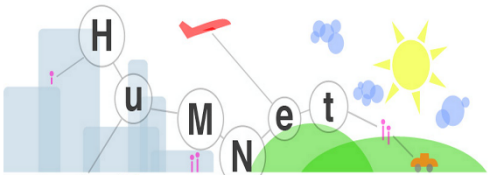
Big Data to tackle Urban Traffic

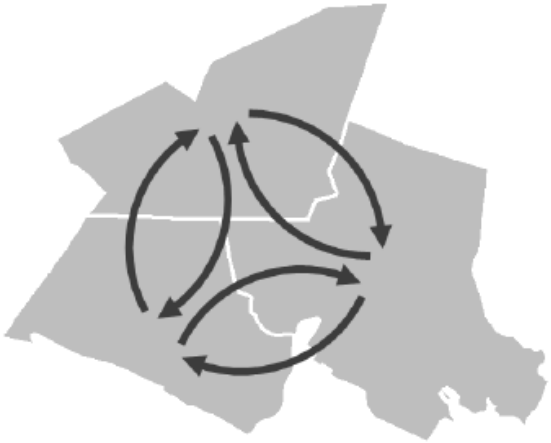
Marta C. Gonzalez

Massachusetts Institute of Technology



CITIES ARE NOW HOME TO HALF OF THE WORLD'S 7 BILLION HUMANS

Overview



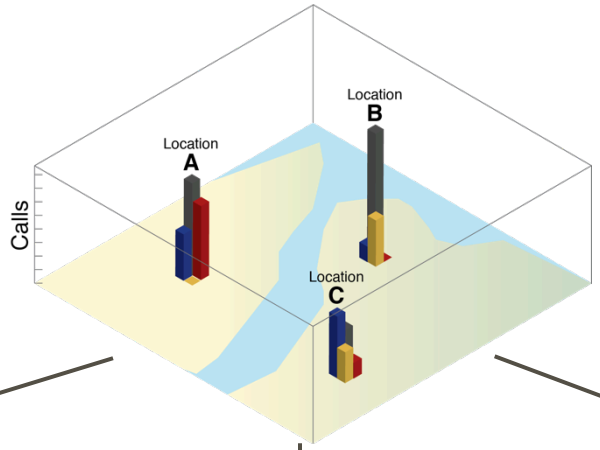
Knowledge
From Massive
& Passive Data

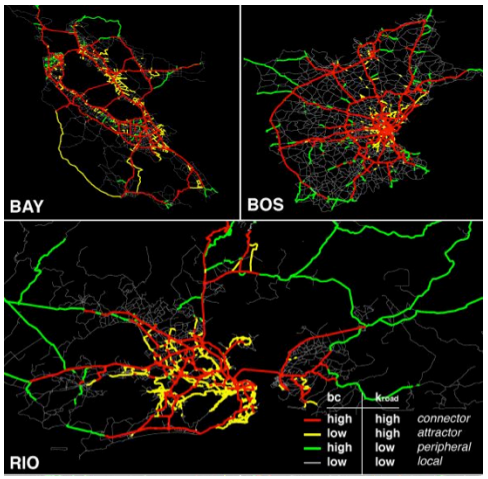
Information
To Users

Planning Tools
at Urban Scale

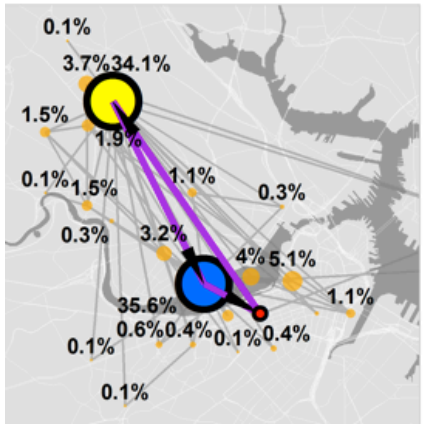
Today

Information and Communication Technologies (ICT) Passive Data

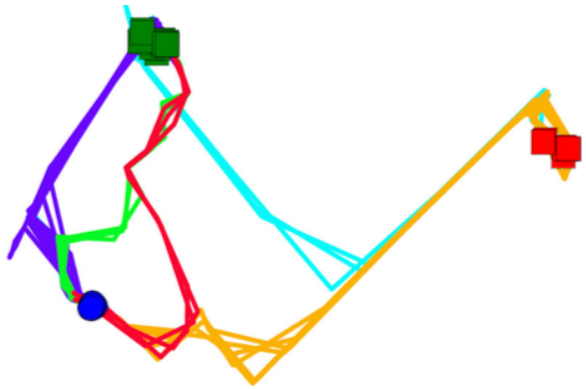




Urban Traffic



Individual Mobility



Route Choice Behavior

Overarching Goal

Opportunities:

Massive spatiotemporal information

- **millions of individuals in a given metro area**
- **long time period of observation (in months).**

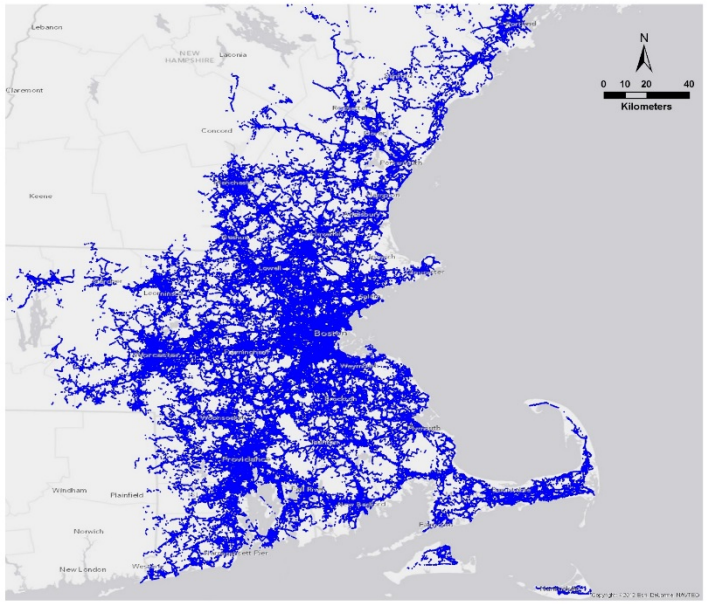
Obstacles:

Massive, and passive data with lots of noise

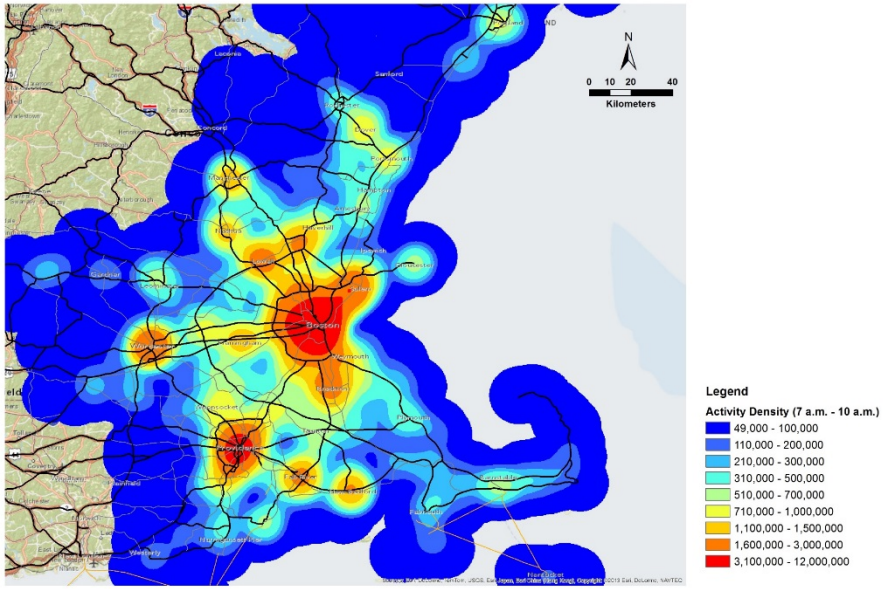
- **anonymity of individuals**
- **missing information**
- **no social demographic characteristics**
- **potentially biased sample**

Overarching Goal

How to extract human daily activities (e.g., types, sequences, and chains) from these massive, passive and noisy Big Data that are comparable to travel demand models from travel surveys? and asses the role of Social Routing?



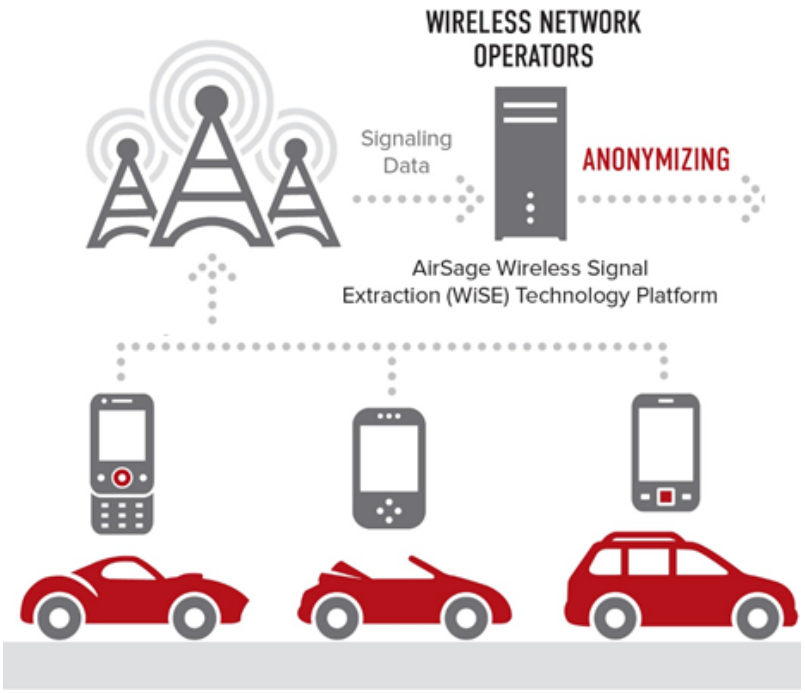
1.9 million total users observed in the 2 months, in Boston 2010.



Human Activity Density 4 P.M.-7 P.M.

Raw Data Description

Traces of People – Where and When



- **800 million of historical location records for 1 million anonymous individuals who use phones in the Boston metropolitan area**
- **Data for one anonymous user:**

Longitude	Latitude	Time
-71.059998	42.356132	1266513700
-71.059730	42.356391	1266513800
-71.063884	42.355315	1266513900
-71.063884	42.355315	1266514200
...	...	...

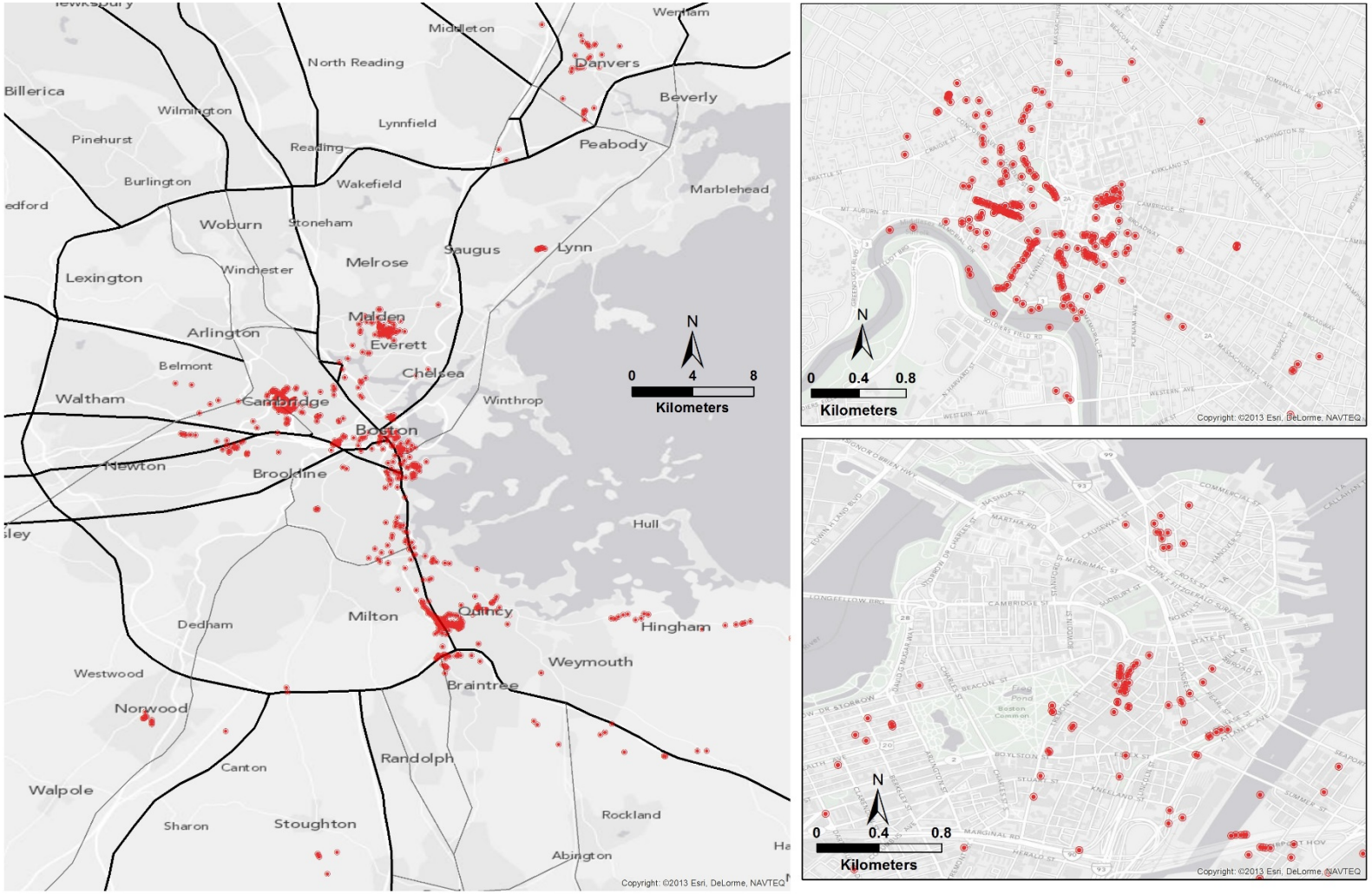
- **Estimation precision error**:
  ~ 300 meter

Reference: http://www.airsage.com/Technology/How-it-works/


Wifi Activity
Sun Jan 31 00:00:00 EST 20
2665
Google

Raw Data Description

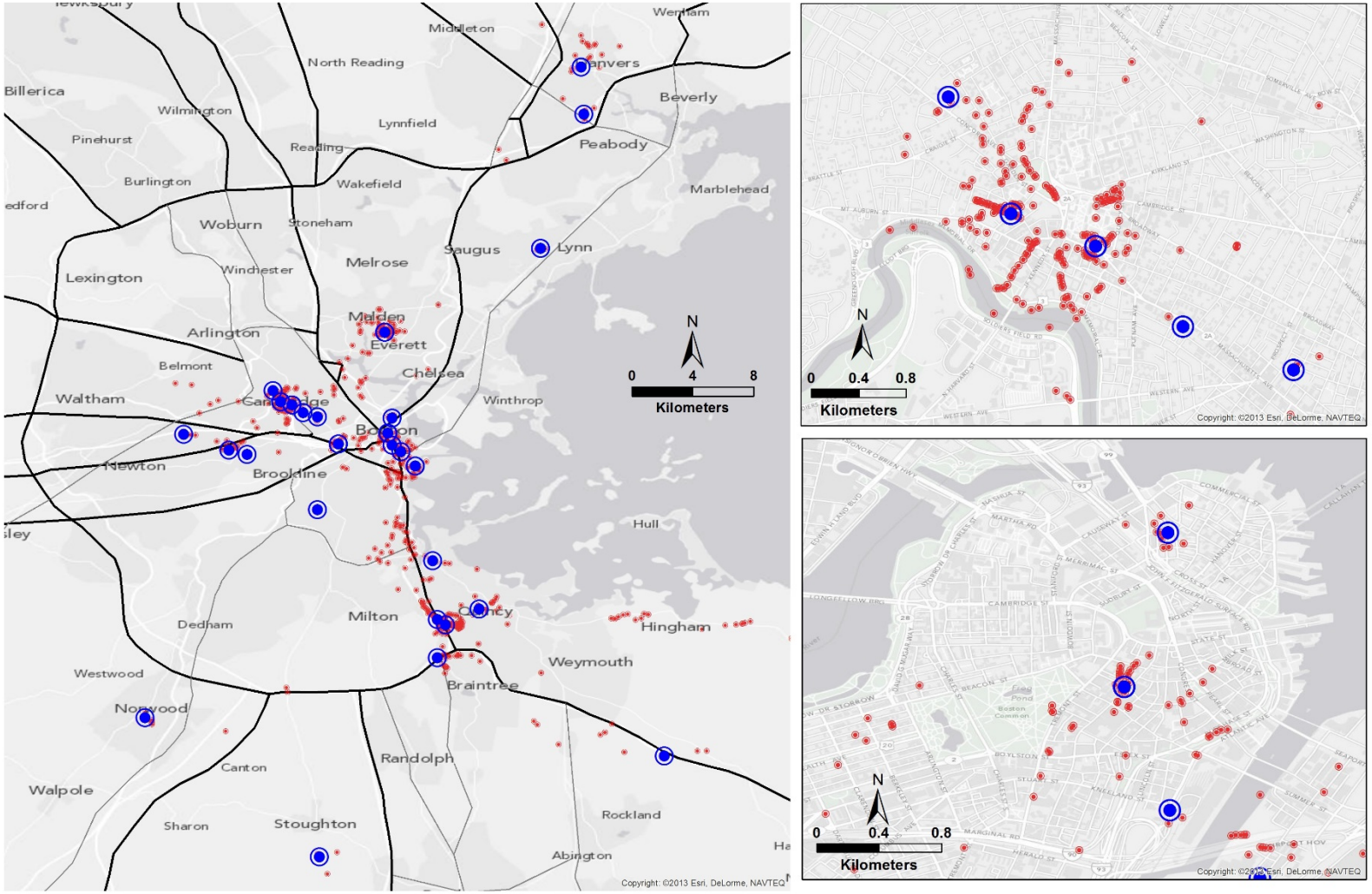
Example of one anonymous cell phone user



1776 phone records for one anonymous user in 2 months, February and March, 2010

Extraction of Daily Trajectories

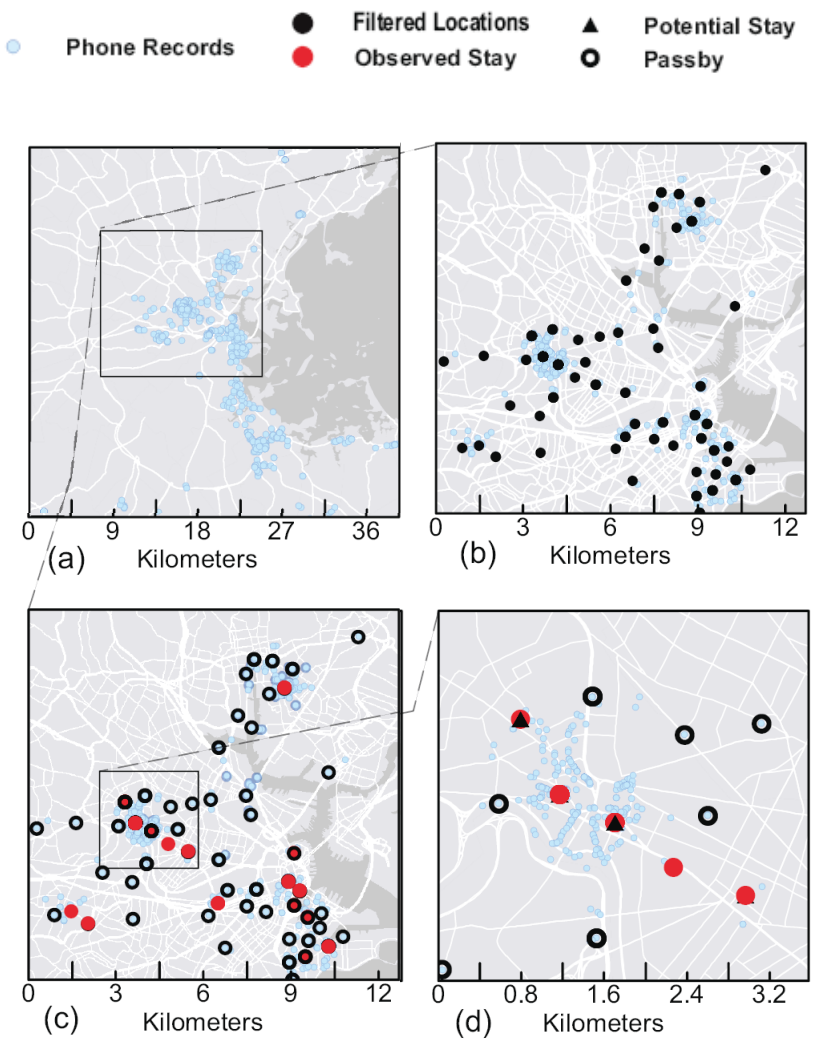
Example of one anonymous cell phone user



Final agglomerated activity destination points for this user: 28 points.

Chapter 3: Extraction of Daily Trajectories

Extracting Point-based Stay, Pass-by and Potential Stays



Algorithms

(Adapted and revised from Hariharan and Toyama, 2010)

(a) Raw cell phone records as input

(b) Reducing noisy jumps

- set roaming distance : 300 meter
- agglomerative clustering to consolidate points that are spatially close to their medoids. (cluster radius=250 meter)

(c) Detecting "Stay", and "Pass-by" areas

- set time duration: 10 minutes

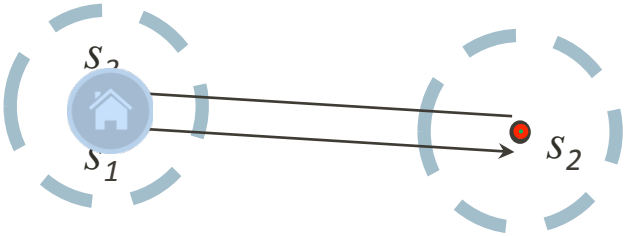
(d) Detecting "Potential Stay" areas

- extract distinct "Stay" area as destinations;
- flag pass-by points collocating with any of the destinations as "potential stay" areas.

Reference:
R. Hariharan and K. Toyama. Project lachesis: parsing and modeling location histories. In Geographic Information Science, pages 106–124. Springer, 2004.

Measuring Individual Activities: Home, Work, and Other

- A phone user's "home" is defined as
  - the most frequently used tower during nights of weekdays & days of weekends
  - over the study period
  - Night time: a parameter (e.g., from 7 pm to 7 am)



A phone user's "work" is defined as

- the most frequently used tower working hours of the weekday

A phone user's "other" is defined as the rest

Origin-Destination Generation

Departure Time Estimation

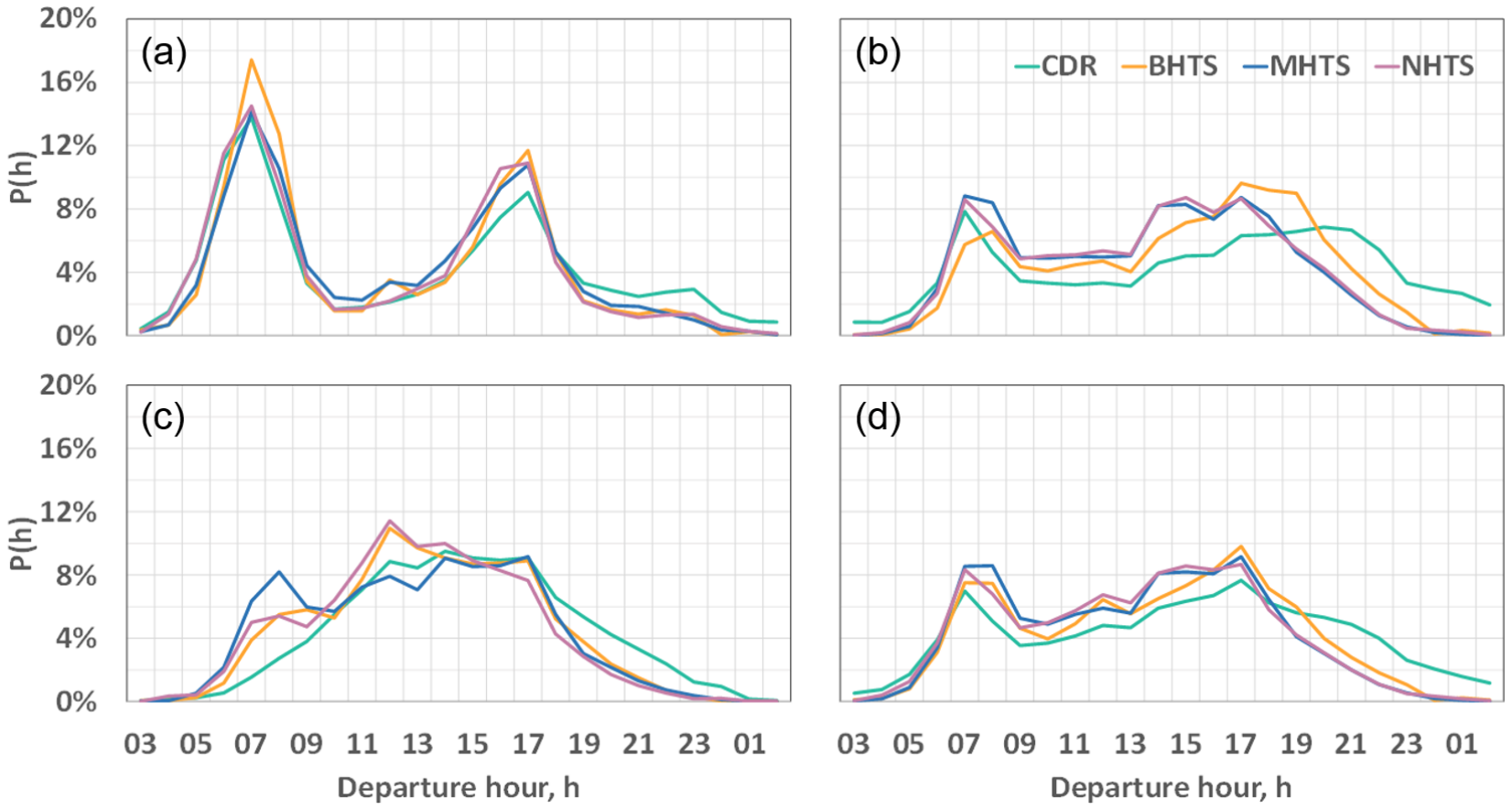
Trip Purposes

Trip Purposes by Time-of-Day

- **HWB**
- **HBO**
- **NHB**

- **AM**
- **MD**
- **PM**
- **RD**

Comparison with Traditional Models



Origin-Destination Generation

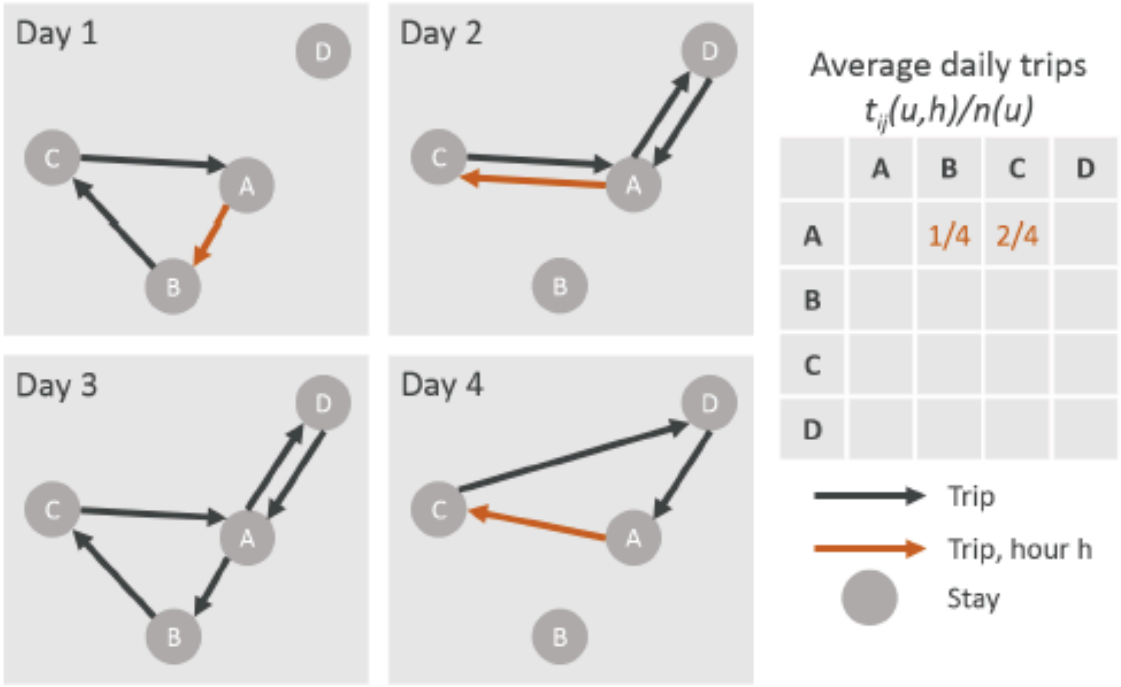
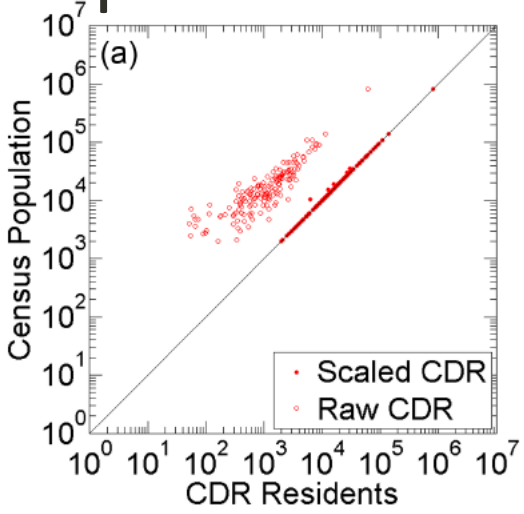
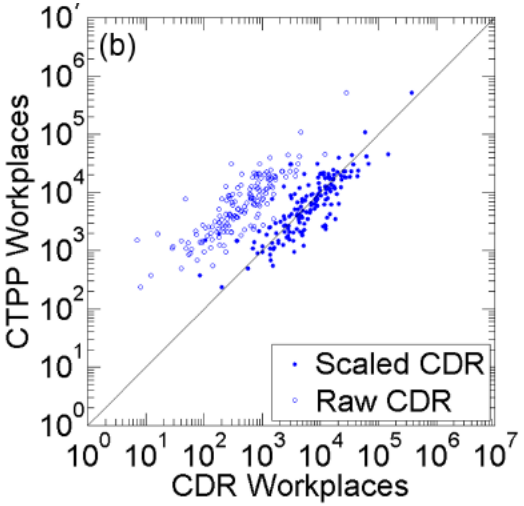


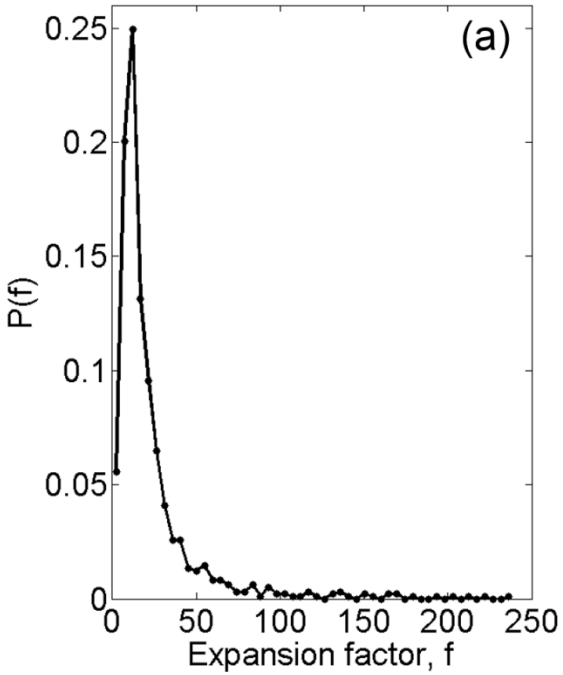
Figure 1: User u makes trips between four unique locations over four days. The user's trips $t_{ij}(u,h)$ in hour h, shown in orange, are converted to average daily trips in hour h by dividing by the number of days $n(u) = 4$ we observe the user.

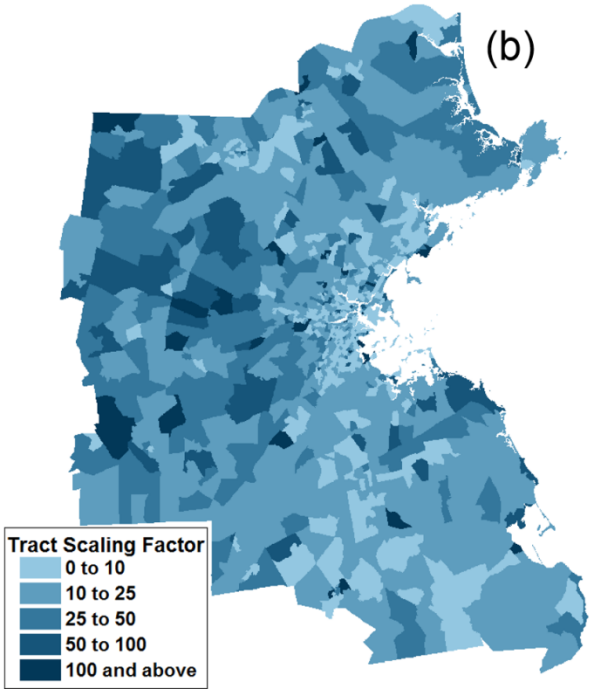
Origin-Destination Generation

Expansion Factors

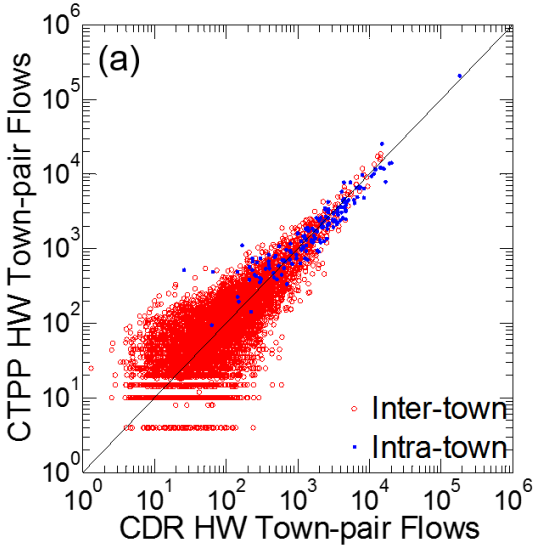


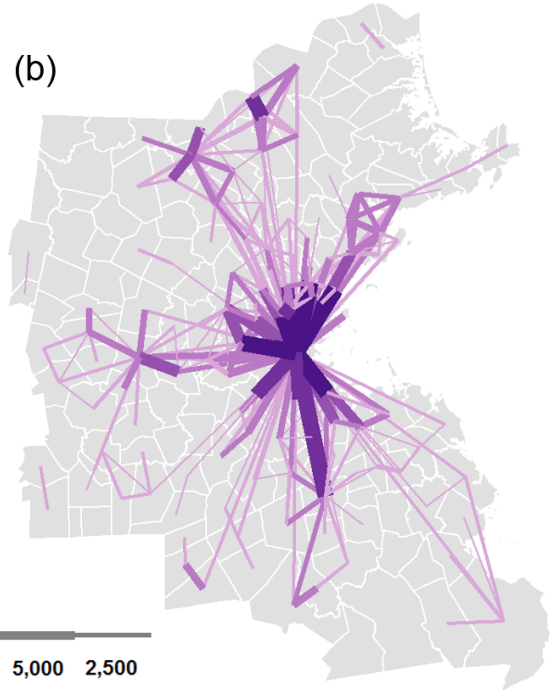


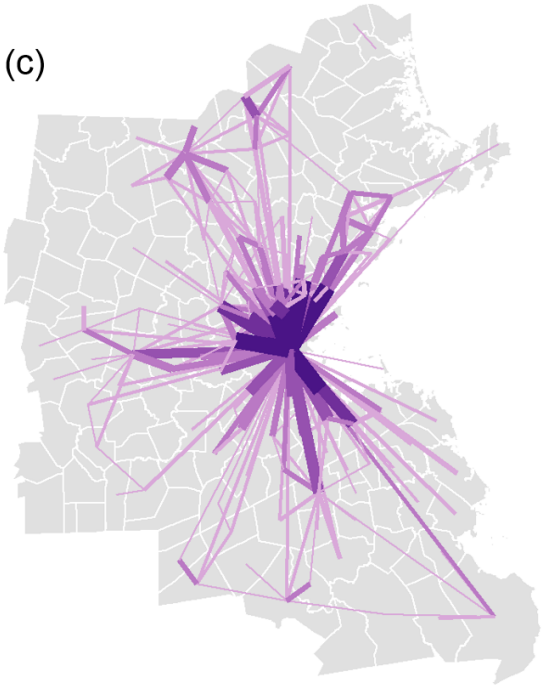




Comparison with Traditional Models







HW Town-pair Flows
1,000 to 2,500
2,500 to 5,000
5,000 to 7,500
7,500 to 10,000
10,000 and above

10,000 5,000 2,500

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Origin–destination trips by purpose and time of day inferred from mobile phone data

Lauren Alexander [a,*], Shan Jiang [b], Mikel Murga [a], Marta C. González [a]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States
[b] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States

Route Assignment

1. Road networks from OpenStreetMap data.
2. Algorithm B, implements equilibration on a directed acyclic graph (DAG).
3. Keep track of where flow is sent two and from.

$$r_g = 1 - \frac{\sum_{o,d} t_{od} d_{od}}{\sum_{e \in E} t_e v_e},$$

where t_{od} and d_{od} represent the demand and the travel time between an origin and a destination, and t_e and v_e represent the travel time and the volume on a road segment e.

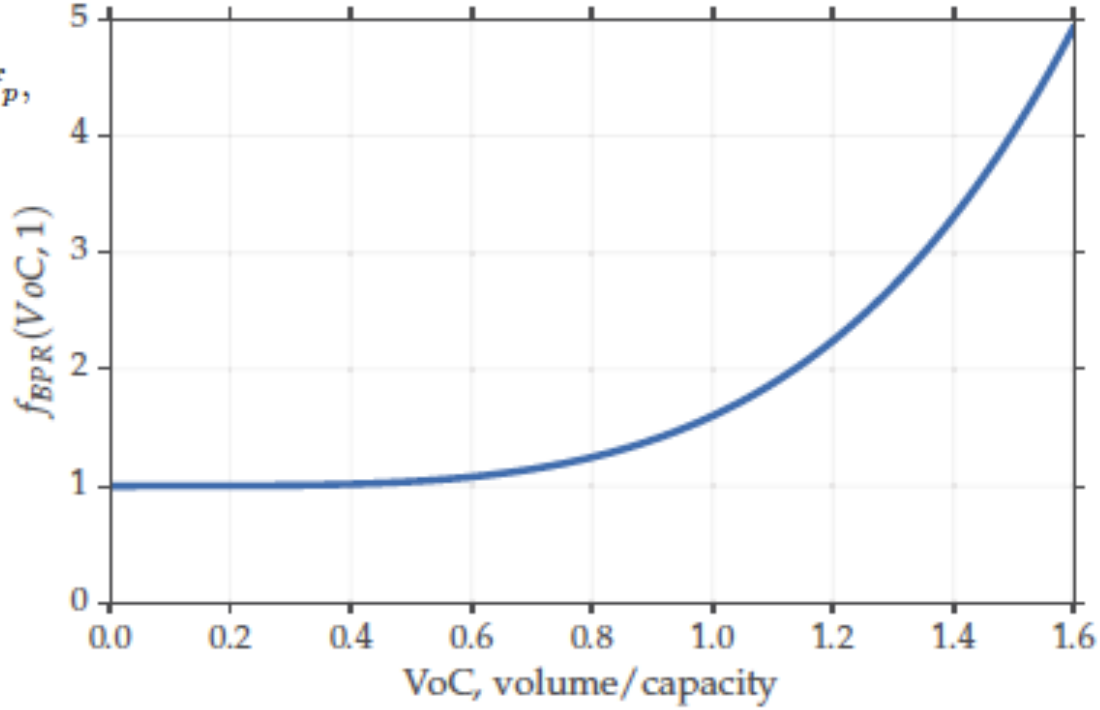
This ensures that all drivers in the system are in fact taking the shortest possible routes,

Assigned volumes are converted to link travel times using a standard BPR function

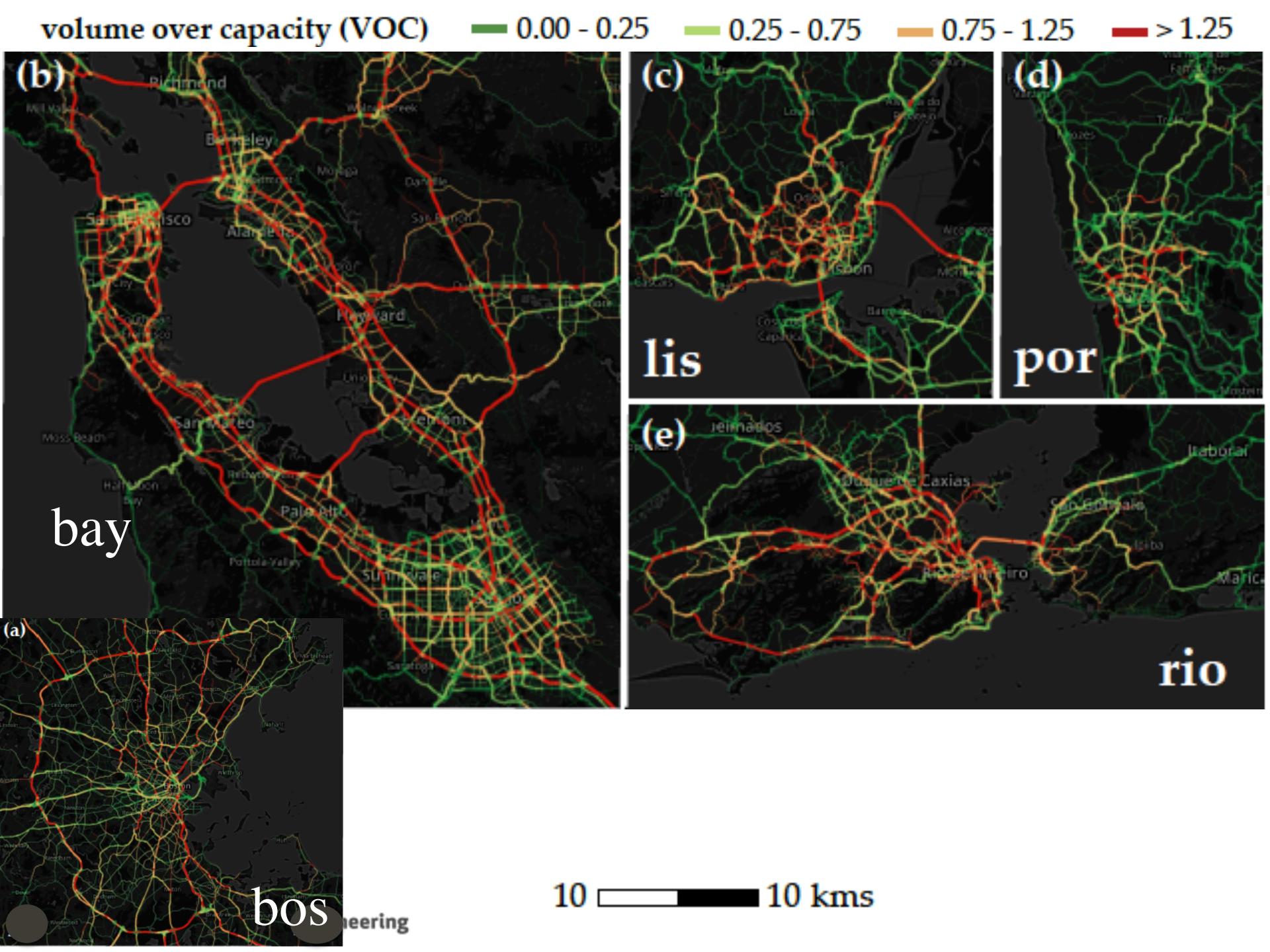
$$f_p^{rio} = f_p^{boston} = f_p^{sfbay} = 1.3 \text{ and } f_p^{lisbon} = f_p^{porto} = 1.1.$$

$$f_{BPR}(VoC, f_p) = t_f * \left(1 + \alpha\,(VoC)^{\beta}\right) * f_p,$$

$$\alpha = 0.6 \text{ and } \beta = 4$$

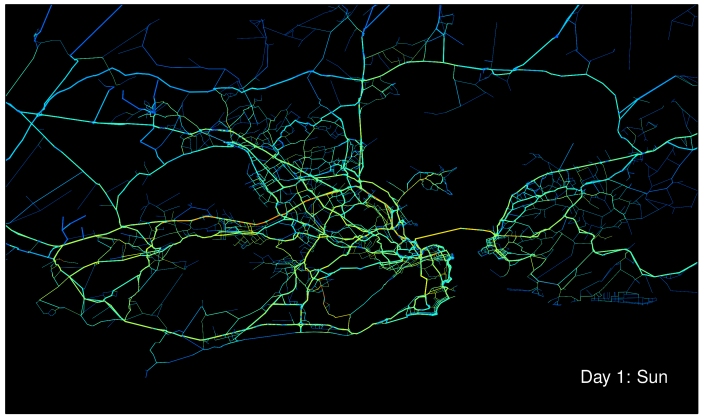


Note: The results of validated travel time at the level of routes act as a validation of the OD flows and show an application of the urban mobility platform to compare cities and the cause of their congestion.

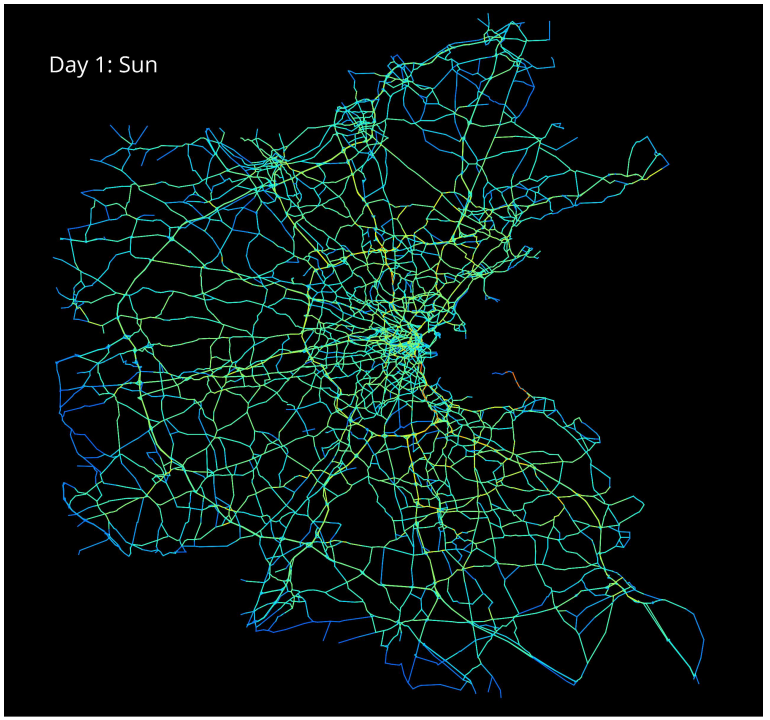

volume over capacity (VOC)
0.00 - 0.25
0.25 - 0.75
0.75 - 1.25
> 1.25
(b)
bay
(c)
lis
(d)
por
(e)
rio
(a)
bos
10
10 kms

Paper 2: Unraveling Urban Traffic

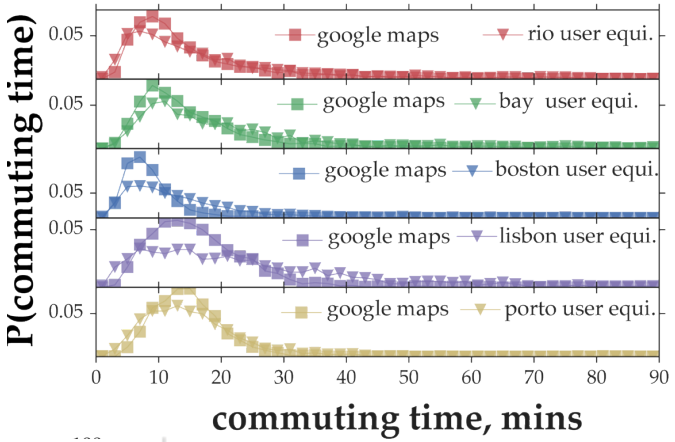
CDR Rio de Janeiro + Waze Data

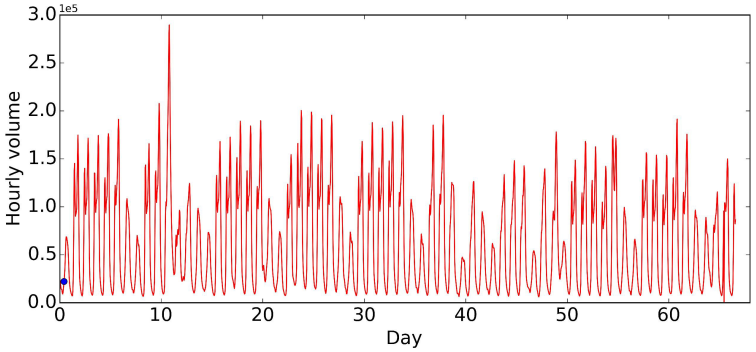


CDR Boston + TomTom Speed reads

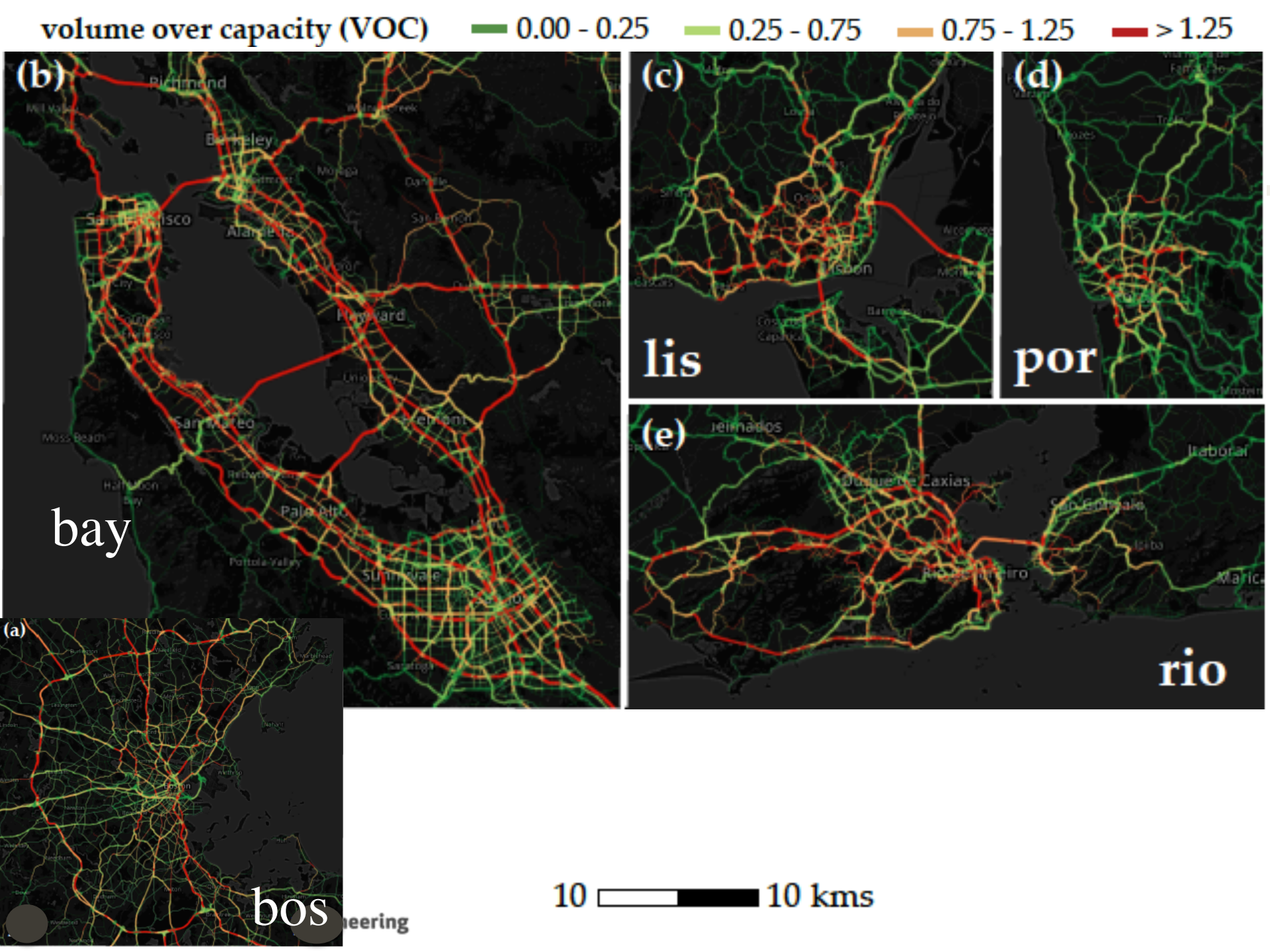


Travel Times Validation

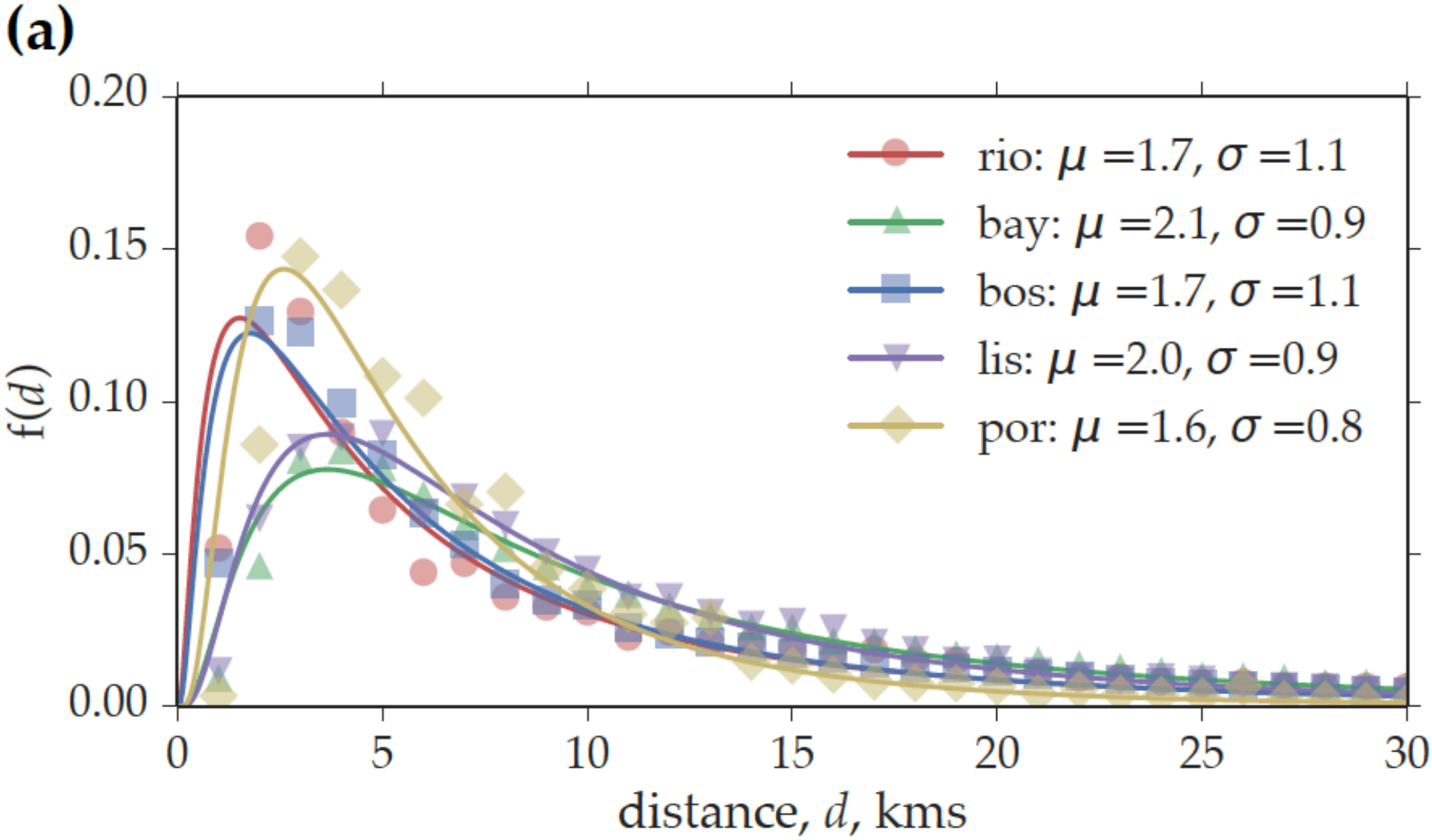




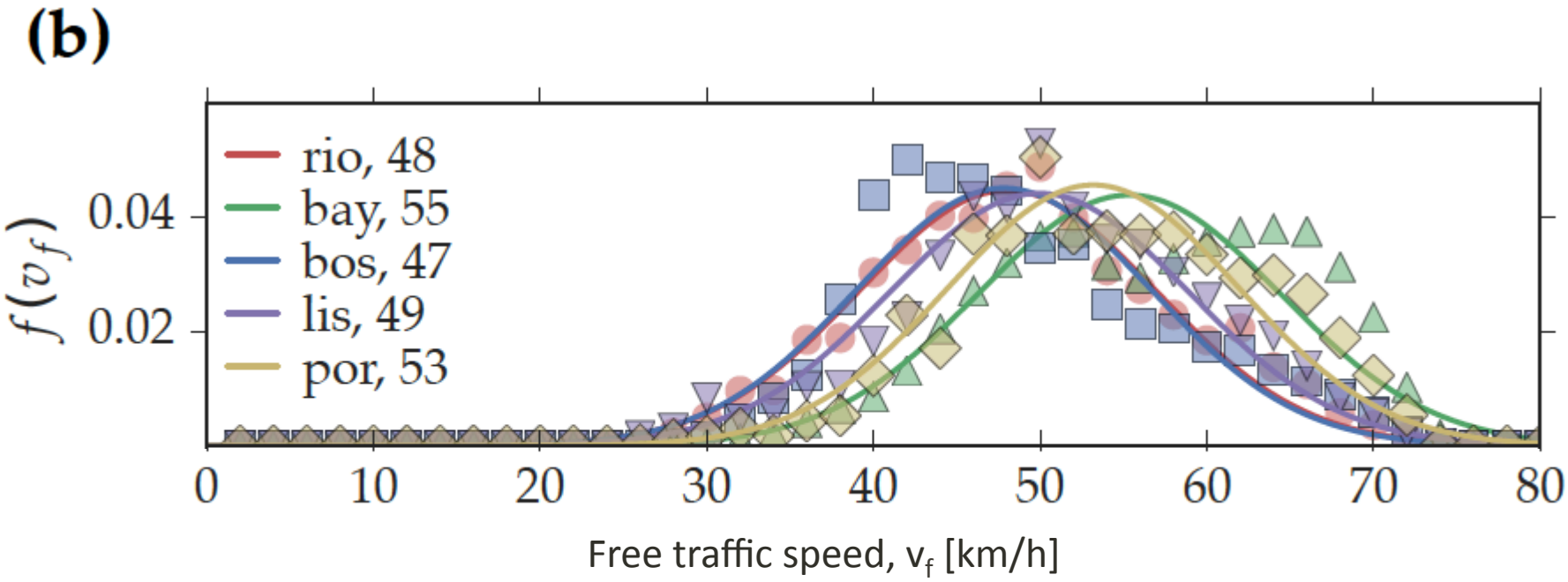
School of Engineering


volume over capacity (VOC)
0.00 - 0.25
0.25 - 0.75
0.75 - 1.25
> 1.25
(b)
bay
(c)
lis
(d)
por
(e)
rio
(a)
bos
10
10 kms

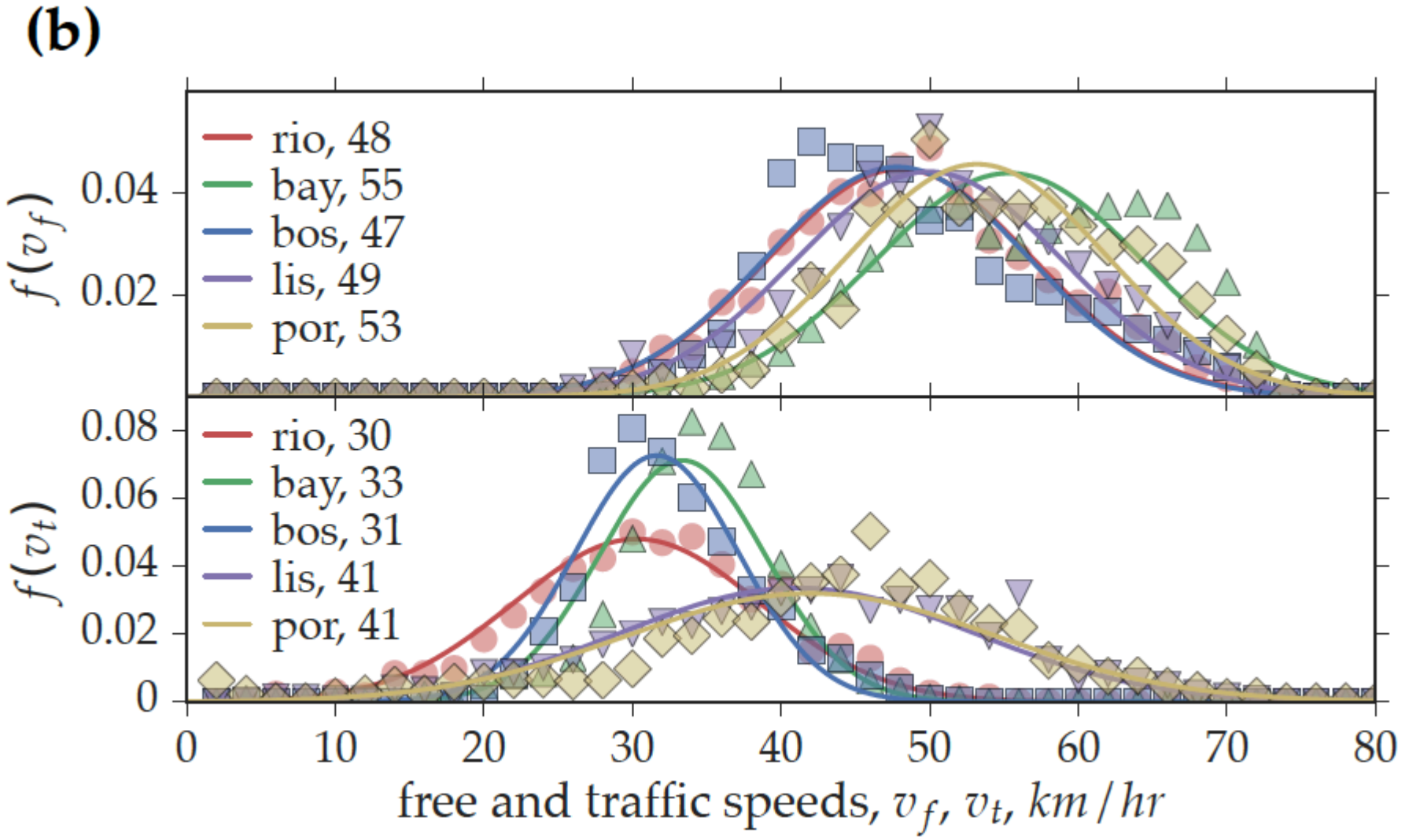
Commuting Distance



Free Traffic Speed Comparison

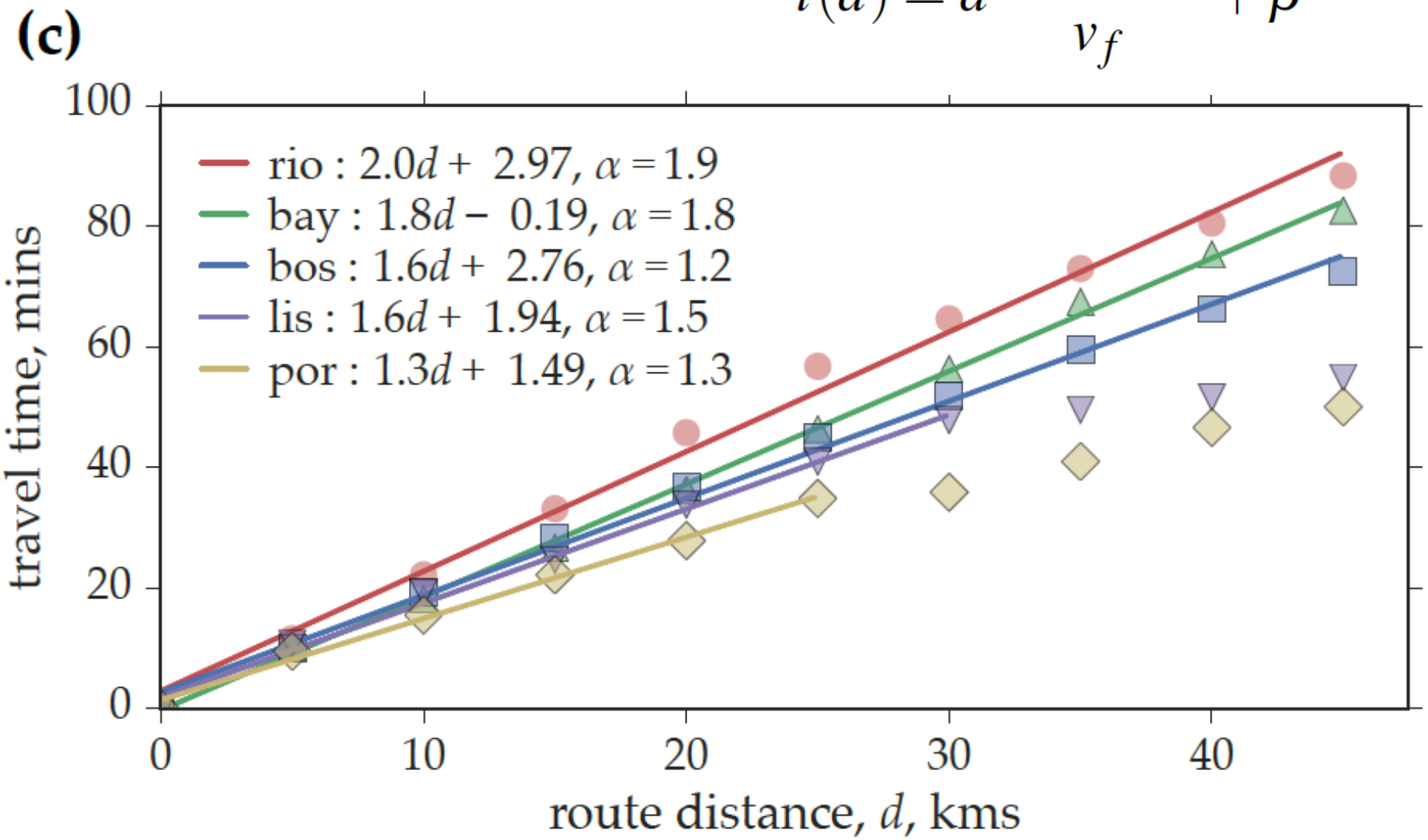


Traffic Speed Comparison

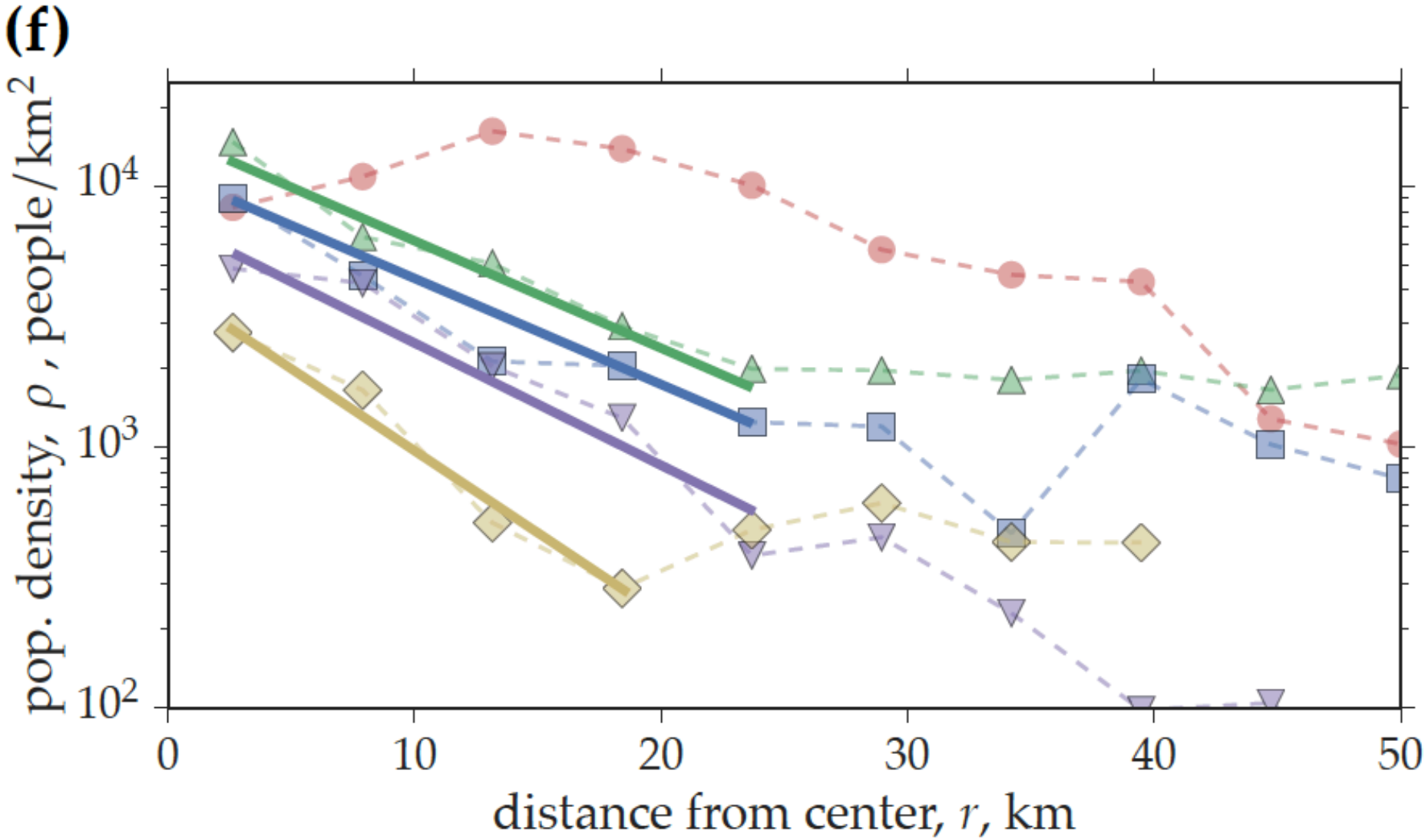


Commuting Time

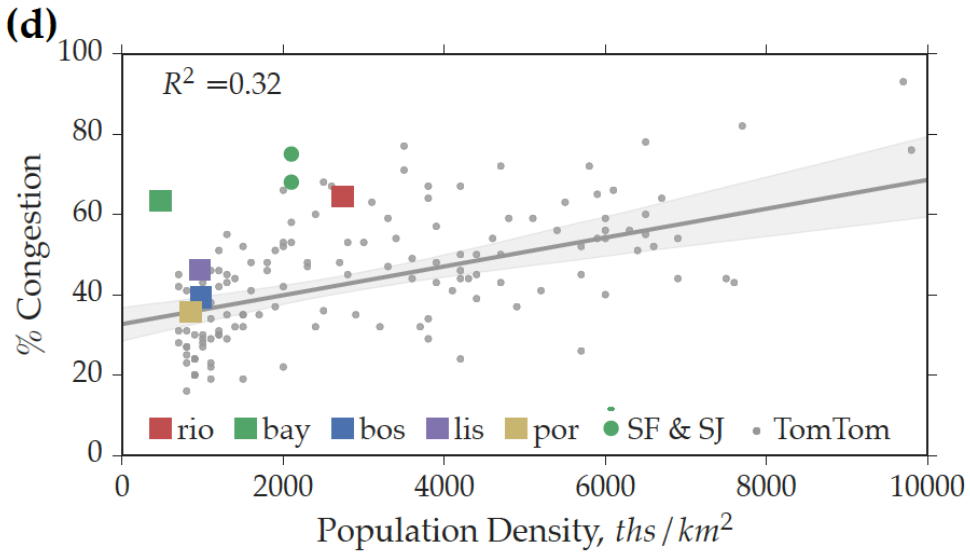
$$t(d) = d\frac{(1+\Gamma)^{\alpha}}{v_f} + \beta$$

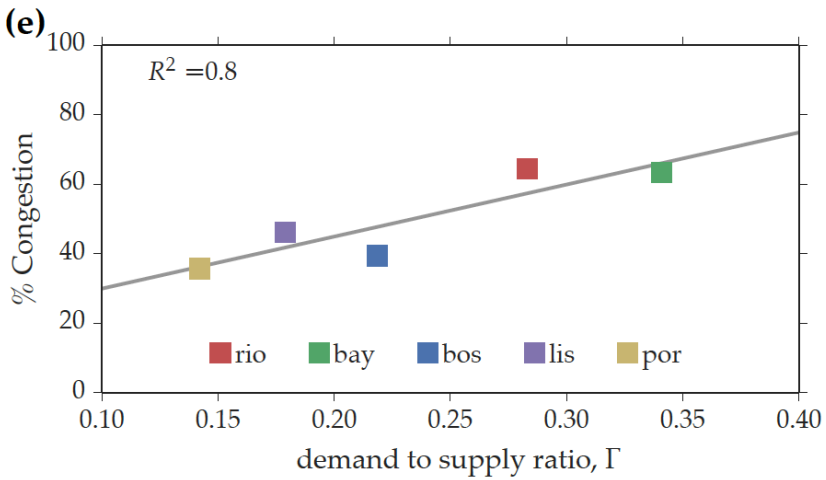


Spatial Density



% Time Lost vs. Pop. Density





$$\Gamma = \frac{\sum_{e \in E} l_e x_e}{\sum_{x_e > 0, e \in E} l_e C_e}.$$

x_e # Cars in the road link e

l_e Road link length in miles

C_e Capacity in the road link [cars/miles^2]

Smart-app (routing)

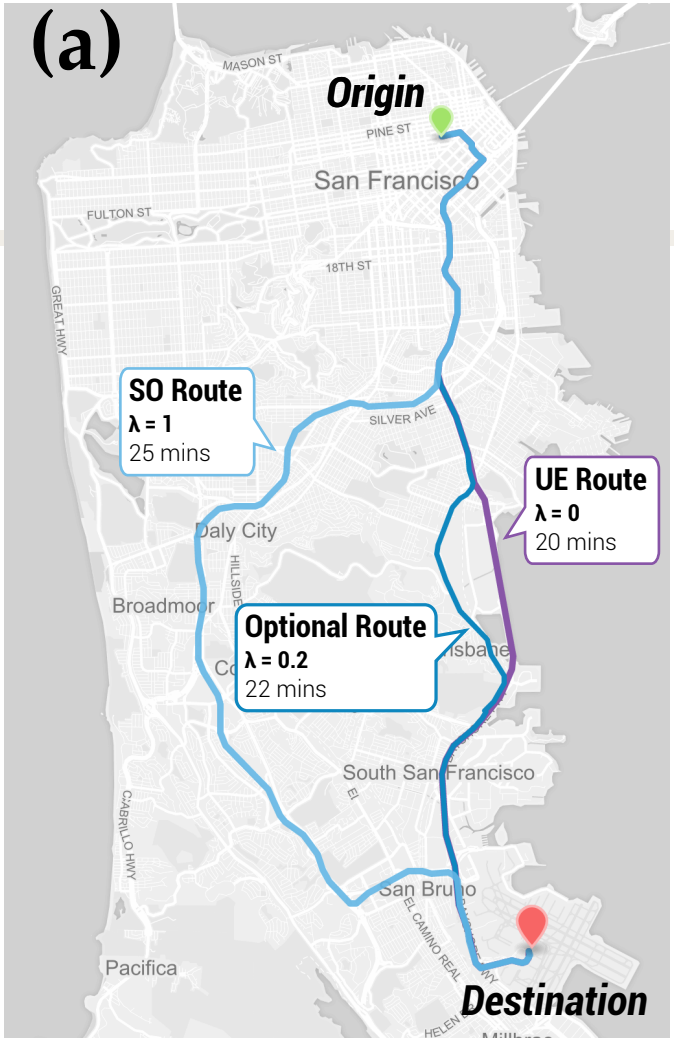
Modifications on the level of altruism:

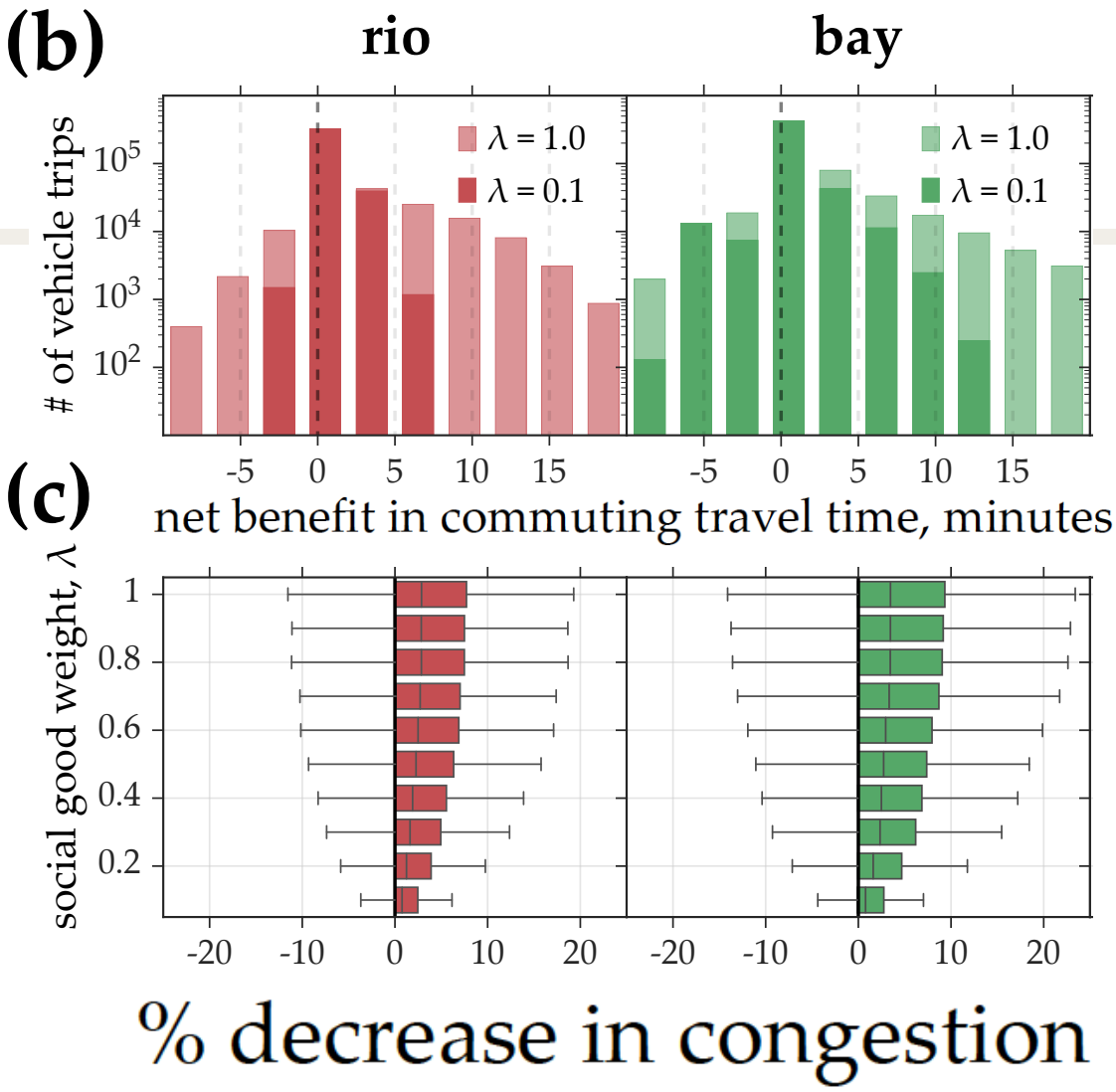
$$c_e^{\lambda}(x_e) = (1-\lambda)t_e(x_e) + \lambda \frac{d\left[x_e t_e(x_e)\right]}{dx_e}$$

$\lambda=[0..1]$

User Equilibrium component

Social Optimun component


(a)
Origin
Destination
SO Route
λ = 1
25 mins
UE Route
λ = 0
20 mins
Optional Route
λ = 0.2
22 mins
San Francisco
Daly City
Broadmoor
South San Francisco
San Bruno
Pacifica
MASON ST
PINE ST
FULTON ST
18TH ST
SILVER AVE
GREAT HWY
HILLSIDE
CABRILLO HWY
EL CAMINO REAL


(b)
rio
bay
# of vehicle trips
10^5
10^4
10^3
10^2
λ = 1.0
λ = 0.1
-5
0
5
10
15
net benefit in commuting travel time, minutes
(c)
social good weight, λ
1
0.8
0.6
0.4
0.2
-20
-10
0
10
20
% decrease in congestion

Implementation Approach

Stage 1: Strategy

- Publish research approach
  - Incentives
  - Success
- Form partnerships
- Secure funding

Stage 2: Pilot

- Solicit proposals
- Select host city
- Conduct Pilot

Stage 3: Expand

- Advertise findings
- Implement Smart Commute in cities nationwide

Volpe

Further Details

1)

Origin–destination trips by purpose and time of day inferred from mobile phone data

CrossMark

Lauren Alexander [a,*], Shan Jiang [b], Mikel Murga [a], Marta C. González [a]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States
[b] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States

OD generation and Validation (Boston)

2)

The path most traveled: Travel demand estimation using big data resources

Jameson L. Toole [a,1], Serdar Colak [b,*,1], Bradley Sturt [a], Lauren P. Alexander [b], Alexandre Evsukoff [c], Marta C. González [a,b]

OD generation and Validation (Portable Platform)

3)

Understanding the limits of socially aware routing in urban areas

Serdar Colak, Antonio Lima and Marta C. Gonzalez

Under Review, Nature Communications @humnetlab.mit.edu

Understanding Individual Routing Behaviour

Antonio Lima, Marta González
Civil and Environmental Eng. - Massachusetts Institute of Technology

Motivations

It is the natural next step in understanding human mobility.

Route choice is a fundamental step of traffic modeling, the task of transforming a set of travel demand (OD matrix) into flows and travel times.

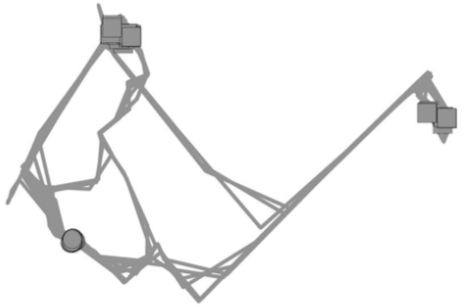
The assumption that “people choose the minimum cost path”, although widely accepted in academic and commercial enviroments, has little empirical support.

How do people navigate in the city?

We analyse 1,5 M GPS trajectories, driven by a set of individuals within four major Spanish cities during a period of 18 months.

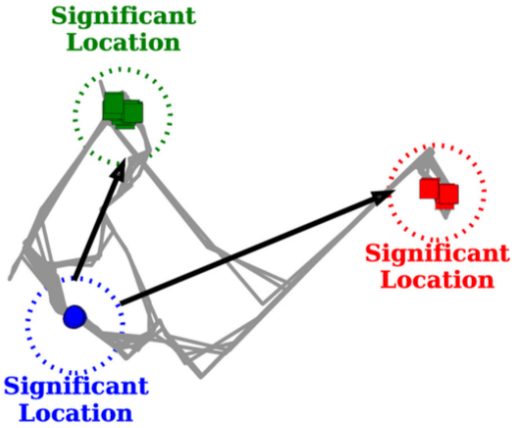
- How many routes a driver uses typically.
- If the routes performed by users are “optimal”.
- Whether some routes are predominant over others.

From trajectories to route choices



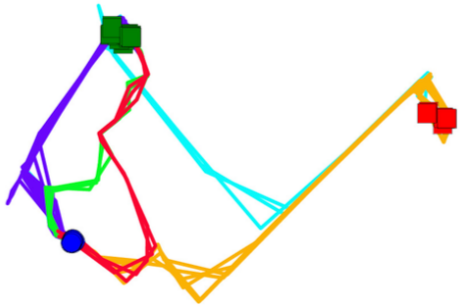
Each trajectory is composed by an arbitrary number of points, every N seconds.

We cluster each driver's source / destination points into a set of **significant locations**, here shown as dotted circles.



We group trajectories by source-destination pair into **routine trips**, here shown as black arrows.

Finally we further cluster the trajectories in each routine trip into a set of **route choices**, color coded in figure.



From *messy* trajectories to route choices

Clustering algorithms typically require the number of clusters to be specified. We instead use non-parametric algorithms, like **MeanShift** and **DBSCAN**.

Trajectories have an heterogeneous number of points (even on the same routes, because of traffic jams, delays, …). It is not trivial to compare them. We used **Dynamic Time Warping** to establish a matching between the two sets of trajectory points.

Given two paths $A = [a_1, a_2, ..., a_N]$ and $B = [b_1, b_2, ..., b_M]$

following recursive definition, for $i = 1 \dots N-1, j = 1 \dots M-1$, is minimized:

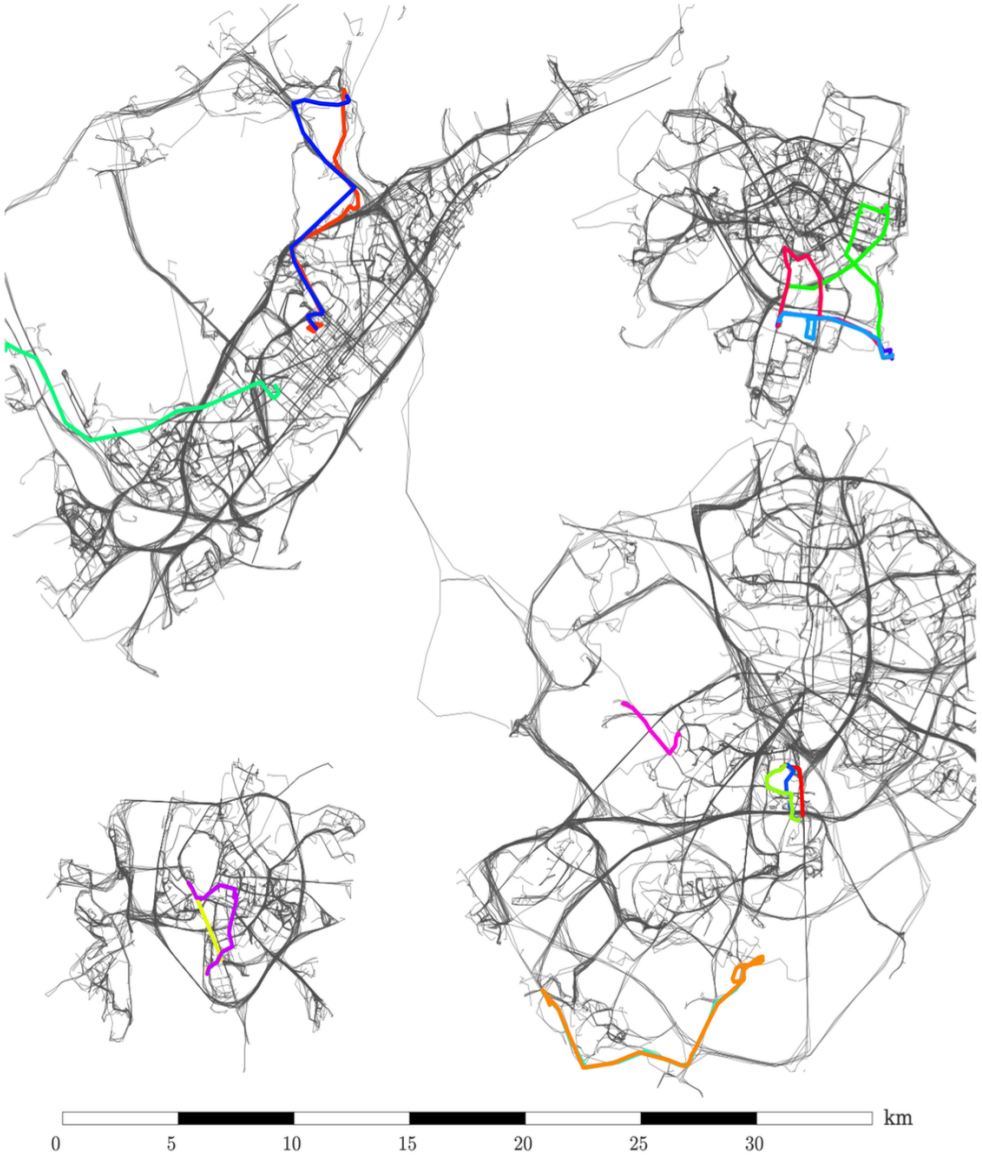
$$W(A_i, B_j) = d(a_i, b_j) + \min \begin{cases} W(A_{i+1}, B_j) \\ W(A_i, B_{j+1}) \end{cases}$$

where A_i and B_j are subsequences containing all the elements $1 \dots i$ from A and $1 \dots j$ from B, respectively;

From *messy* trajectories to route choices

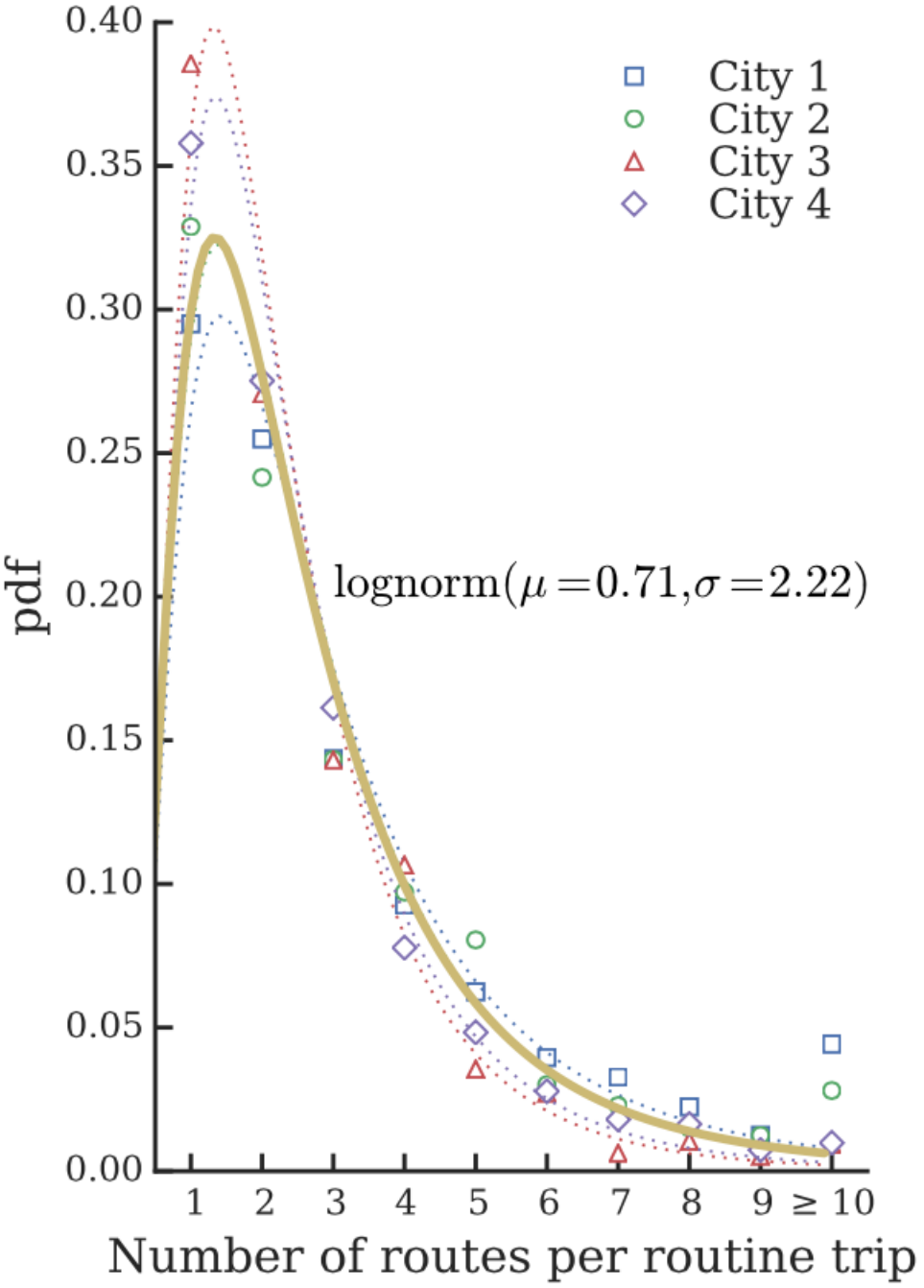
This methodology is agnostic of the underlying urban network.

It can be used to transform unstructured location sequences into route choices between significant locations in any city.



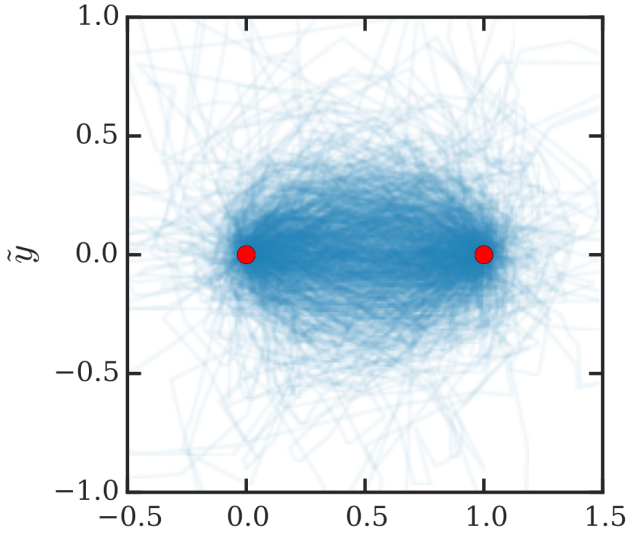
Results

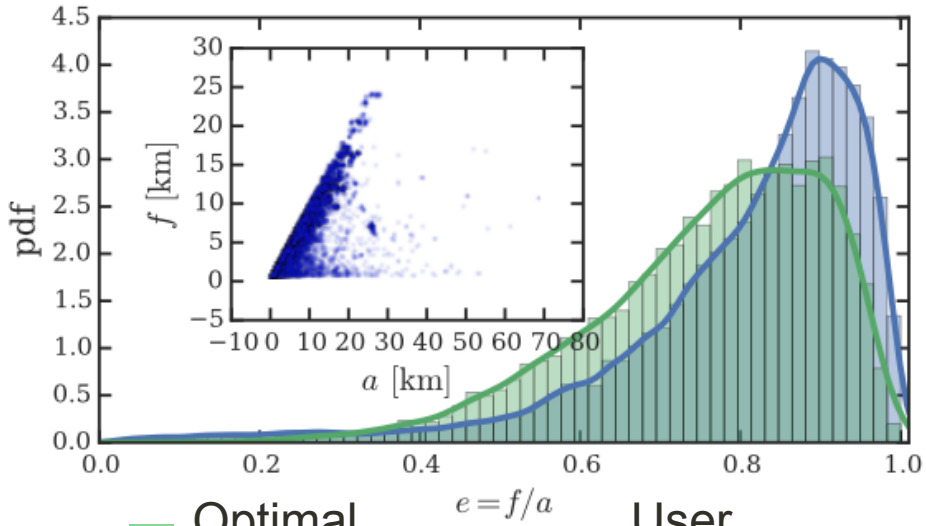
1. Most people use few routes, despite the total period under consideration is 18 months.
2. We compared user trips to trips returned by Google Directions API, which accounts for distance and traffic conditions.
3. We found that **53% of the preferred trips ever used are not optimal**.
4. And the more often people travel between two locations, the more likely is for them to have a **preferred route**.

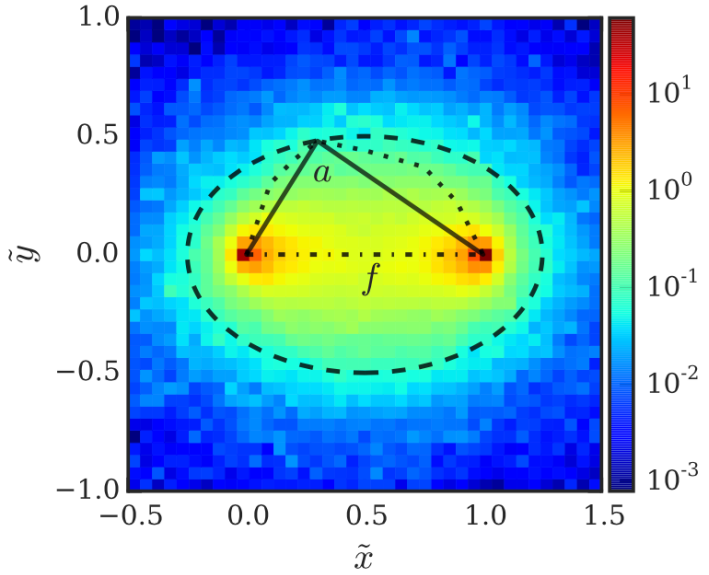


The boundaries of human routes

1. We rototranslated and scaled every trajectory to the same reference system, having source (0, 0) and destination (1, 0). 95% of the positions are contained within an ellipse of high-eccentricity.
2. Eccentricity measures us how much the user is away from the ideal straight location.
3. We show that detours people are willing to take are bound.







flickr: rogermeyer

Take away messages - Recap

- Drivers often do not choose the shortest path.
- Regardless of the urban network, they drive within an high-eccentricity ellipse, with foci as source / destination.
- For recurring trips, a dominant route is preferred, and some alternative routes are occasionally taken.
- This set of behavioural rules can be used to inform realistic models of routing behaviour that are not based on minimum-cost assumptions.

Understanding the limits of socially aware routing in urban areas,

Antonio Lima, Rade Stanojevic, Dina Papagiannaki, Pablo Rodriguez

Marta C. Gonzalez, Under Review, @humnetlab.mit.edu

TimeGeo: Modeling framework for daily time geography from Sparse Data sources

Data Sources

— 2 millions of individual phone users in Boston (For purchase nationwide in AirSage.com)

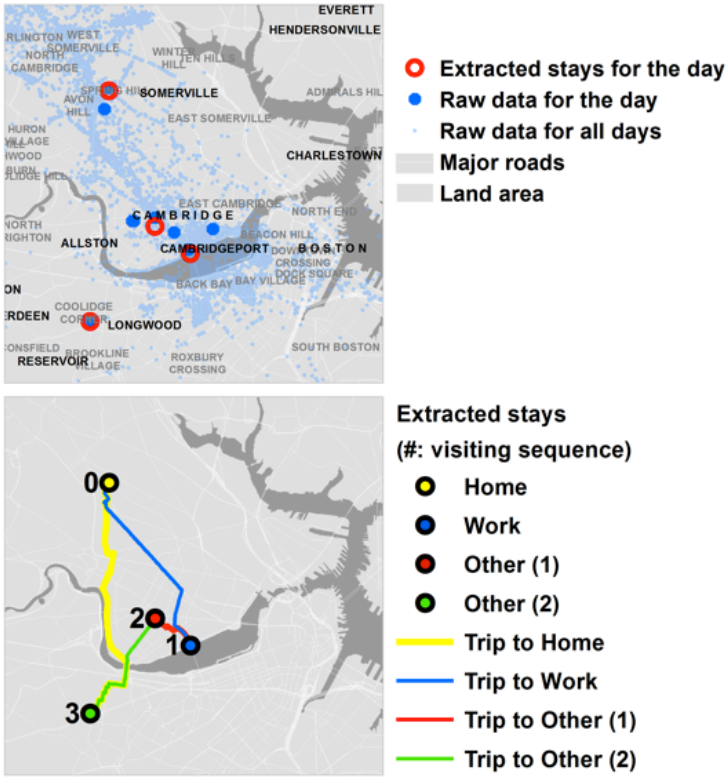
— 14 Months of self-collected **complete** mobile phone data of 1 Student.

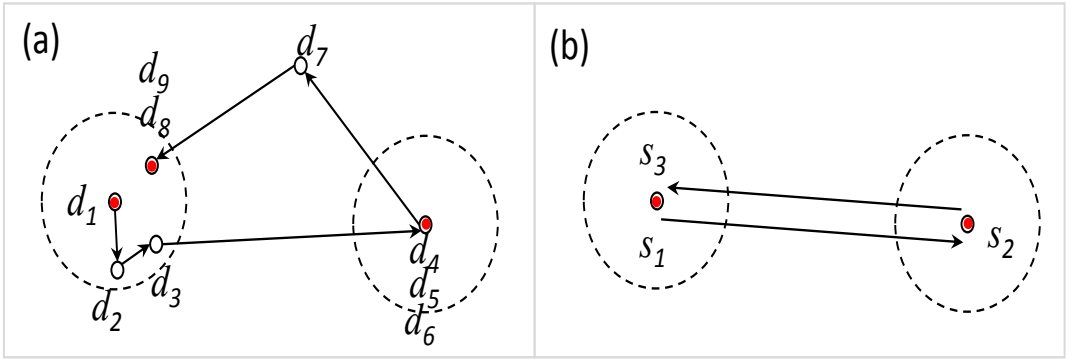
Goal

To model Individual Trajectories (resolution: 10min and 300m radius)

Stay & Pass by Extraction
Home & Work Detection

1 Sample day of a student



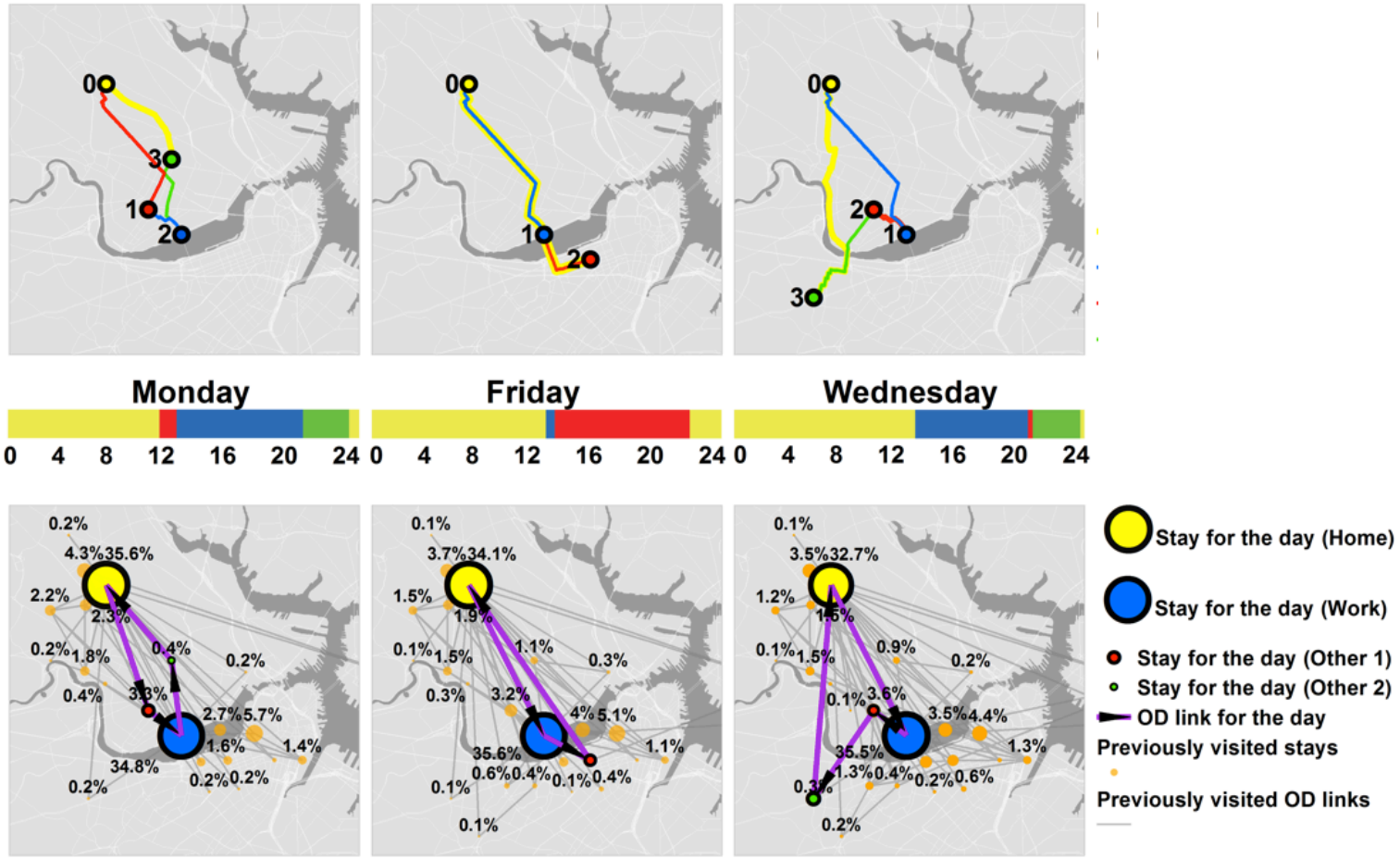


1) With a roaming distance Δd_1 (300 mts.), we cluster spatially close locations within Δd_1 .

2) A time threshold Δt (10 minutes) separate various stay points *Si* .

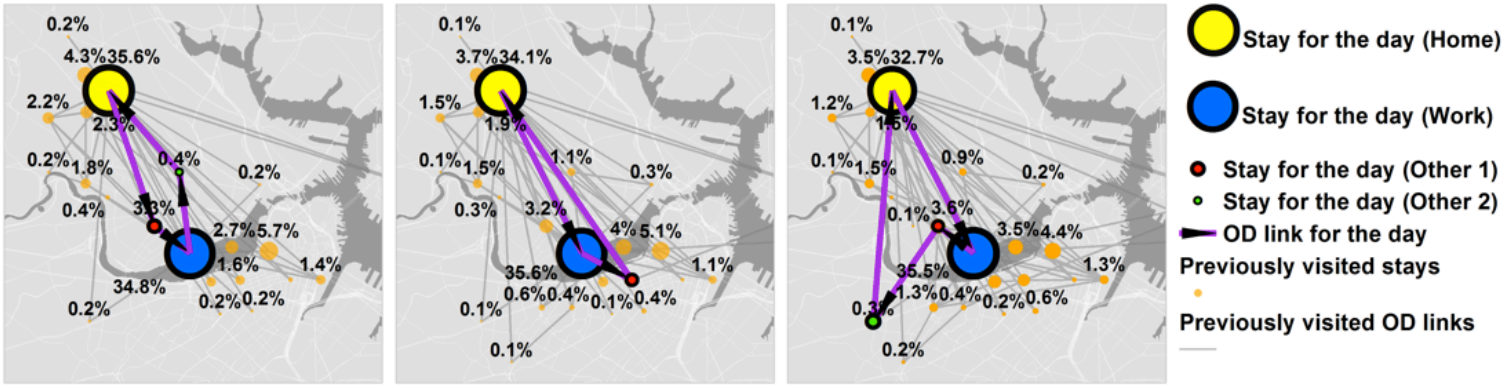
3) Home is defined as the most frequently visited location during nights of weekdays & days of weekends over the study period; A phone user's "work" is defined as the most frequently visited loaction working hours of the weekday

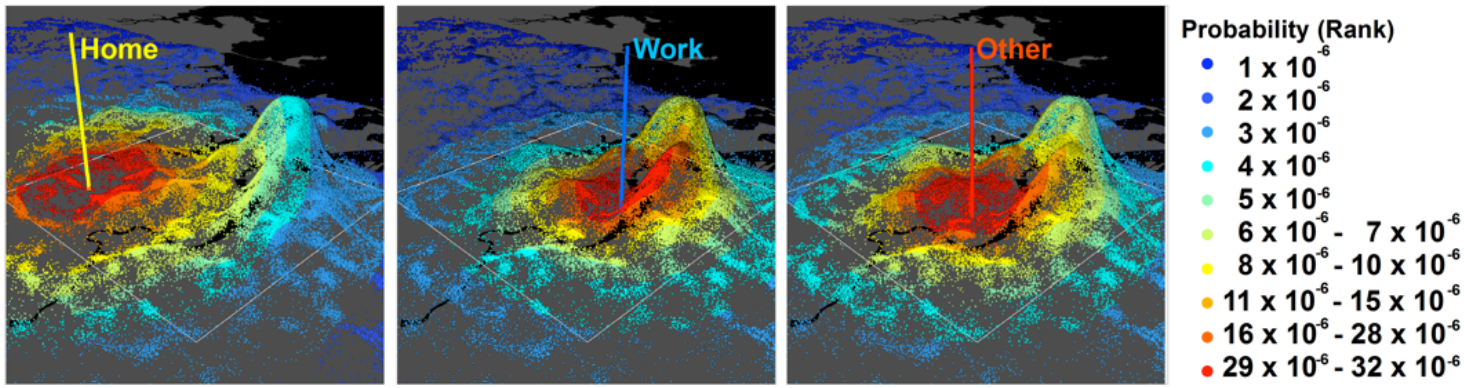
Spatial Mobility Networks based on frequency of returns to few preferred locations



Explorations are selected based on a Ranking Function

Three days of student



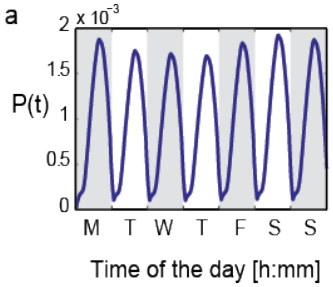


Colors represent the P(rank), height is POIs (point of interests) numbers.

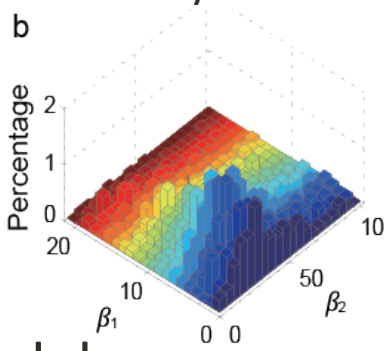
The Model

Features Extracted from data of Active Users

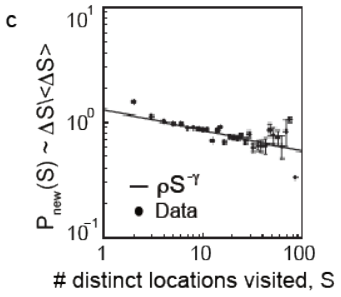
Circadian Rhythm



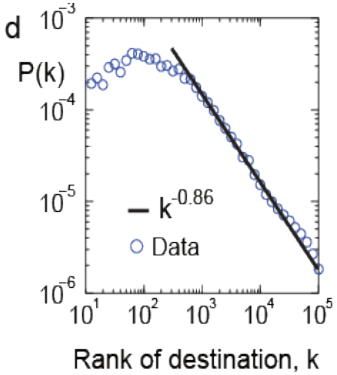
Mobility Rates



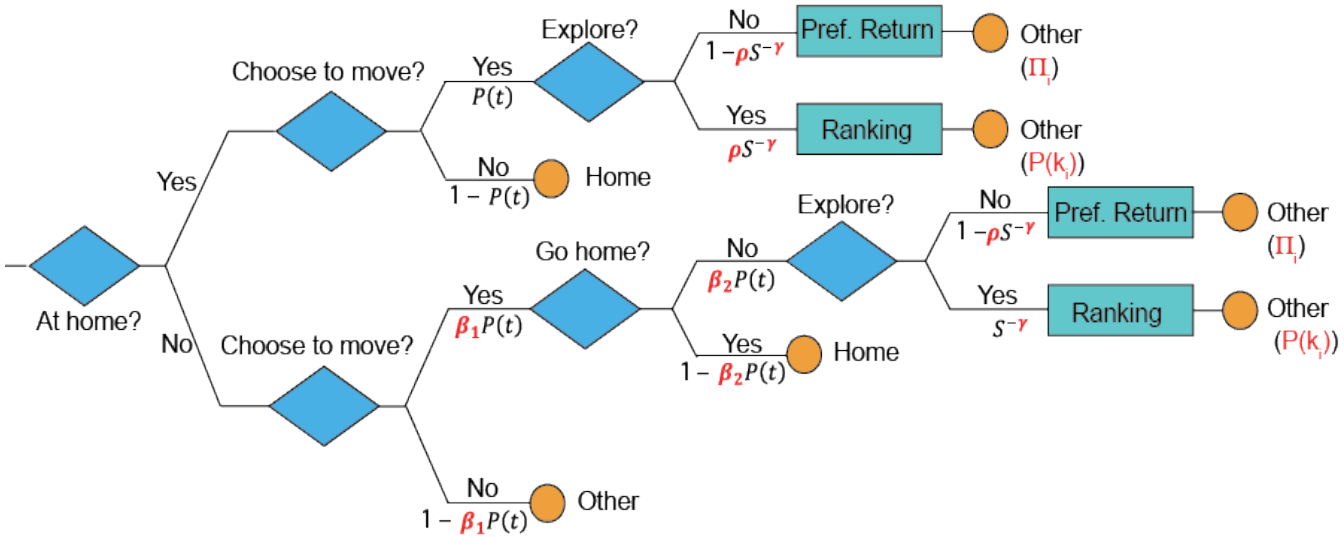
Preferential Return



Ranking of Explorations



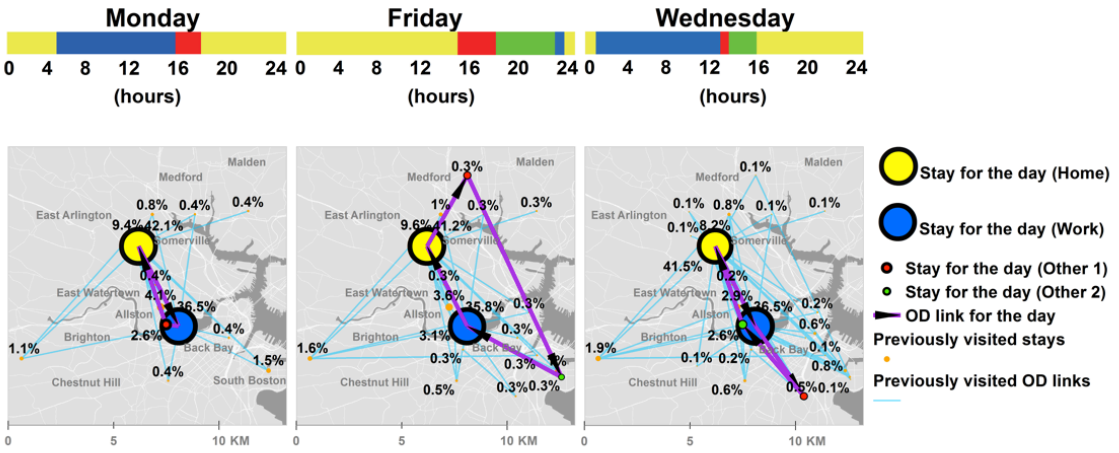
Flowchart of the Model



Models Results

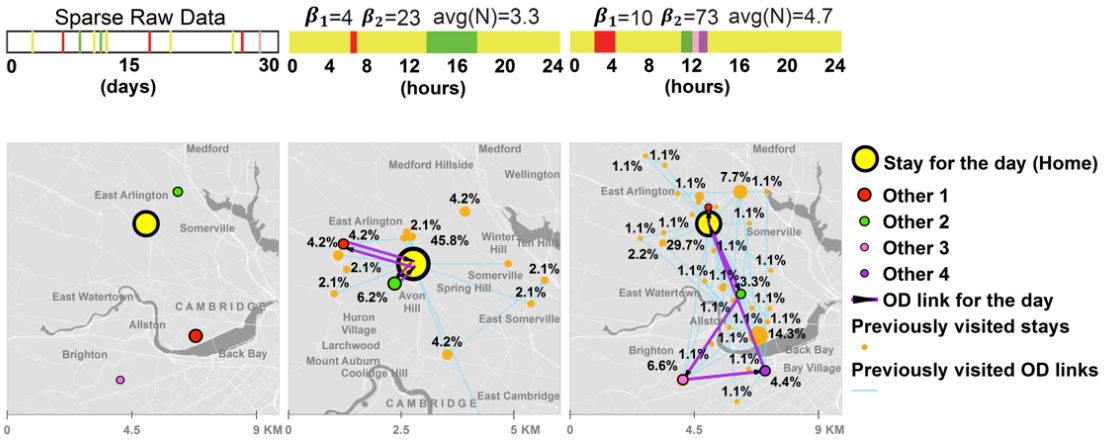
Modeled Trajectories of student (with only home and work data used)

Note: This mechanistic model does not use historic data for training; it can be enriched with existing methods to "predict next locations" (work in progress for KDD'16)



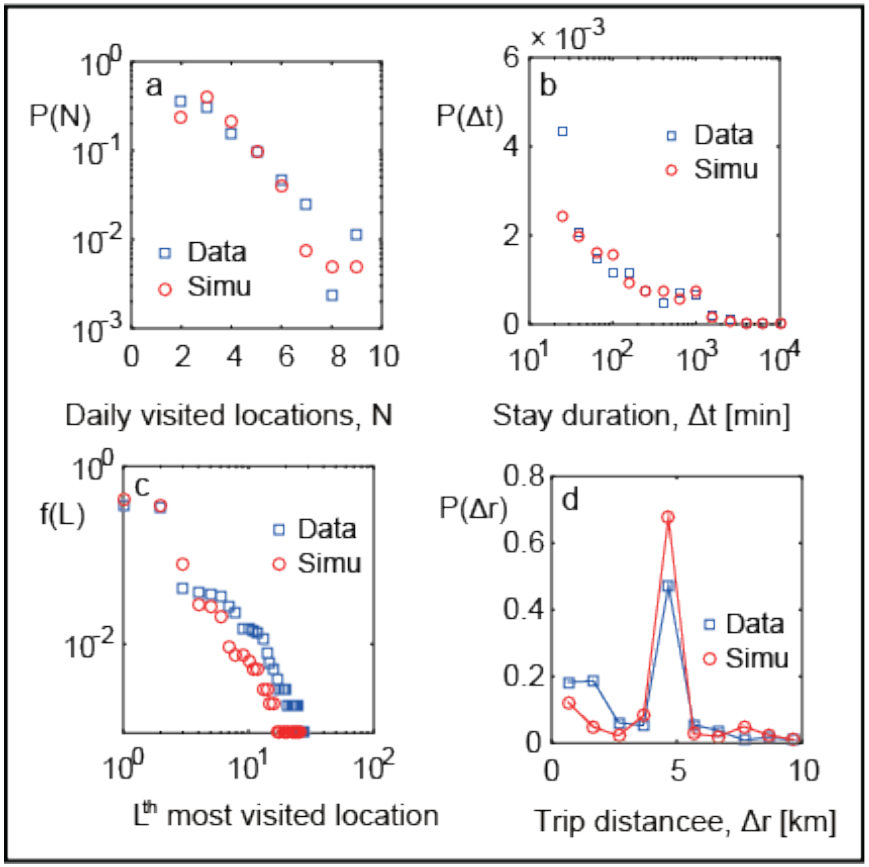
Modeled Trajectories from **sparse data of a sample user** (with previous locations used).

Note: On sparse data "next location" prediction with machine learning methods fail (AAAI-16, submitted)

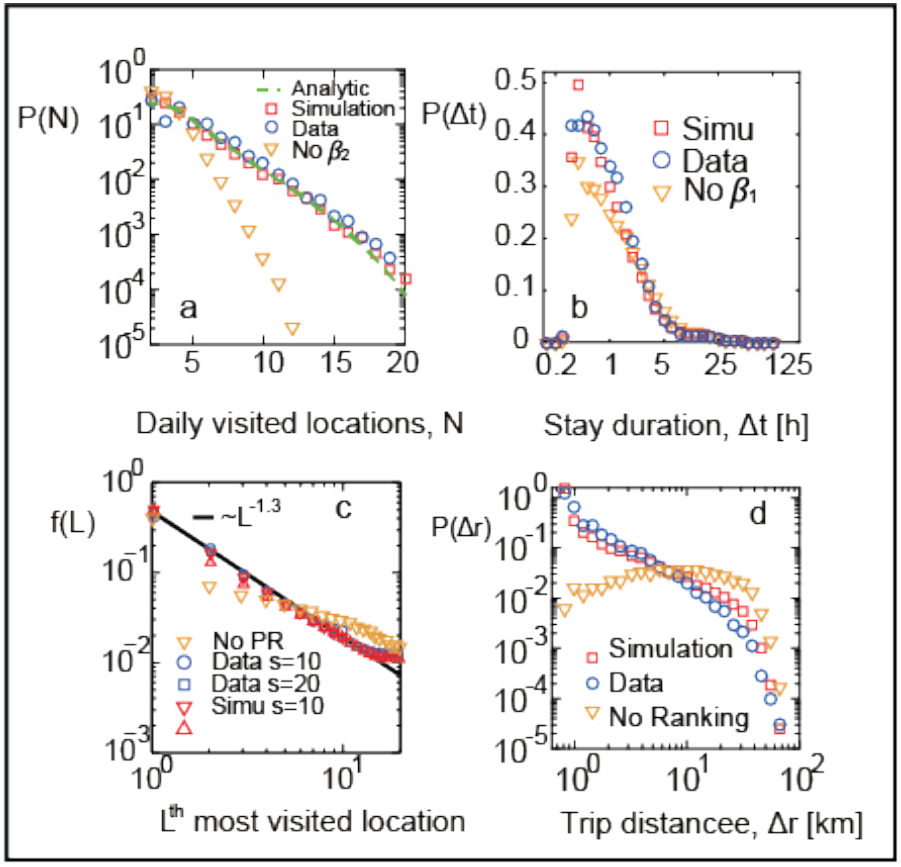


Models Results

Individual user

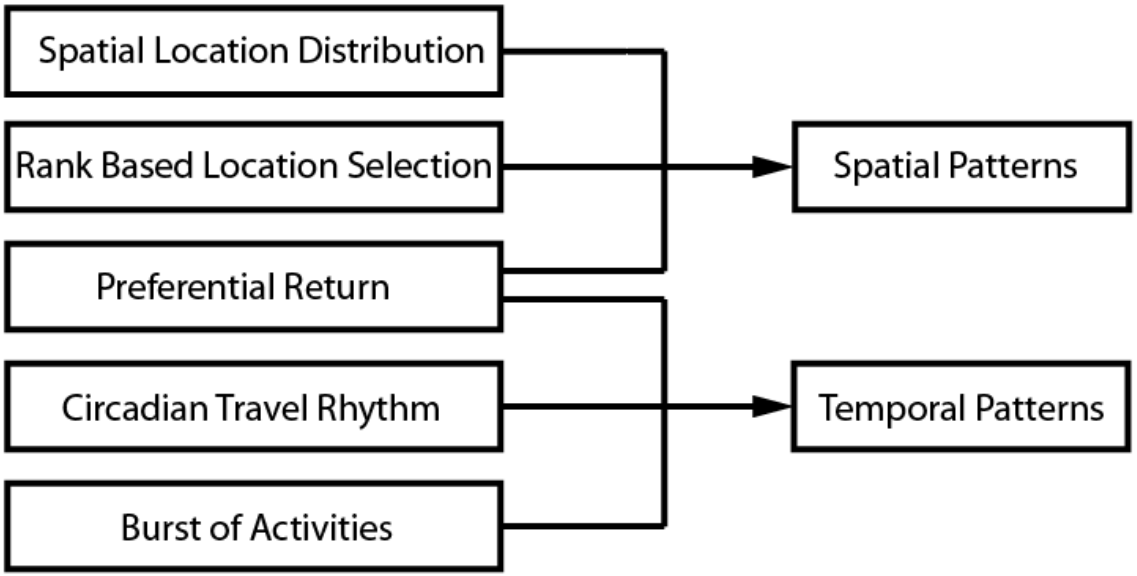


Aggregated results: Boston Trajectories



Summary

Fundamental Mechanisms to generate synthetic trajectories from Mobile phone data.



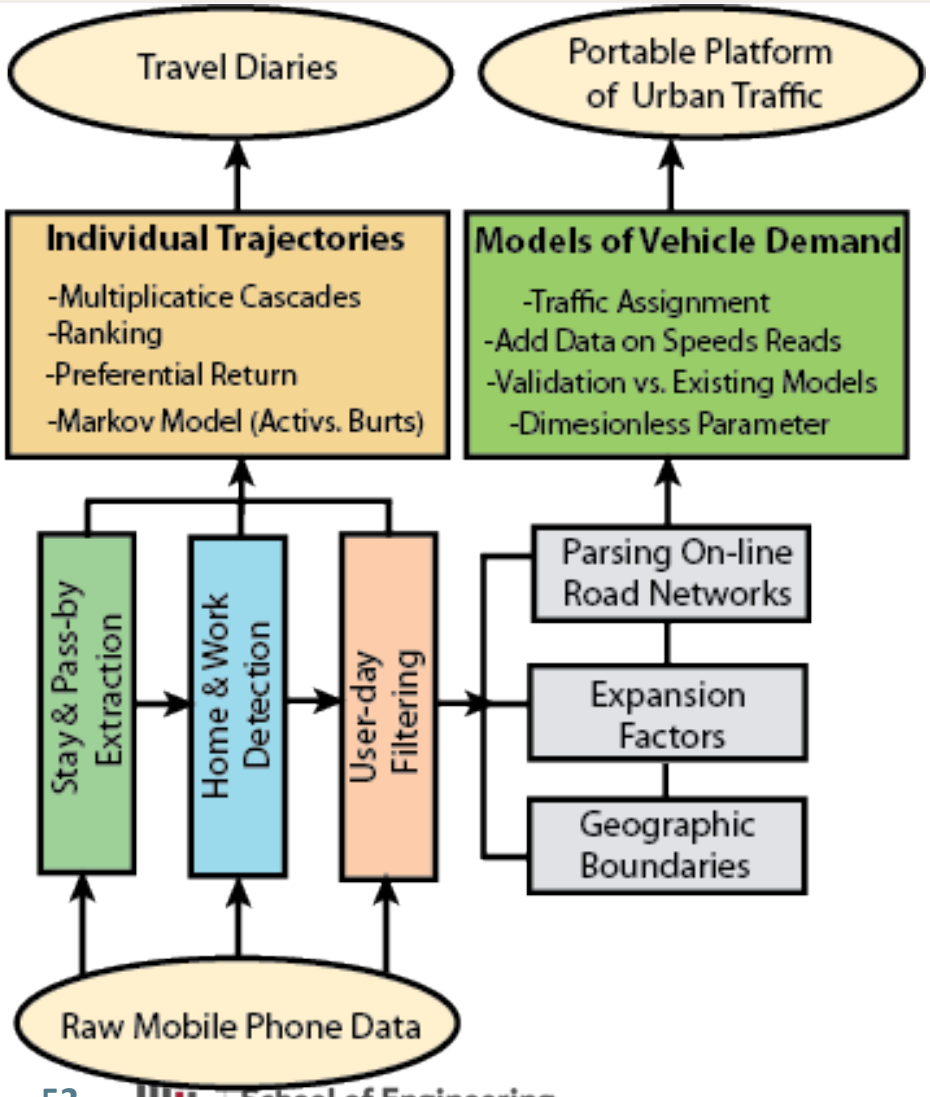
The combined algorithm

Input:
$p(t)$, ρ, γ, α, β_1, β_2, the home location and the set of other locations
Output:
Location at each time step;
Set $t = 0$; $l = home$; $S = 0$; //S is the number of visited location
while $t < t_{max}$ **do**
  if $l == Home$
    if $rand < p(t)$ //decide to move
      if $rand < \frac{\rho S^{-\gamma}}{\rho S^{-\gamma} + 0.6 \times (1 - \rho S^{-\gamma})}$ //Choose a previously unvisited location;
        Choose rank k location with probability $p(k) = \frac{k^{-\alpha}}{\sum_{i=1}^{M} i^{-\alpha}}$;
        $S++$;
        $l = other$;
      else
        Choose a previously visited location k with $p(k) = f(k)$;
        $f(k)++$;
      end if
    end if
  else
    if $rand < \beta_1 p(t)$ //decide to move
      if $rand < \beta_2 p(t)$ //go to a place other than home
        if $rand < \frac{\rho S^{-\gamma}}{\rho S^{-\gamma} + 0.6 \times (1 - \rho S^{-\gamma})}$ //Choose a previously unvisited location.
          Choose rank k location with probability $p(k) = \frac{k^{-\alpha}}{\sum_{i=1}^{M} i^{-\alpha}}$;
          $S++$;
        else
          Choose a previously visited location k with $p(k) \sim f(k)$;
          $f(k)++$;
        end if
      else //go home
        $l = home$;
      end if
    end if
  end if
  $t++$;
end while

HumnetLab, https://github.com/humnetlab/individual mobilitymodel," (2015). >>>> Code Available

Modeling daily trajectories: Universal mechanisms regional effects and individual differences (Yingxiang Yang, Daniele Veneziano, Chaoming Song, Shounak Athavale, Marta C. Gonzalez), Submitted, 2015. [pdf] http://humnetlab.mit.edu/wordpress/publications/

Summary



- HuMNet works at the intersection of statistical physics and machine learning methods to generate urban transportation models. That intersection enables the generation of knowledge from data that cannot be extracted from one discipline alone.
- Our work converts raw mobility data into **surveyless models** of **trip diaries, urban traffic** and **social behavior** at urban scale. This is key for urban and transportation planning.

HUMNET is looking for Collaborations in Funded Research Proposals!!!

Comparison with Traditional Models

	HBW	HBO	NHB	AM 6a-9a	MD 9a-3p	PM 3p-7p	RD 7p-6a	Total
CDR Trips (in Millions)	2.81	7.84	4.73	2.46	4.12	4.15	4.65	15.37
MHTS Trips (in Millions)	2.14	8.99	7.18	3.99	6.24	6.06	2.31	18.61
Tract-pair Correlation	0.30	0.64	0.58	0.42	0.65	0.54	0.40	0.58
Town-pair Correlation	0.96	0.97	0.98	0.97	0.98	0.97	0.96	0.98

Table 2.2: Average daily trips by purpose and period from CDR data and the 2010/2011 Massachusetts Travel Survey (MHTS) [61], as well as the correlation coefficients of CDR and MHTS tract-pair and town-pair trips.

Source	Daily HBW Trips, Millions	Inter-tract Share, %	Inter-town Share, %	Average Trip Length, Miles
CDR	2.11	94	68	9.67
Census	2.10	90	68	10.72

Table 2.3: Comparison of average weekday HW CDR and 2006-2010 CTPP [85] flows.