Activity-based travel demand modeling with cellular data

Prof Alexei Pozdnoukhov

Systems & Transportation
UC Berkeley
alexep@berkeley.edu

This is total chaos!
Let’s call an engineer!
And so we did...
The art of forecasting as “being able to forecast what will happen and then explain why it did not happen”....

...fully applies to travel demand forecasting.
Urban mobility meets big data

1. For travel demand models and transportation system design
2. For system management and traffic flow control

- The data we need are on individual travellers, not infrastructures
- Such mobility data are expensive to collect with traditional surveys
- Traditional demand models are difficult to build, takes up to 5 years
- Traffic sensing is based on ageing infrastructure with high running costs
Urban Mobility Data: The Ecosystem

Private sector mobility providers
Total investment by 2014(*):

- Uber $5.9B
- Lyft $1.0B
- Hailo $100.6M
- Sidecar $35M
- Flywheel $12M
- Leap Bus $100T
- Carma $10.1M
- Zirx $36.4M
- Cabify $15M
- Gett $200M
- Getaround $43M
- Cabify $15M
- Leap Bus $100T
- Zirx $36.4M

Net revenue 2014(*):

- Uber $2B
- Lyft $130M
- Zipcar $68M
- Getaround x7

Public sector agencies
Operations/year:
- Caltrans $10B
- SFMTA $1B
- MTC $400M

Infrastructure costs:
- $6.4B Bay Bridge
- $68B High Speed Rail

Net revenue 2014(*):

- $12M + $300K/year

Disruptive Mobility Services
Citizens
Choice Context
City

Private GeoSocial Data
Travel Decision Support Products
GTFS, Maps

Public Infrastructure Planning and Policy
Demand Model
Travel Surveys

MTC/DOT, etc.

~ 6 -12 months
~ 5-10 years
$12M + $300K/year

Influence
Data Flow
Data

Cellular data (CDR* and more)
- collected by telecoms for billing purposes
- high volumes
- reasonable space-time resolution
- implicit social network of users
- no access to content

* Call Detail Records

Social media data
- big internet players, “public” API
- growing volumes
- variable space-time resolution
- partly networked
- content
Cellular data

Aggregates over population present within a given cell in the network

<table>
<thead>
<tr>
<th>Day</th>
<th>8am</th>
<th>9:30am</th>
<th>12pm</th>
<th>1:30pm</th>
<th>6pm</th>
<th>8:30pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 2</td>
<td>8:30am</td>
<td>10am</td>
<td>11:30am</td>
<td>13pm</td>
<td>5:30pm</td>
<td>6:30pm</td>
</tr>
<tr>
<td>Day 3</td>
<td>8am</td>
<td>9:30am</td>
<td>12pm</td>
<td>1:30pm</td>
<td>6pm</td>
<td>8:30pm</td>
</tr>
<tr>
<td>Day 4</td>
<td>8:30am</td>
<td>11am</td>
<td>1pm</td>
<td>5:30pm</td>
<td>7:30pm</td>
<td></td>
</tr>
</tbody>
</table>
1. Calibrate a “sensing function” (prior on infrastructure & signal strength)

\[ p(x|x_i, r_i) \]

2. IPFP (iterative proportional fitting), with daytime/nighttime groundtruth

3. Downscale to a known map

\[ R_{vic}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \int f(x)p(x|x_i, r_i)dx)^2 \]

\[ f(x) = \sum_{j=1}^{\ell} \alpha_j \int K(x, x')p(x'|x_j, r_j)dx' \]

[Pozdnoukhov & Kaiser, 2011; Pozdnoukhov et al (in progress)]
Spatial kernel methods

Kernel methods

- Support Vector machines,
- Support Vector regression,
- Kernel ridge regression,
- Support Vector novelty detection

[Vapnik, 1995; Scholkopf and Smola, 2002]

Kernels on spatial support

\[
\tilde{k}(x, x') = k(x, x') - k_x^T (I + \gamma L K)^{-1} \gamma L k_{x'}
\]

[Pozdnoukhov, 2010]
Incremental learning of a Support Vector dictionary (hypothesis space)

\[ f(.) \approx \sum_{i=1}^{M} \alpha_i K(., x_i) \]

\( \{x, y\} \)

\( \{\alpha_i\} \)

\( \langle \Phi(\cdot), \Phi(x_i) \rangle_H \)

\( x_{OLD} \)

\( O(m^2) \): size of sensing infrastructure

\( O(n) \), const time per sample in a stream

Population density

10:00 am

[Kaiser & Pozdnoukhov, 2012]
Population density

11:00 am

[Kaiser & Pozdnoukhov, 2012]
Population density

12:00 am

[Kaiser & Pozdnoukhov, 2012]
Population density

1:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

2:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

3:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

4:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

5:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

6:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

7:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

8:00 pm

[Kaiser & Pozdnoukhov, 2012]
Population density

[Kaiser & Pozdnoukhov, 2012]
Traffic density

[Yadlowsky et al, 2015]
Traffic density

Calibrated cellular coverage

Non-calibrated cellular coverage
Connecting the dots into OD flows

<table>
<thead>
<tr>
<th>Day</th>
<th>8am</th>
<th>9:30am</th>
<th>12pm</th>
<th>1:30pm</th>
<th>6pm</th>
<th>8:30pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 2</td>
<td>8:30am</td>
<td>10am</td>
<td>11:30am</td>
<td>13pm</td>
<td>5:30pm</td>
<td>6:30pm</td>
</tr>
<tr>
<td>Day 3</td>
<td>8am</td>
<td>9:30am</td>
<td>12pm</td>
<td>1:30pm</td>
<td>6pm</td>
<td>8:30pm</td>
</tr>
<tr>
<td>Day 4</td>
<td>8:30am</td>
<td>11am</td>
<td>1pm</td>
<td>5:30pm</td>
<td>7:30pm</td>
<td>8:30pm</td>
</tr>
</tbody>
</table>

“Home”    “Work”    “Home”
Travel Demand: OD aggregates

Californian Household Transportation Survey

Origin

Destination

1454 x 1454 OD tables
SF Bay Area

9 counties

Population: 7’500’000

Commuters: 3’400’000

Transit: 350’000 (10%)

Time to work: 28 mins
Individual level data: connecting the dots into route flows
Route flow inference

All flows are in 1000 vehicles/hour.

Pointpath flows:
\[
\begin{align*}
  f_{p1234} &= 1 &= x_1 \\
  f_{p1654} &= 4 &= x_2 \\
  f_{p654} &= 10 &= x_3 + x_4
\end{align*}
\]

OD demands:
\[
\begin{align*}
  d_{AB} &= 5 &= x_1 + x_2 \\
  d_{CB} &= 10 &= x_3 + x_4
\end{align*}
\]

Link flow: \( b = 9 = x_2 + x_3 \)

\[(Ux = f): \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} x = \begin{bmatrix} f_{p1234} \\ f_{p1654} \\ f_{p654} \end{bmatrix} \]

\[(Tx = d): \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} x = \begin{bmatrix} d_{AB} \\ d_{CB} \end{bmatrix} ; \quad (Ax = b): \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix} x = b \]

Sol. with pointpaths: \( x^* = [1 4 5 5]^T \); sol. with ODs: \( x = x^* + [1 -1 1 -1]^T t, \forall t \in [-1, 4] \)
Route flow inference

Helps answering key questions from data (with no user equilibrium or other traffic assignment assumptions involved):

1. How to quantify individual contributions to congestion (from where to where)
2. In case of link obstruction, percentage of commuters affected by it
3. What should policies be to avoid “un-coordinated” reroutes?

Route flows: from any O to any D

Obstructed link

Green (less than 5%, red: more than 30%)
Route flow inference

[Wu et al, 2015]
Individual level data: connecting the dots into sequences of activities

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8am 9:30am 12pm 1:30pm 6pm 8:30pm</td>
<td>8:30am 10am 11:30am 13pm 5:30pm 6:30pm 9pm</td>
<td>8am 9:30am 12pm 1:30pm 6pm 8:30pm 8am 9:30pm</td>
<td>8:30am 11am 1pm 5:30pm 7:30pm</td>
</tr>
<tr>
<td>“Home”</td>
<td><img src="image" alt="Home" /></td>
<td><img src="image" alt="Home" /></td>
<td><img src="image" alt="Home" /></td>
<td><img src="image" alt="Home" /></td>
</tr>
<tr>
<td>“Work”</td>
<td><img src="image" alt="Work" /></td>
<td><img src="image" alt="Work" /></td>
<td><img src="image" alt="Work" /></td>
<td><img src="image" alt="Work" /></td>
</tr>
<tr>
<td>“Shop”</td>
<td><img src="image" alt="Shop" /></td>
<td><img src="image" alt="Shop" /></td>
<td><img src="image" alt="Shop" /></td>
<td><img src="image" alt="Shop" /></td>
</tr>
<tr>
<td>“Home”</td>
<td><img src="image" alt="Home" /></td>
<td><img src="image" alt="Home" /></td>
<td><img src="image" alt="Home" /></td>
<td><img src="image" alt="Home" /></td>
</tr>
</tbody>
</table>
Modelling choices

Big Data vs

(10\textsuperscript{th} of DCMs with 100\textsuperscript{th} of parameters in utility functions)
Activity-based modelling

1) Home, Work, School, + secondary places
   - Heuristics, time thresholds
   - Mixture models
   - Hidden Semi-Markov Models

2) Travel mode recognition
   - Clustering
   - Supervised classification
Learning activity sequences

Neural-network Hidden Markov Model (IO-HMM)
The SmartBay

Berkeley

UNIVERSITY OF CALIFORNIA
The SmartBay project.

Micro-simulation of SF Bay Area mobility from cell phone data.

1M agents, 500 shown in visualization

(c) http://www.matsim.org/
Validation on freeway volume counts

[McArdle et al., 2012]
Future work - I

Model travel choices (mode, destination) in the context of accessibility and social influence; mode adoption (TNCs, EVs)
Future work – II: NSF CRISP

Multi-scale Infrastructure Interactions with Intermittent Disruptions: Coastal Flood Protection Infrastructure, Transportation and Governance Networks

5 meters SL rise + 100y storm event
Future Work III - TNCs (Uber/Lyft)

Scenario: TNC fleets serving a given population
Supply: 2 fleets with dynamic vehicle routing, demand responsive
Demand: calibrated from SFMTA taxi trips data
Network: SF with background traffic for a typical weekday
Conclusions

• accurate demand forecasting, both for planning and operations
• high fidelity activity-based traffic micro-simulations, “what if” scenarios
• social influence in destination choice, mode choice, route choice, novel concepts of demand management informed by data
• insights for novel mobility concepts: on-demand transportation, car/ride-sharing, flexible services and multi-modality
• quantitative behavioral studies, geography of social networks, influence propagation, travel behavior and mobility lifestyle adoption
Thank you!

alexeip@berkeley.edu

SmartBay team
Mogeng Yin
Andrew Cambell
Sid Feygin
Maddie Sheehan
Sudatta Mohanty

MegaCell team
Alex Bayen
Cathy Wu
Steve Yadolovsky
Jerome Thai

Jean-Francois Paiement
Colin Goodall
Ann Skudlark
Chris Volinsky

Jeff Pang
(now with Databricks)