Big Data to Change Urban Demand Modeling Paradigms

Marta C. Gonzalez
Why do we need a Transportation Model?

Founded in 1978 has 16 Companies around the World

- Bogota
- Bologna
- Boston
- Leeds
- London
- Los Angeles
- Madrid
- Mexico City
- New Delhi
- New York
- Rome
- San Juan
- Santiago
- Sao Paulo
- Toronto
- Vancouver

- Development planning
- Economics
- Infrastructure
- Local & regional transport
- Rail
- Research & Innovation
- Sports & major events
- Sustainable transport
- Urban transit
Where do they apply a Transportation Model?

Projects

- Transport planning and design for London’s new financial district, Canary Wharf
- Eurotunnel (largest Public Private Partnership project in the world)
- New Lines Programme – developing the case for high-speed rail in the UK
- Multi-modal transport masterplan for Abu Dhabi
- Design of Bogota’s award-winning Transmilenio bus rapid transit network
- 30-year transit masterplan for Sacramento
- Review of air passengers’ rights for European Commission
- Transport planning for the 2012 London Olympics
- Transport plans for World Cup 2014 in Brazil
- Best practice interchange guidelines for London
- Bikeability cycle training
Transportation demand modeling

- Choice models based on the attributes of the transport alternatives and characteristics of travelers

- Attributes
  - Mode (bus, train, auto, air, ships etc.)
  - Level of service
    - Travel time
    - Travel cost
    - Frequency
    - Reliability
    - ... 

- Characteristics of travelers
  - Trip purpose (work, school, social activities etc.)
  - Professional activity
  - Education
  - Age
  - ...
13. **ESTIMATED COST**

Cost estimates must be broken down by the basic project phases. There are two basic classifications, but a breakdown below this level to tasks will be required:

A. Instrument development and design, including determining the sampling technique and preparation for the survey.

B. Survey fielding, retrieval, checking and coding.

It is the intention at present, to choose the contractor as a group of MPOs. The contract will at that point be broken up into 5 contracts with the legal contracts for each MPO being entered into. The development costs will at that stage be pro-rated (based on sample size) to each of the participants, in essence as a per completed household overhead rate. In the event that any one of the participants withdraws at that stage, it is important that there is a mechanism for re-allocating the shared development costs.

14. **PROJECT RESOURCES:**

A. Metro: approximately $550,000
B. State of Oregon (Statewide) approximately $ 500,000
C. Southwest Washington RTC approximately $ 200,000

Source: DOT, 1996, prepared by Cambridge Systematics
Opportunity:
Call data records and ICT (information and communication technologies) from mobile devices is a proxy for individual movement.

Challenge:
Can we use Information and ICT to Simplify the Generation of Urban Transportation Models and change the current paradigm?

Greater Good: Optimize the whole, not an individual!
This will a review work of all our methods, to be published in 2016.
Our Work

• How do people move around the city?
  - Individual Trip Profiles
  - Contextual Mobility Patterns
  - Probabilistic Dynamic Model of City Traffic

• How do we optimize solutions?
  - Portable platform
  - “Routing” Test Case

Mobile Data is the next frontier, in which people and spaces are connected in time!
Information and Communication Technologies (ICT) Discrete Data

- Inference of daily Journeys
- Validated Models of Urban Traffic
- Optimization: Routing Applications

From discrete data to social benefits
What are we using Travel Model for?

We have developed a portable pipeline to generate urban demand models from mobile phone data and we are interested now in its applications for better cities.

Examples of current projects:

1- WithinBlocks: Building Occupancy Models for Energy Modeling
2- Metro: Helping to plan the first metro in Riyadh
3- Planning for Electric Vehicles in the Bay Area
4- Understanding Travel time Reliability
5- Understanding Air quality in Beijing integrating AQ Sensors and a Traffic Model
6- Distributing Travel Demand during the Olympic Games
Doctoral Students

Toole

Masters Students

Alexander

Former Postdocs

Schneider
Belik
Wang
Spatiotemporal Mapping of Mobile Phone for Urban Demand Modeling

Travel Diaries

Portable Platform of Urban Traffic

Individual Trajectories
- Multiplicative Cascades
- Ranking
- Preferential Return
- Markov Model (Activs. Burts)

Models of Vehicle Demand
- Traffic Assignment
- Add Data on Speeds Reads
- Validation vs. Existing Models
- Dimensionless Parameter

Stay & Pass-by Extraction
Home & Work Detection
User-day Filtering

Raw Mobile Phone Data

Parsing On-line Road Networks
Expansion Factors
Geographic Boundaries
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/
Detailed Results Paper 1
Modeling Individual Trajectories

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
Model Individual Trajectories
(resolution 10min and 300mts radius)
Stay Extraction
How to detect a stay?

- Consecutive records within a circle of certain radius (300m)
- The time interval between the first/last record is larger than certain threshold (10min)
Stay region extraction

From stay to stay region

- Stay Region: stays from different trajectories might represent the same location
- Interchangeable with “location”
## Stay region extraction

### How to get stay regions?

- Grid based clustering
- Divide the entire area into a grid
- Merge stays in neighboring grid cells

![Diagram of grid-based clustering for stay region extraction](image)
Stay Extraction
Stay Extraction
Measuring Individual Activities: Home, Work, and Other

• A phone user’s “home” is defined as
  – the most frequently region 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 stay region working hours of the weekday

A phone user’s “other” is defined as the rest
Stay Extraction

Same user can re-locate by rechecking the criteria periodically
Reconstructing daily journeys is easier for active users.
Distribution of Phone Records per individual

Group II >> Sparse users

Group III and IV >> Representative Active Users
We can learn from the behavior of active users to effectively model sparse users?

1. What is the likelihood of a trip to occur at different times of the days?

2. How long do they stay in flexible activities?

3. What is the like-hood to visit new Locations?

4. How do they select them?

5. How many trips I likely do per week?
Explore vs. Return

Like hood to be found in the most visited Location by time of the day

Number of Locations by Time of the Day

Probability of Exploring having S locations


Ranking of POIs to select New Destination
What is the probability of departing from home to do a flexible activity at time $t$?

$$P_i(t) = n^i_w P_d(t)$$
Time independent Markov Model

\[ P(t) = n_w P_d(t) \]

\[ P_1 = 1 - P(t) \]  
Moving from “home”

\[ P_2 = 1 - \beta_1 P(t) \]  
Moving from “other”

\[ P_3 = \beta_1 P(t) \beta_2 P(t) \]  
Moving from “other” to “other”

\[ \beta_1 \]  
Generates shorter stays in the flexible location state.

\[ \beta_2 \]  
Generate Different number of activities in a row per active cycle
Measure the 3 key features from Active Users
The Model

Features Extracted from data of Active Users

Global Trip prob.  Individual Mobility Rates  Preferential Return  Ranking of Explorations

Flowchart of the Model

Choose to move?  Explore?  
Yes  Yes  No  No

1 - \( P(t) \)  \( \beta_1 P(t) \)  \( \beta_2 P(t) \)  \( 1 - \beta_2 P(t) \)

Yes  No  Yes  No

At home?  Exploring  Pref. Return  Ranking  Other

Home  Other  Other  Other

\( P(K) \)  \( k^{-0.86} \)
Models Results

Modeled Trajectories of our Student and see the results (only home and work data used)
Individual Mobility Patterns

Student vs. his synthetic version

- Daily visited locations, $N$ vs. $P(N)$
- Stay duration, $\Delta t$ [min] vs. $P(\Delta t)$
- $L^{th}$ most visited location vs. $f(L)$
- Trip distance, $\Delta r$ [km] vs. $P(\Delta r)$
Models Results

Modeled Trajectories From Sparse Data of sample User (previous locations used)
Models Results

The combined algorithm

Input:
- $x_0$, the home location and the set of other locations
- $P$, the probability of moving to other locations
- $\beta$, the home location and the set of other locations
- $\gamma$, the number of visited locations
- $\epsilon$, the probability of staying at home

Output:
- $\hat{x}$, the predicted location

1. Set $\hat{x} = x_0$.
2. For $i = 1, 2, \ldots, \gamma$:
   a. If $\epsilon < \frac{1}{\epsilon}$, go to step 3.
   b. Choose a location from $P$ (a previously visited location).
   c. If $i = \epsilon$, set $\hat{x}$ to the chosen location.
   d. Else, choose a location from $\gamma$.
   e. If $\epsilon < \frac{1}{\epsilon}$, set $\hat{x}$ to the chosen location.
   f. $i = i + 1$.
3. End while.
Are Active users Representative?
Comparison with Traditional Models
The path most traveled: Travel demand estimation using big data resources

Jameson L. Toole a,1, Serdar Colak b,2, Bradley Sturt a, Lauren P. Alexander b, Alexandre Evseevoff b, Marta C. González a,b

a Engineering Systems Division, MIT, Cambridge, MA 02139, United States
b Department of Civil and Environmental Engineering, MIT, Cambridge, MA 02139, United States

COPYRIGHT Federal University of Rio de Janeiro, Brazil

Contents lists available at ScienceDirect
Transportation Research Part C
journal homepage: www.elsevier.com/locate/trc
Commuting Distance

(a)

\[ f(d) \]

\[ \text{straight-line distance, } d, \text{ kms} \]

- \( \text{rio: } \mu = 1.6, \sigma = 1.2, KS = 0.049 \)
- \( \text{bay: } \mu = 2.1, \sigma = 0.9, KS = 0.023 \)
- \( \text{bos: } \mu = 1.7, \sigma = 1.1, KS = 0.028 \)
- \( \text{lis: } \mu = 2.0, \sigma = 0.9, KS = 0.021 \)
- \( \text{por: } \mu = 1.6, \sigma = 0.7, KS = 0.018 \)
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

\[ f_{\text{rio}} = f_{\text{boston}} = f_{\text{sf bay}} = 1.3 \quad \text{and} \quad f_{\text{lisbon}} = f_{\text{porto}} = 1.1. \]

\[ f_{\text{BPR}}(V_{oC}, f_p) = t_f \cdot (1 + \alpha (V_{oC})^\beta) \cdot f_p, \]

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

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.
Validation of CDR OD assignment
Free Traffic Speed Comparison

(b)

\[ f(v_f) \]

Free traffic speed, \( v_f \) [km/h]

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

(b)

free and traffic speeds, $v_f, v_t, \text{km/hr}$
Commuting Time

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

(c) Travel time vs. route distance, \( d, \text{kms} \)

- rio: 2.0\(d\) + 2.97, \(\alpha = 1.9\)
- bay: 1.8\(d\) - 0.19, \(\alpha = 1.8\)
- bos: 1.6\(d\) + 2.76, \(\alpha = 1.2\)
- lis: 1.6\(d\) + 1.94, \(\alpha = 1.5\)
- por: 1.3\(d\) + 1.49, \(\alpha = 1.3\)
% 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:

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

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

User Equilibrium component

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

UE Route $\lambda = 0$ 20 mins

Optional Route $\lambda = 0.2$ 22 mins

(a)

(b)

(c)

% decrease in congestion

net benefit in commuting travel time, minutes

# of vehicle trips

social good weight, $\lambda$

rio

bay

$\lambda = 1.0$ $\lambda = 0.1$

$\lambda = 1.0$ $\lambda = 0.1$

$\lambda = 1.0$ $\lambda = 0.1$

(b)
$\lambda=0.3$ is a good compromise
Next steps: Travel Time Reliability
Recommendations for the new Riyadh Metro System (METRO)

- Methodology for optimizing catchment zones based on potential demand of the new metro system and to support the design of the bus system.

- Scenario analysis of various travel time outputs from diverse coupling mechanisms of the road vehicle trips (cars and buses) with the metro system.
Coupled Networks approach: A small network 2 (metro) is coupled with the larger network 1 (road network). We will explore various coupling mechanisms that result in the best performance in terms of travel times for existing conditions of OD demand and travel times.
Current Application: Help Distribute Demand in the Rio Olympic Games

Rio de Janeiro
1.2 million people
average of ~16 calls per day
1,411 cell phone towers
Overview of research

• Working at the intersection 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.

• Current models of travel demand are based in Census and Activity Surveys which get outdated and are expensive and limited in size. While massive digital traces (phone usage, on-line activities, GPS records) are passively generated geo-located signals of human activities. But these signals are sparse and incomplete.

• We can convert raw data into surveyless models of trip diaries, urban traffic and social behavior at urban scale. This is key for urban and transportation planning.
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

Logos:
- CDC
- Massachusetts Institute of Technology
- Google
- TomTom