Big Data to tackle Urban Traffic

Marta C. Gonzalez
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

Marta C. Gonzalez
Today

Information and Communication Technologies (ICT) Passive Data

Urban Traffic

School of Engineering

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 assess 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:

<table>
<thead>
<tr>
<th>Longitude</th>
<th>Latitude</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>-71.059998</td>
<td>42.356132</td>
<td>1266513700</td>
</tr>
<tr>
<td>-71.059730</td>
<td>42.356391</td>
<td>1266513800</td>
</tr>
<tr>
<td>-71.063884</td>
<td>42.355315</td>
<td>1266513900</td>
</tr>
<tr>
<td>-71.063884</td>
<td>42.355315</td>
<td>1266514200</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Estimation precision error:

\~ 300 meter

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

- (a) Graph showing variation of P(h) with departure hour, h
- (b) Graph comparing CDR, BHTS, MHTS, NHTS models
- (c) Graph showing another comparison of P(h) with departure hour, h
- (d) Graph illustrating differences among CDR, BHTS, MHTS, NHTS 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

(a)

- Census Population vs. CDR Residents
- Scaled CDR
- Raw CDR

(b)

- CTPP Workplaces vs. CDR Workplaces
- Scaled CDR
- Raw CDR

Graphs show the relationship between population and workplace counts for origination and destination, highlighting scaled and raw CDR data.
Comparison with Traditional Models
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

\[ f^{\text{rio}}_p = f^{\text{boston}}_p = f^{\text{sfbay}}_p = 1.3 \text{ and } f^{\text{lisbon}}_p = f^{\text{porto}}_p = 1.1. \]

\[ f_{\text{BPR}}(V_oC, f_p) = t_f \times \left(1 + \alpha (V_oC)^\beta\right) \times f_p, \]

\[ \alpha = 0.6 \text{ and } \beta = 4. \]

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.
Paper 2: Unraveling Urban Traffic

CDR Rio de Janeiro + Waze Data

CDR Boston + TomTom Speed reads

Travel Times Validation
Commuting Distance

(a)

\[ f(d) \]

\[ \text{distance, } d, \text{ kms} \]

- rio: \( \mu = 1.7, \sigma = 1.1 \)
- bay: \( \mu = 2.1, \sigma = 0.9 \)
- bos: \( \mu = 1.7, \sigma = 1.1 \)
- lis: \( \mu = 2.0, \sigma = 0.9 \)
- por: \( \mu = 1.6, \sigma = 0.8 \)
Free Traffic Speed Comparison

(b)

Free traffic speed, $v_f$ [km/h]
Traffic Speed Comparison

(b)

\[ f(v_f) \]

\[ f(v_t) \]

free and traffic speeds, \( v_f, v_t, \text{km/hr} \)

- rio, 48
- bay, 55
- bos, 47
- lis, 49
- por, 53

- rio, 30
- bay, 33
- bos, 31
- lis, 41
- por, 41
Commuting Time

\[ t(d) = d \frac{(1 + \Gamma)^\alpha}{v_f} + \beta \]
Spatial Density

(pop. density, $\rho$, people/km$^2$)

distance from center, $r$, km

(f)
\[ \Gamma = \frac{\sum_{e \in E} \ell_e x_e}{\sum_{x_e > 0, e \in E} \ell_e C_e}. \]

\( x_e \)  # Cars in the road link \( e \)
\( \ell_e \)  Road link length in miles
\( C_e \)  Capacity in the road link [cars/miles\(^2\)]
Smart-app (routing)

Modifications on the level of altruism:

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

\[ \lambda = [0..1] \]

User Equilibrium component

Social Optimum component
SO Route $\lambda = 1$
25 mins

UE Route $\lambda = 0$
20 mins

Optional Route $\lambda = 0.2$
22 mins

 Origin

 Destination

(b) 

Rio

Bay

# of vehicle trips

(c) 

Net benefit in commuting travel time, minutes

Social good weight, $\lambda$

% decrease in congestion
## Implementation Approach

<table>
<thead>
<tr>
<th>Stage 1: Strategy</th>
<th>Stage 2: Pilot</th>
<th>Stage 3: Expand</th>
</tr>
</thead>
</table>
| Publish research approach  
  • Incentives  
  • Success  
  • Form partnerships  
  • Secure funding | Solicit proposals  
  • Select host city  
  • Conduct Pilot | Advertise findings  
  • Implement Smart Commute in cities nationwide |

![Logos of CDC, EPA, Massachusetts Institute of Technology, Google, and TomTom](image_url)
Further Details

1) Origin–destination trips by purpose and time of day inferred from mobile phone data
   Lauren Alexander a,b, Shan Jiang b, Mikel Murga a, Marta C. González a
   aDepartment of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States
   bDepartment of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States

2) The path most traveled: Travel demand estimation using big data resources
   Jameson L. Toole a, Serdar Colak a, Bradley Sturt b, Lauren P. Alexander b, Alexandre Evsukoff c, Marta C. González a
   a Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States
   bDepartment of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States
   cDepartment of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States

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

OD generation and Validation (Boston)

OD generation and Validation (Portable Platform)
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 environments, 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 $i = 1 \ldots N - 1$, $j = 1 \ldots M - 1$, is minimized:

$$W(A_i, B_j) = d(a_i, b_j) + \min \left\{ \begin{array}{l} W(A_{i+1}, B_j) \\ W(A_i, B_{j+1}) \end{array} \right.$$

where $A_i$ and $B_j$ are subsequences containing all the elements $1 \ldots i$ from $A$ and $1 \ldots j$ 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.
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.

Results
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.
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.
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)
1 Sample day of a student

1) With a roaming distance $\Delta d_1$ (300 mts.), we cluster spatially close locations within $\Delta d_1$.

2) A time threshold $\Delta t$ (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 location 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

- Daily visited locations, N
- Stay duration, Δt [min]
- f(L)
- Trip distance, Δr [km]

- P(N)
- P(Δt)
- f(L)
- P(Δr)
Summary

Fundamental Mechanisms to generate synthetic trajectories from Mobile phone data.

- Spatial Location Distribution
- Rank Based Location Selection
- Preferential Return
- Circadian Travel Rhythm
- Burst of Activities

Spatial Patterns
Temporal Patterns
The combined algorithm

Input:
\( p(t), \gamma, \alpha, \beta_1, \beta_2 \), the home location and the set of other locations

Output:
Location at each time step;
Set \( t = 0; \ l = \text{home}; S = 0; \) // \( S \) is the number of visited location

while \( t < t_{\text{max}} \) do
  if \( l = \text{Home} \)
    if \( \text{rand} < p(t) \) // decide to move
      if \( \text{rand} < \frac{1}{2}p(t)(1-p(t)) \) // Choose a previously unvisited location:
        Choose rank \( k \) location with probability \( p(k) = \frac{k^{-\alpha}}{\sum_{i=1}^{n} i^{-\alpha}} \);
        \( S + 1; \)
        \( l = \text{other}; \)
      else
        Choose a previously visited location \( k \) with \( p(k) = f(k); \)
        \( f(k) + 1; \)
    end if
  end if
  else
    if \( \text{rand} < \beta p(t) \) // decide to move
      if \( \text{rand} < \frac{1}{2}p(t)(1-p(t)) \) // Choose a previously unvisited location.
        Choose rank \( k \) location with probability \( p(k) = \frac{k^{-\alpha}}{\sum_{i=1}^{n} i^{-\alpha}} \);
        \( S + 1; \)
        \( l = \text{other}; \)
      else
        Choose a previously visited location \( k \) with \( p(k) = f(k); \)
        \( f(k) + 1; \)
    end if
    else // go home
      \( l = \text{home}; \)
    end if
  end if
  \( t + 1; \)
end while


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/
• 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.
is looking for Collaborations in Funded Research Proposals!!!
Comparison with Traditional Models

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

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.

<table>
<thead>
<tr>
<th>Source</th>
<th>Daily HBW Trips, Millions</th>
<th>Inter-tract Share, %</th>
<th>Inter-town Share, %</th>
<th>Average Trip Length, Miles</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDR</td>
<td>2.11</td>
<td>94</td>
<td>68</td>
<td>9.67</td>
</tr>
<tr>
<td>Census</td>
<td>2.10</td>
<td>90</td>
<td>68</td>
<td>10.72</td>
</tr>
</tbody>
</table>

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