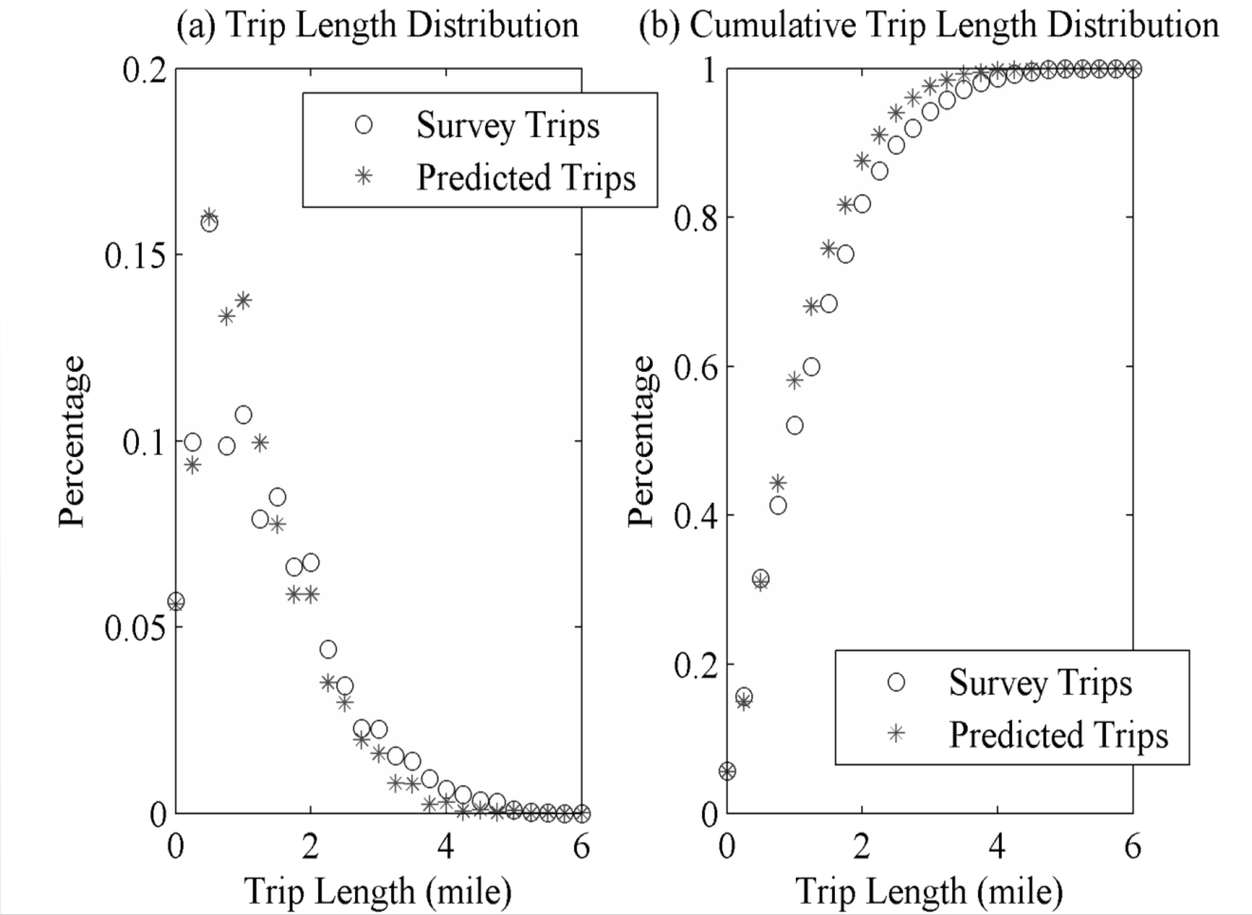
F_{ij}: Friction function

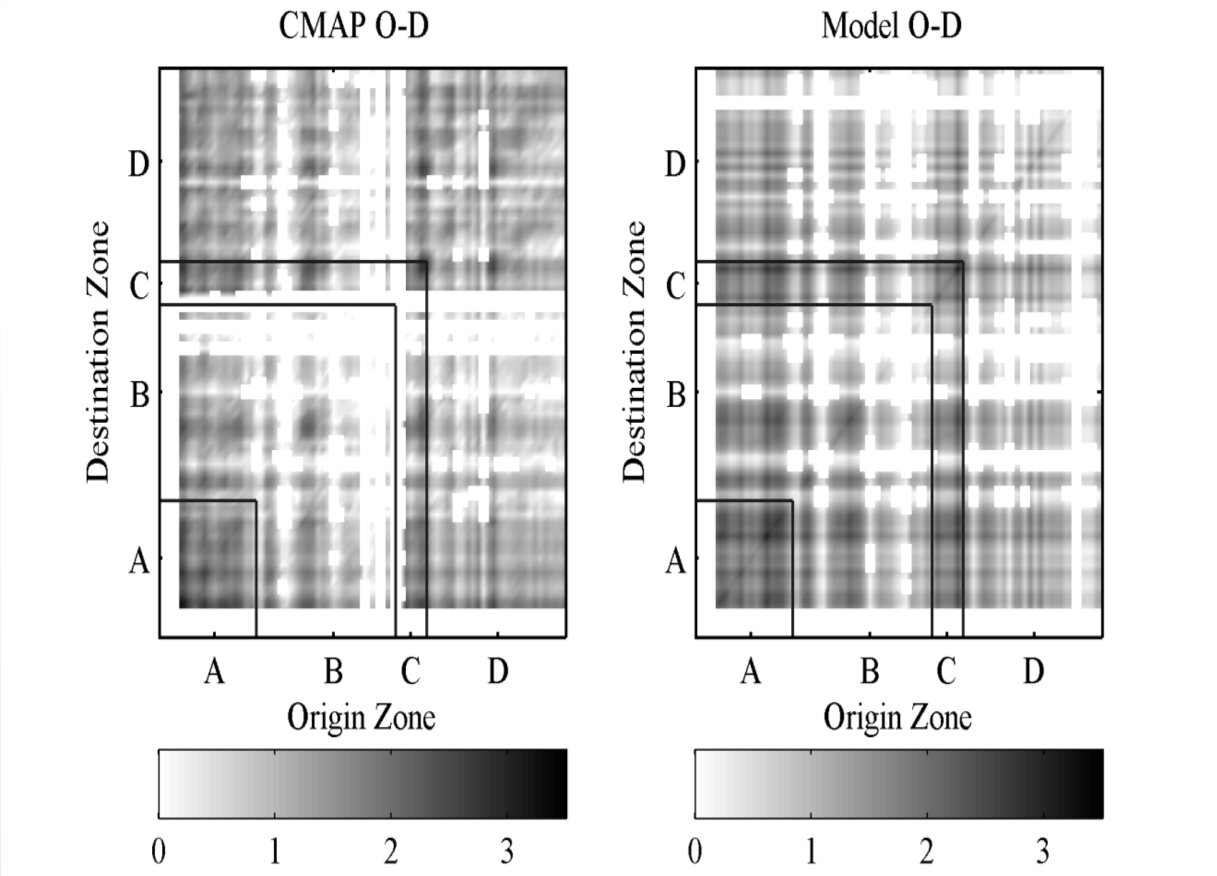
n=8	n=5	n=3	n=2	n=1
College & Univ.	College & Univ.	College & Univ.	College & Univ.	College & Univ.
Homes & Work	Homes & Work	Homes & Work	Homes & Work	Homes & Work
Art & Entertain.	Art & Entertain.	Art & Entertain.	Art & Entertain.	Art & Entertain.
Nightlife Spots	Nightlife Spots	Nightlife Spots	Nightlife Spots	Nightlife Spots
Shops	Shops	Shops	Shops	Shops
Food	Food	Food	Food	Food
Great Outdoors	Great Outdoors	Great Outdoors	Great Outdoors	Great Outdoors
Travel Spots	Travel Spots	Travel Spots	Travel Spots	Travel Spots

$$CR = \frac{\sum_i \min(p_i^M, p_i^O)}{\sum_i \max(p_i^M, p_i^O)}$$

Calibration Measure

MODELING RESULTS

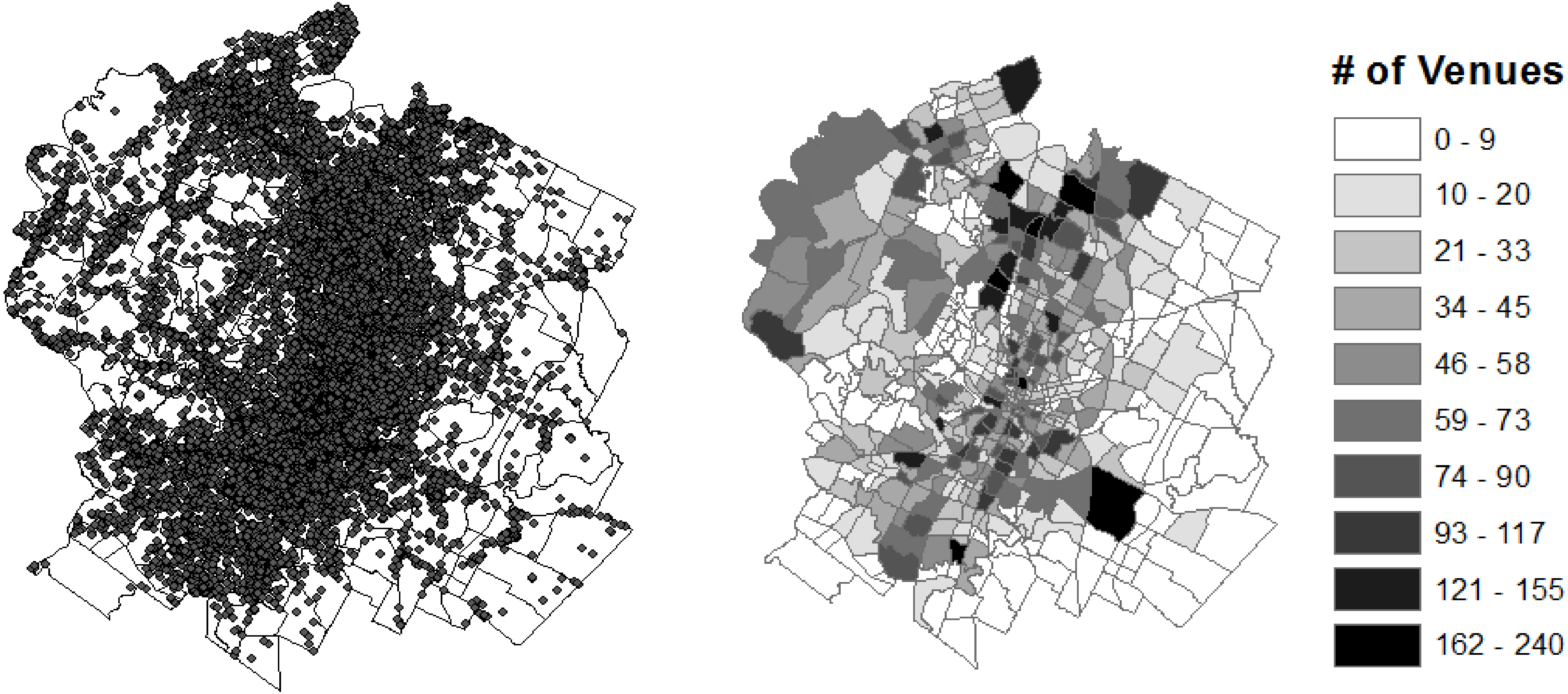




STATIC ORIGIN-DESTINATION ANALYSIS

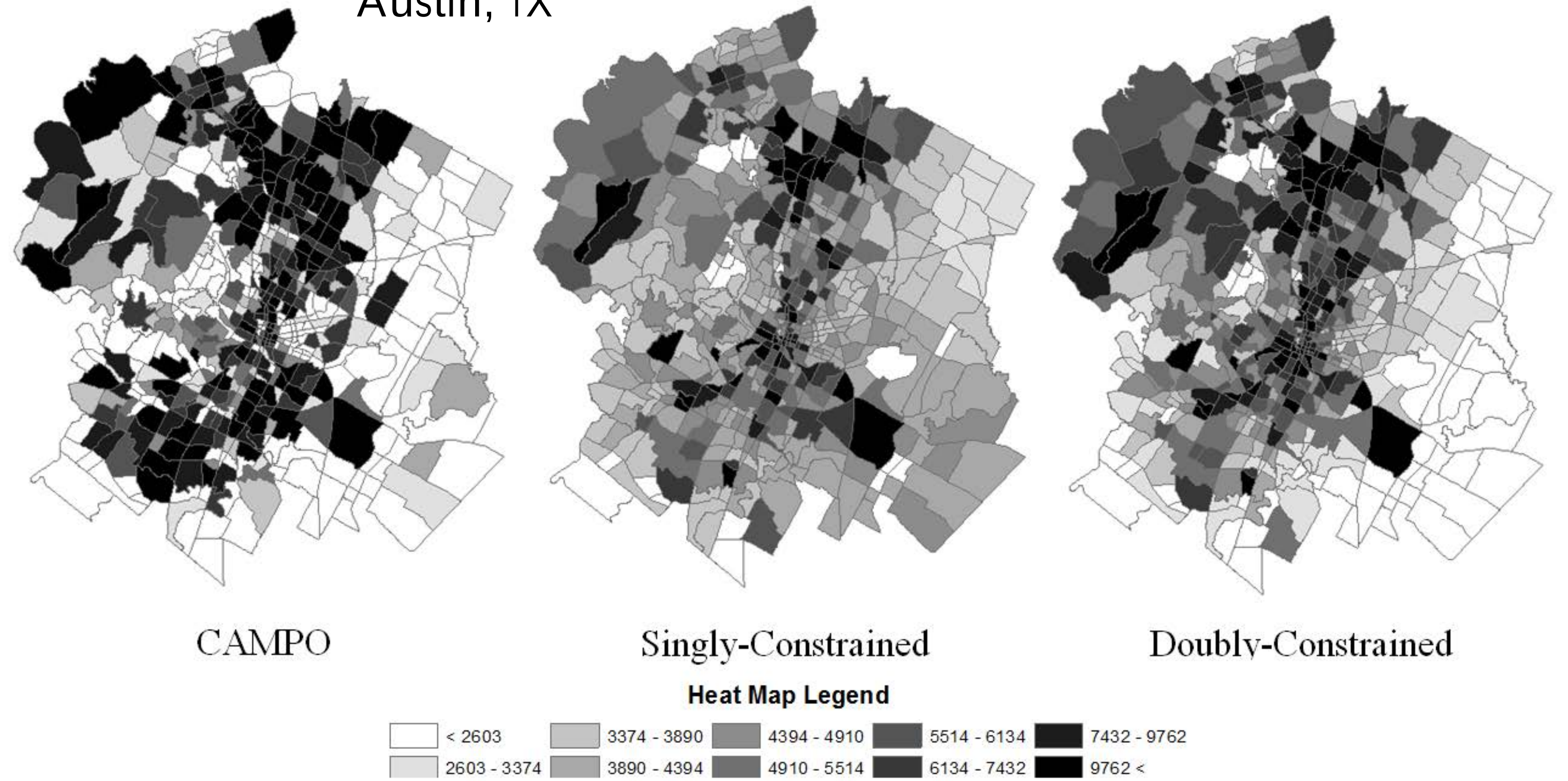
- Motivations: Improve OD Estimation
- Methodologies: Apply Locational-factors + Doubly-Constrained Gravity model
- LBSN Data:
  - Bi-hourly Check-in Counts in Austin Area, 19,170 venues,
- Reference Data:
  - 2010 CMAP (Chicago Metropolitan Agency for Planning) OD Matrices

FOURSQUARE VENUE DISTRIBUTION AUSTIN, TX



LBSN PRODUCTION RESULTS

Austin, TX

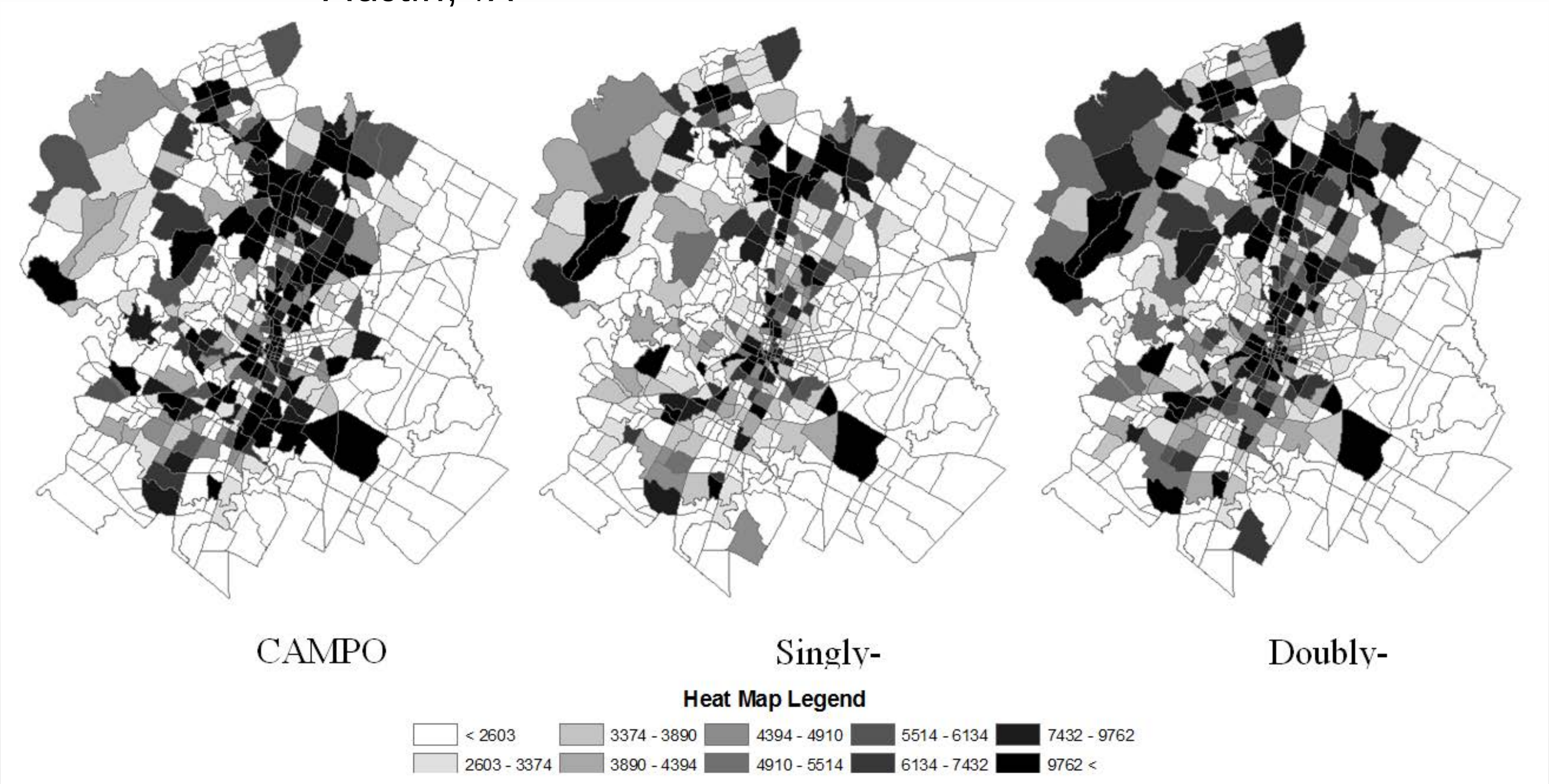


(a) Production Comparison Maps

- Source: [7] IJST 2014, [8] TRR 2014

LBSN ATTRACTION RESULTS

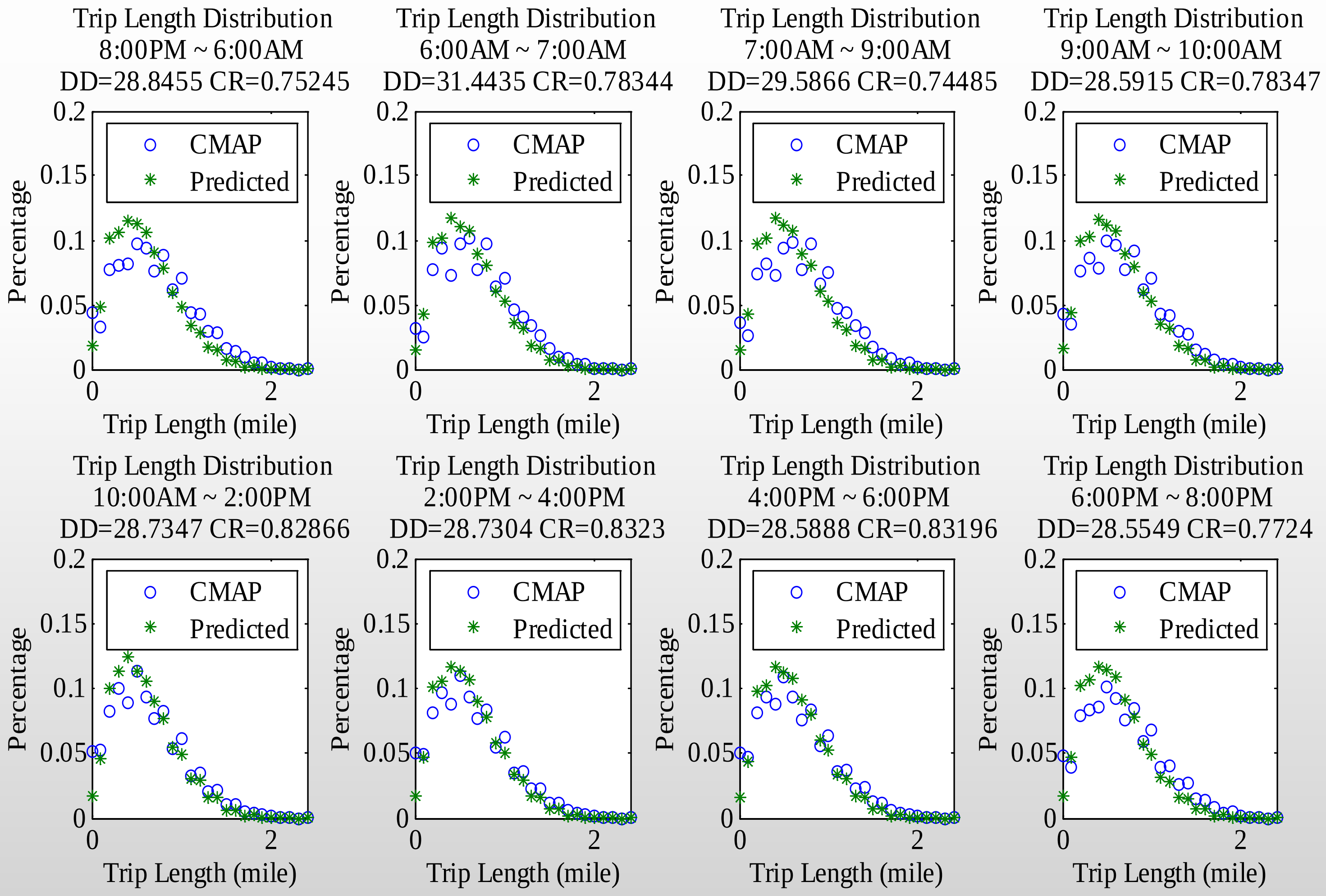
Austin, TX



(b) Attraction Comparison Maps

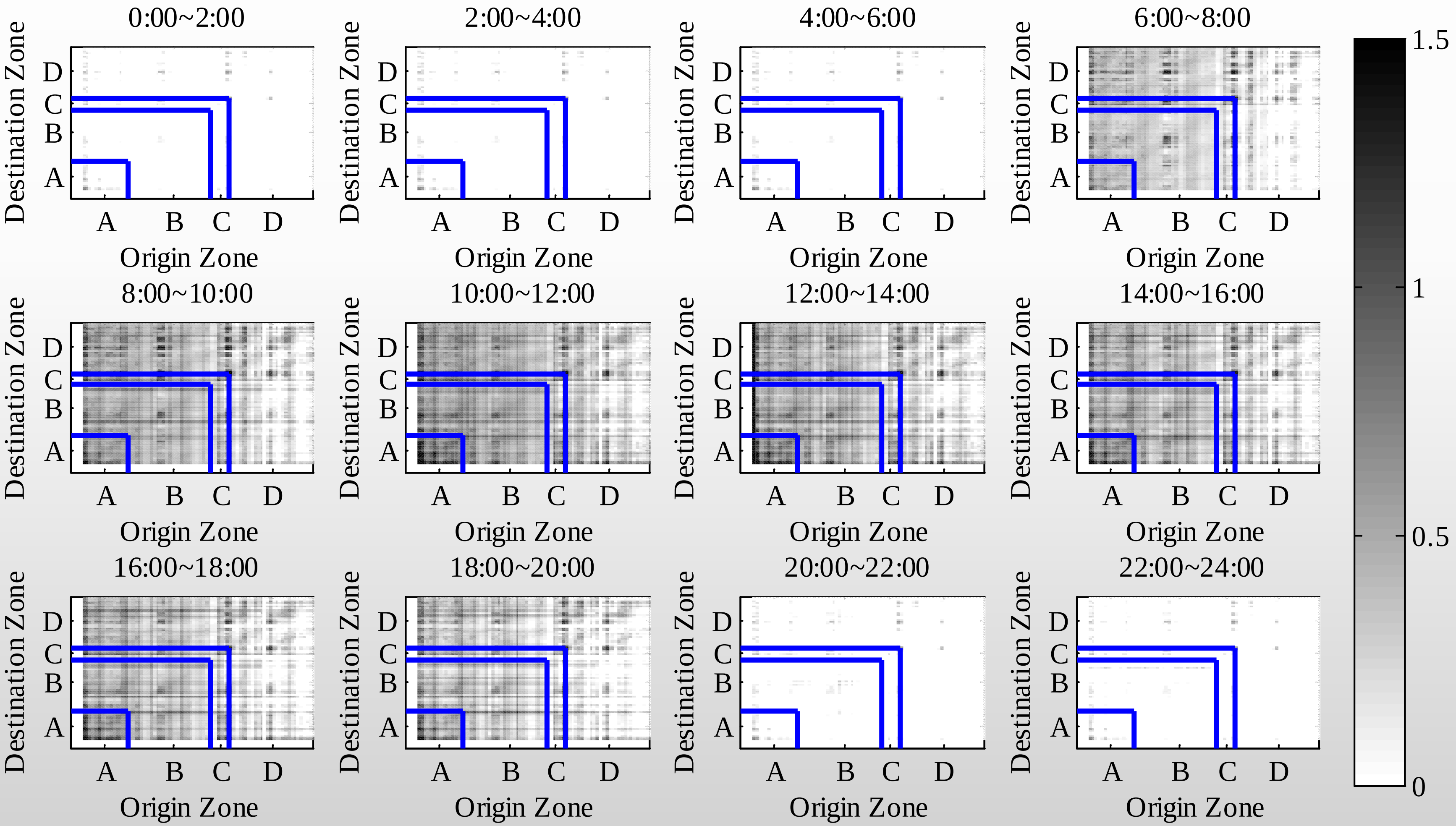
DYNAMIC OD ESTIMATION

- Apply similar methodologies to bi-hourly OD compare with MPO Time-of-day Factor Results



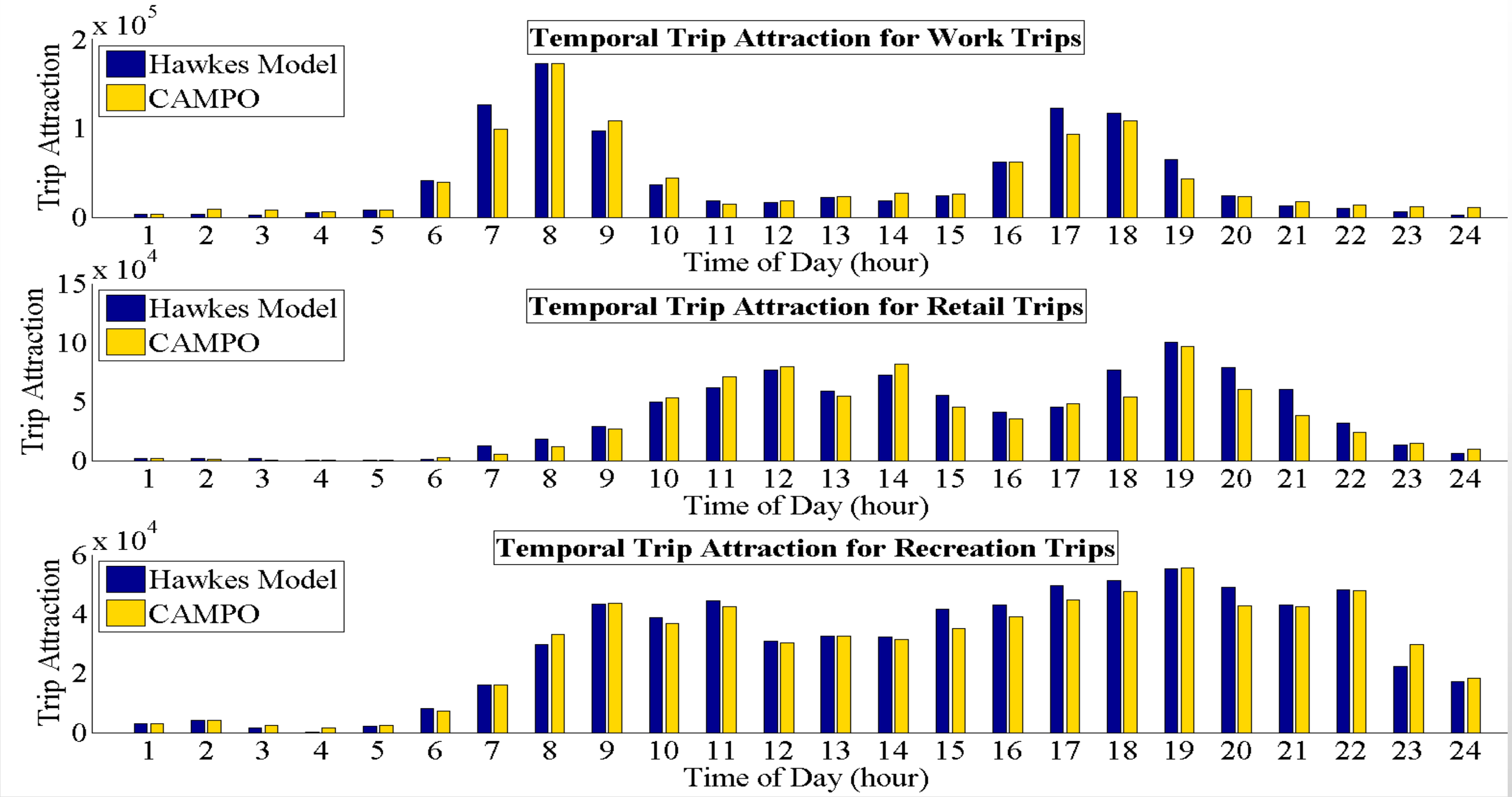
DYNAMIC OD ESTIMATION

- Apply similar methodologies to bi-hourly OD compare with MPO Time-of-day Factor Results



ADDING HAWKES PROCESS TO LBSN ESTIMATION

- Changing from **Uniform** to **Hawkes Random** Arrivals



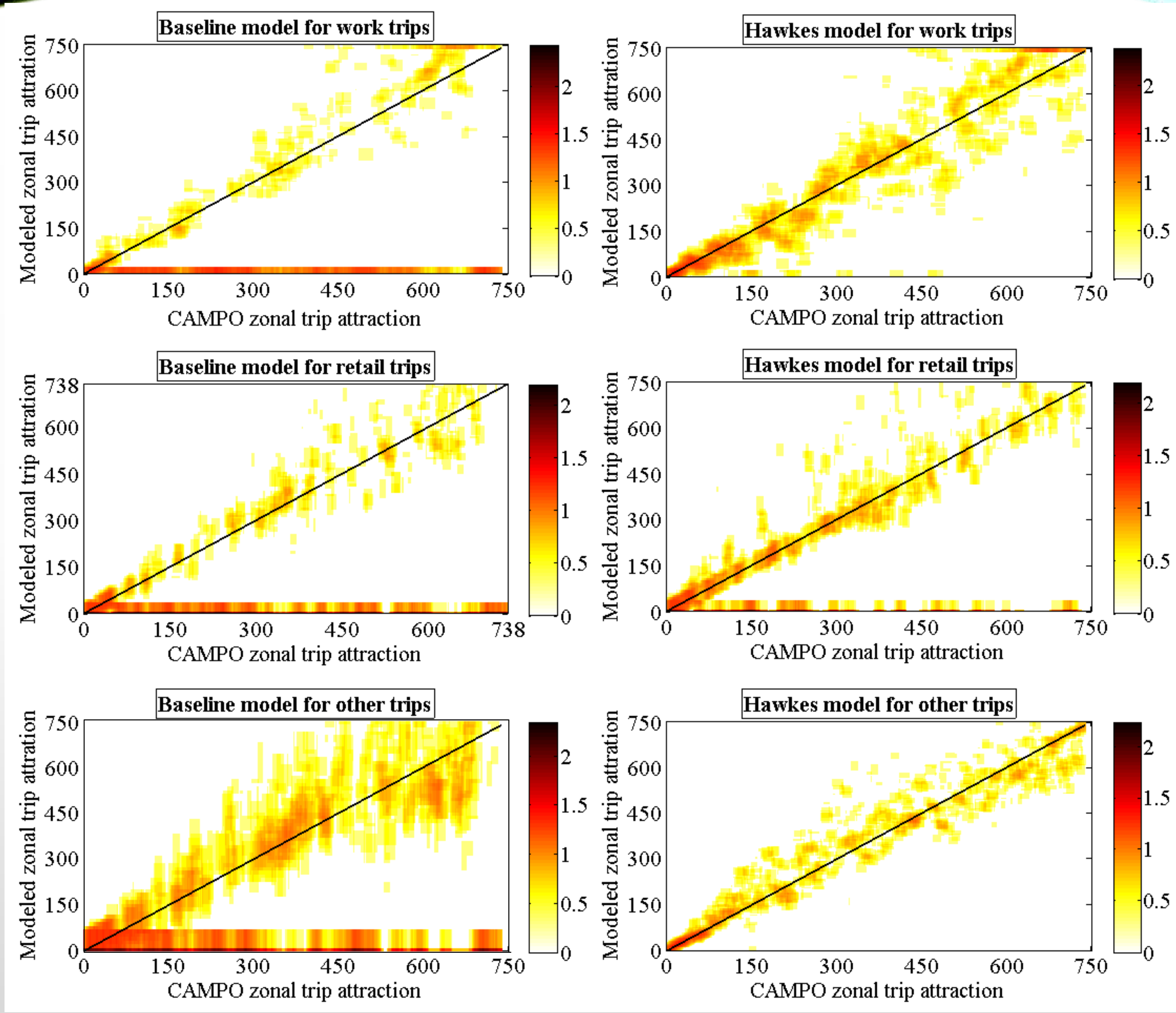
HAWKES MODEL PRODUCTION/ATTRACTION RESULTS

Reference Model: The previous simple random sampling: A = p*C

A: Attraction

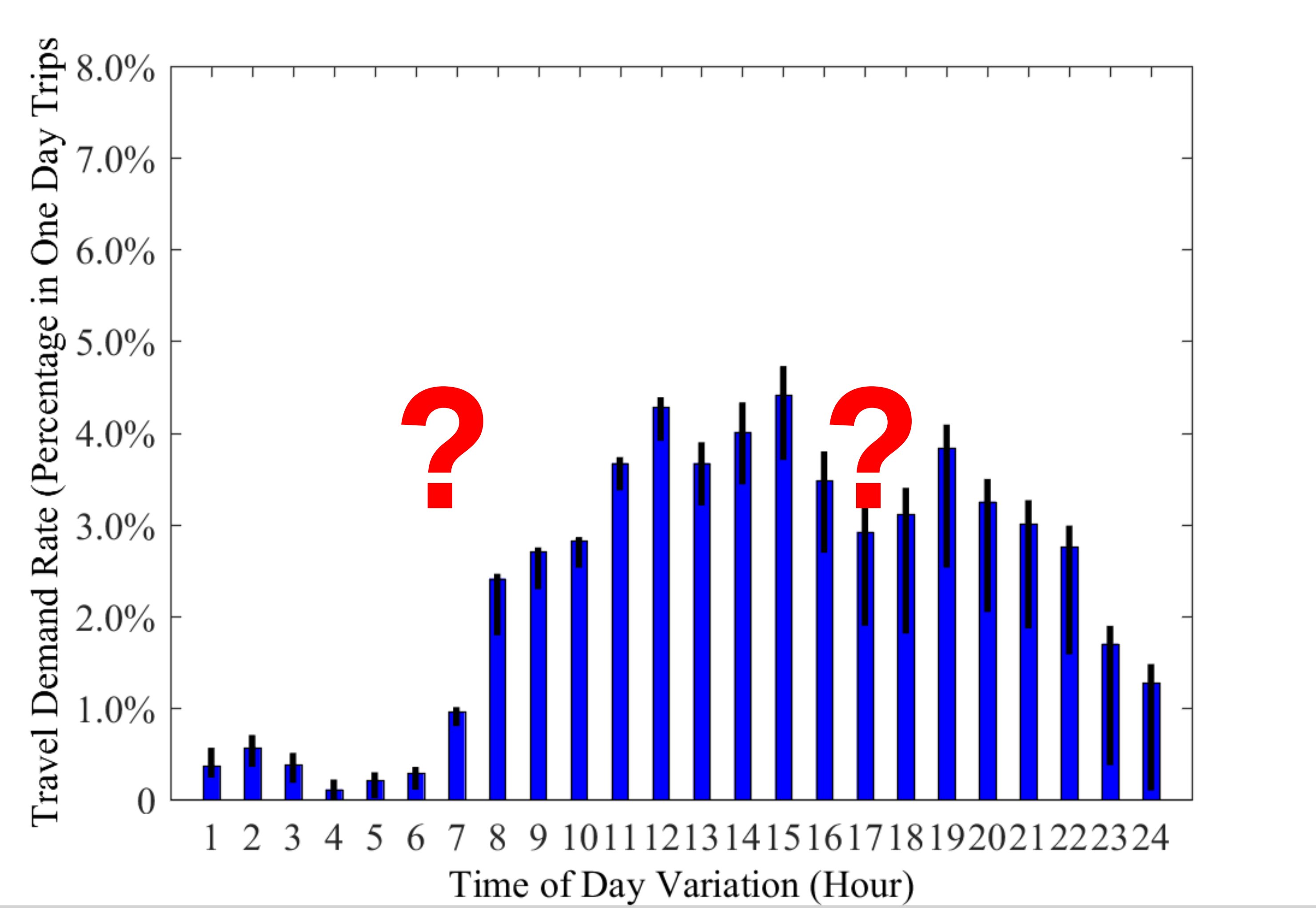
p: Scaling factor (assuming uniform arrivals)

C: check-in counts



KEY CHALLENGES

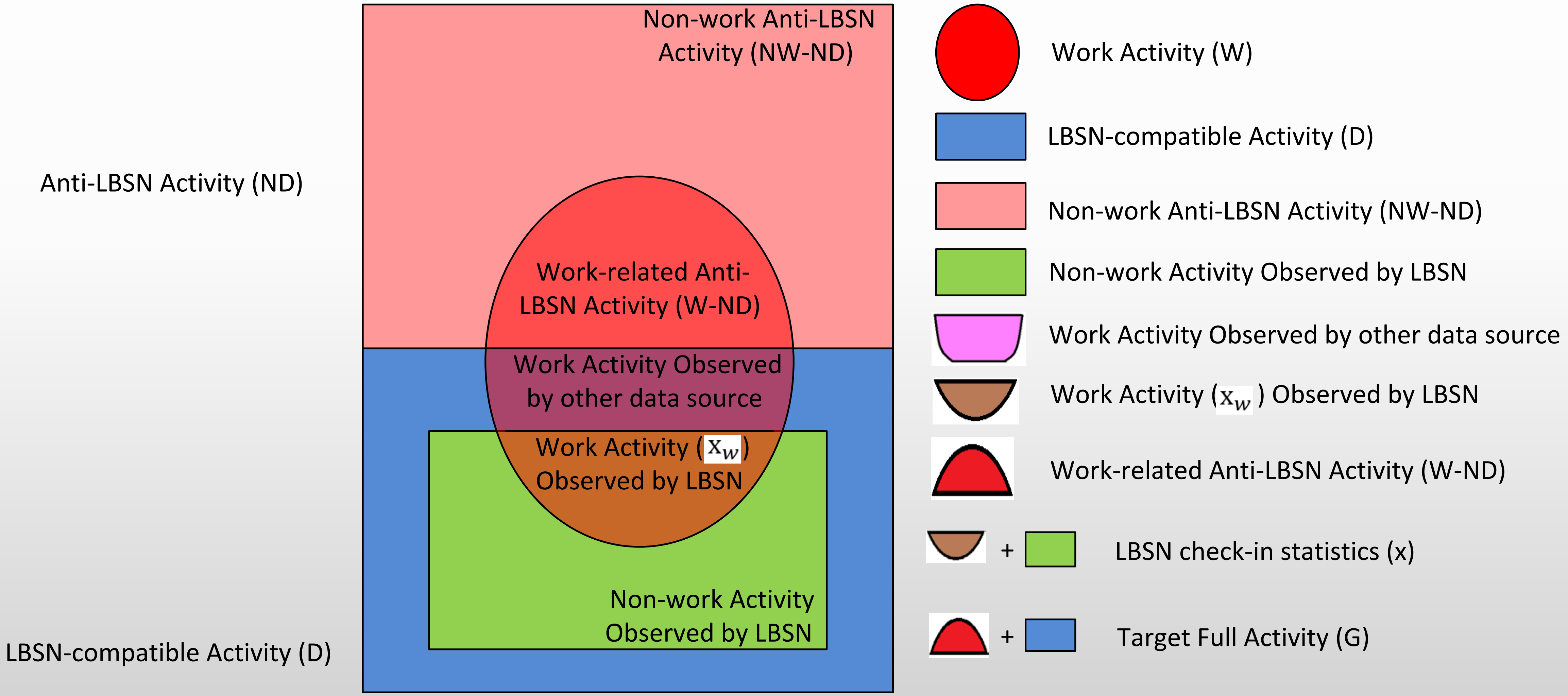
- LBSN Sampling: Not full social network activities
- Sampling bias especially for Home/Work Trips
- Dynamic Estimation is limited by hourly sampling rate and zone resolution.



COMPATIBILITY BETWEEN TRAVEL AND SOCIAL NETWORK ACTIVTIES

- Compatible definition:
  - A travel activity that can more likely to result in a check-in/social network event at the destination
- Compatible trip purposes:
  - Shops, restaurants, night life, outdoor activities etc.
- Incompatible trip purposes:
  - Work/Home (Commuting trips)
- Idea: Assuming a total activity limit for different time of day, compatible trips and LBSN activities shares the time frame (Direct Sensing) while incompatible trips exclude LBSN activities (Anti-Sensing, less LBSN activities => more trips)

ACTIVITY SET IN SENSING AND ANTI-SENSING MODEL (WORK TRIPS)



THE ANTI-SENSING MODEL FOR URBAN TRAVEL DEMAND

- **LBSN-compatible Activity Pattern Estimation Model**

$$D \sim Dist(f(\mathrm{x}_{t,w}, \mathrm{x}_t; \theta_d, \beta_d))$$

- The number of travel demand in a time interval $[t, t+\Delta t]$ is nonhomogeneous Poisson with mean

$$\mu = \int_t^{t+\Delta t} \lambda(\tau) d\tau$$

- Where $\lambda(\tau)$ is the intensity function

$$\lambda(\tau) = \theta_d * (\mathrm{x}_t - \mathrm{x}_{t,w}) + \beta_d$$

$$P(D_i|\mu_i) = \frac{e^{-\int_{t_i}^{t_i+\Delta t} \lambda(\tau) d\tau} (\int_{t_i}^{t_i+\Delta t} \lambda(\tau) d\tau)^{D_i}}{D_i!}$$

- Where x_t is the social media statistics, θ_d is the converting parameters, β_d is the bias factors (e.g. hourly pattern, location, trip type), and Δt is set of resolution as 15min to 1 hour.

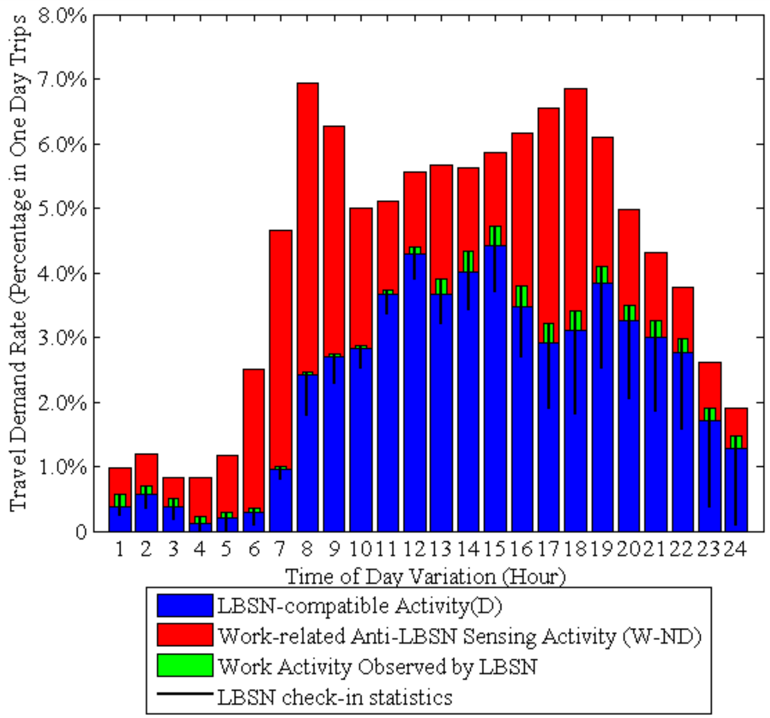
THE ANTI-SENSING MODEL FOR URBAN TRAVEL DEMAND

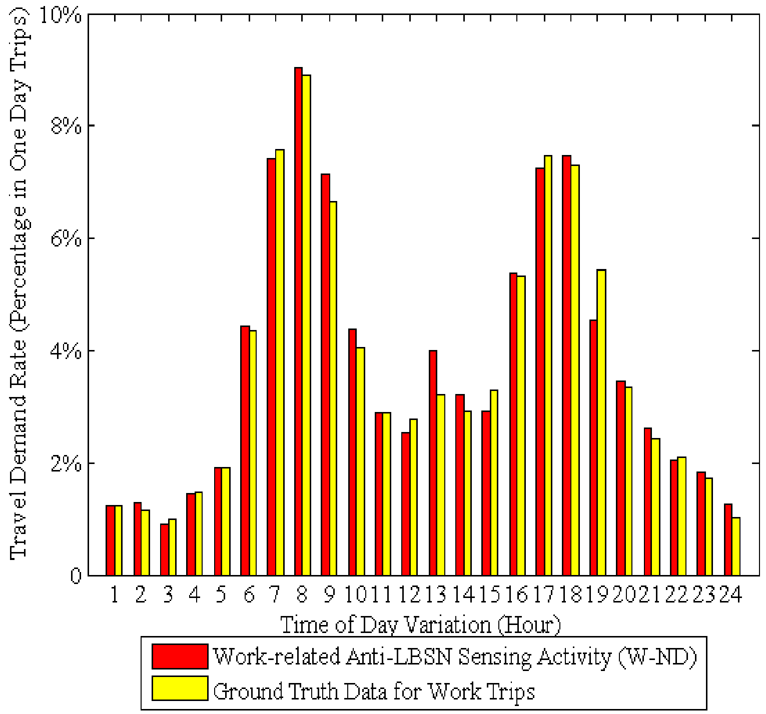
- **Work-related Anti-LBSN Activity Pattern Estimation Model**
- $W - ND \sim g(G, D)$
- Where $g(G, D)$ is a function of the work-related anti-LBSN activity demand $W - ND$ with estimated LBSN-compatible demand D, and full activity pattern G.
- In each time interval, the full activity pattern G has a fixed time budge regarding the human energy, attention, and multi-tasking capabilities.
- $\mathrm{P(W - ND) + P(D) = P(G)}$.
- $\mathrm{P}(W - ND_1 = w - nd_1, \ldots, W - ND_n = w - nd_n | \boldsymbol{\mu}) =$

$$\prod_{i=1}^{n} \mathrm{P}(W - ND_i = w - nd_i | \boldsymbol{\mu}) = \prod_{i=1}^{n} \left(1 - \frac{e^{-\int_{t_i}^{t_i+\Delta t} \lambda(\tau) d\tau} \left(\int_{t_i}^{t_i+\Delta t} \lambda(\tau) d\tau\right)^{d_i}}{d_i!}\right)$$

THE ANTI-SENSING MODEL PRELIMINARY RESULTS

- **PERLIMINARY RESULT ANALYSIS**





Left: The daily activity pattern by LBSN Sensing

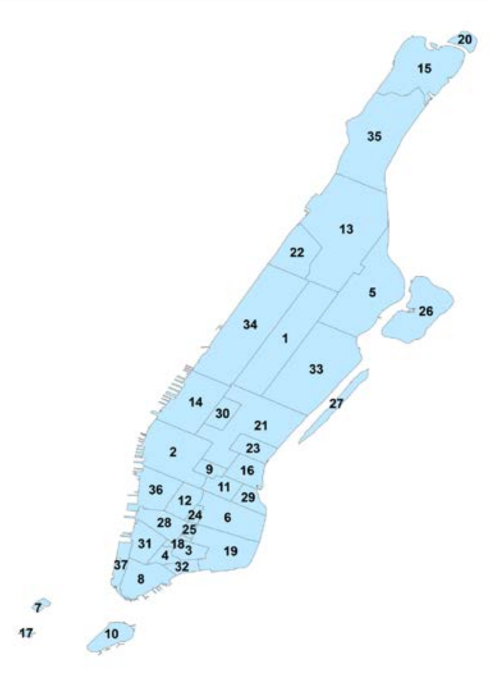
Right: Comparison between the estimated work-related activity pattern and ground truth data

LAND USE CORRELATION BASED ON LBSN

- **The cross-correlation-based method**

- The idea of the pattern recognition of urban travel demand through such technologies is examined.

- **Study area and dataset**



1	Central Park
2	Chelsea
3	Chinatown
4	Civic Center
5	East Harlem
6	East Village
7	Ellis Island
8	Financial District
9	Flatiron District
10	Governors Island
11	Gramercy
12	Greenwich Village
13	Harlem
14	Hell's Kitchen
15	Inwood
16	Kips Bay
17	Liberty Island
18	Little Italy
19	Lower East Side
20	Marble Hill
21	Midtown
22	Morningside Heights
23	Murray Hill
24	NoHo
25	Nolita
26	Randall's Island
27	Roosevelt Island
28	SoHo
29	Stuyvesant Town
30	Theater District
31	Tribeca
32	Two Bridges
33	Upper East Side
34	Upper West Side
35	Washington Heights
36	West Village
37	Battery Park City

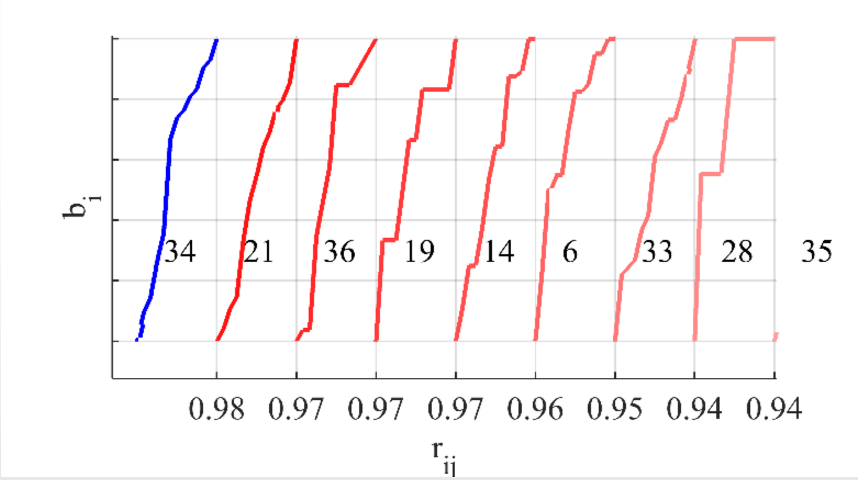
- Neighborhood index and land use snapshot for Manhattan Island, NYC.

- The data set includes one year of tweets posted within Manhattan island of New York City from 11:40 pm of February 25th, 2010 to 04:26 am of January 21st, 2011.

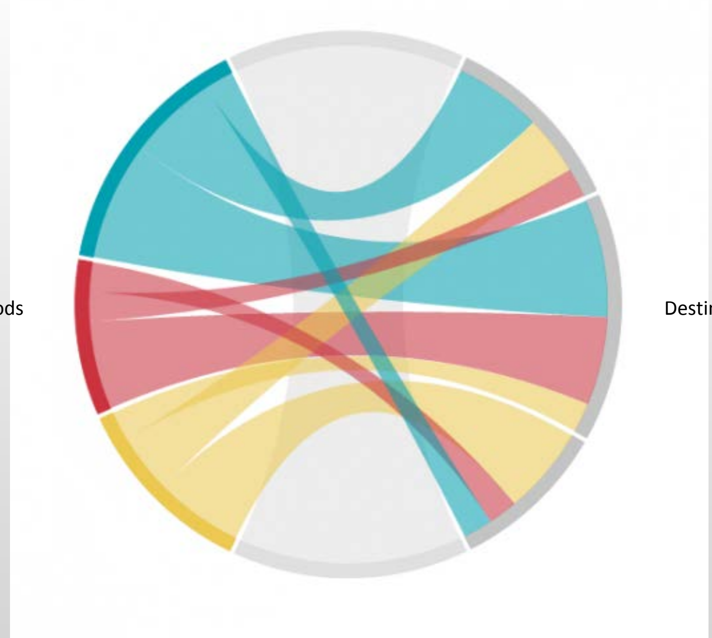
LAND USE CORRELATION BASED ON LBSN

- **Time delay correlation model**

$$r_{i,j}(t,w) = f_i(t,w) * f_i(t + \tau_{tr} + \tau_{dw}, w)$$



* Blue line represents the origin area and red represents the potential destination area.



* Sample of three OD pairs: colorful chords show the flow from the origin to the destination, the width indicate the flow value

POTENTIALS AND LIMITATIONS OF LBSN DATA

- Potentials:
  - Large-scale, High-Resolution Activity Data
  - Estimate static and dynamic travel demand
  - Integration with other Big Data Sources: Operations, Cellphone LBS, video, etc.
  - Integration with Activity-based and Trip-based models
- Limitations:
  - Individual tracking is incomplete
  - Changing in social network market