Travelers' Behavior Modeling, Demand Estimation and Traffic Management: a Data Perspective

Wei Ma
Mobility Data Analytics Canter (MAC)
Carnegie Mellon University
Outline

- Background
- Behavior modeling: Statistical Traffic Assignment
- Demand estimation: Probabilistic O-D Estimation
- Dynamic Network Loading
- Real-time Traffic Management
Background

<table>
<thead>
<tr>
<th>Ranking</th>
<th>City</th>
<th>Hours spent in congestion</th>
<th>Total Cost per Driver</th>
<th>Total Cost per City</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>104</td>
<td>$2,408.00</td>
<td>9.7B</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>New York</td>
<td>89</td>
<td>$2,533.00</td>
<td>16.9B</td>
</tr>
<tr>
<td>3</td>
<td>San Francisco</td>
<td>71</td>
<td>$1,861.00</td>
<td>3.1B</td>
</tr>
</tbody>
</table>

All data and insights provided by INRIX

http://inrix.com/resources/inrix-2016-traffic-scorecard-us/
Background

<table>
<thead>
<tr>
<th>Ranking</th>
<th>City</th>
<th>Hours spent in congestion</th>
<th>Total Cost per Driver</th>
<th>Total Cost per City</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Los Angeles</td>
<td>104</td>
<td>$ 2,408.00</td>
<td>9.7B</td>
</tr>
<tr>
<td>2</td>
<td>New York</td>
<td>89</td>
<td>$ 2,533.00</td>
<td>16.9B</td>
</tr>
<tr>
<td>3</td>
<td>San Francisco</td>
<td>71</td>
<td>$ 1,861.00</td>
<td>3.1B</td>
</tr>
<tr>
<td></td>
<td>Pittsburgh</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Percentage of all driving time spent in congestion per driver: 9%

Hours spent in congestion by the average U.S. commuter: 42 hours

Total cost of congestion to all drivers in the U.S.: 300B

Cost of congestion to the average U.S. driver: $1,400

All data and insights provided by INRIX

Carnegie Mellon University
Civil and Environmental Engineering

http://inrix.com/resources/inrix-2016-traffic-scorecard-us/
## Background

**9%**

Percentage of all driving time spent in congestion per driver

**42**

Hours spent in congestion by the average U.S. commuter

**300B**

Total cost of congestion to all drivers in the U.S.

**$1,400**

Cost of congestion to the average U.S. driver

---

<table>
<thead>
<tr>
<th>Ranking</th>
<th>City</th>
<th>Hours spent in congestion</th>
<th>Total Cost per Driver</th>
<th>Total Cost per City</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Los Angeles</td>
<td>104</td>
<td>$ 2,408.00</td>
<td>9.7B</td>
</tr>
<tr>
<td>2</td>
<td>New York</td>
<td>89</td>
<td>$ 2,533.00</td>
<td>16.9B</td>
</tr>
<tr>
<td>3</td>
<td>San Francisco</td>
<td>71</td>
<td>$ 1,861.00</td>
<td>3.1B</td>
</tr>
<tr>
<td>25</td>
<td>Pittsburgh</td>
<td>33</td>
<td>$ 1,062.00</td>
<td>944M</td>
</tr>
</tbody>
</table>

---

All data and insights provided by INRIX

---

**Carnegie Mellon University**

Civil and Environmental Engineering

http://inrix.com/resources/inrix-2016-traffic-scorecard-us/
Background

What causes traffic congestions?

- Supply:
  - Network capacity --- more roads
  - Road capacity --- broader roads, forbid on-street parking
  - Speed density relationship --- autonomous vehicle

- Demand:
  - Heavy demand --- public transit, ride sharing
  - Imbalanced demand --- connected vehicle, information center
Background

A fundamental problem for all the traffic management models:

- How to get:
  - Network Conditions
  - Traffic Demand
  - Travelers’ Behavior
Background
Background

SUPPLY SIDE

Carnegie Mellon University
Civil and Environmental Engineering
Background

What are the demand and supply for a road network?

DEMAND SIDE

• Origin – Destination demand (O-D) demand

• Previously estimated by the density of residence, not accurate and not flexible

• We want to estimate the O-D demand from various traffic data.
86499 O-D pairs
Traffic flow
Traffic speed
Background

**OBJECTIVE**

Infer the network conditions, travelers’ behavior and traffic demand from partially observed traffic data.

**Input:** partially observed traffic data  
**Output:** whole network condition, underlying behavior
Background

Look the problem as a regression model

\[ Y = f(q, s) + e \]

(speed, flow) = Travelers’ behavior(O-D demand, network supply) + error
Statistical Traffic Assignment

\[(\text{speed, flow}) = \text{Travelers’ behavior (O-D demand, network supply)} + \text{error}\]

Daily time-varying traffic counts on SR41 SB and NB

- Statistical traffic assignment explores the statistical features of recurrent flow patterns

\[f: Q \leftrightarrow X, F\]
Statistical Traffic Assignment

- O-D demand variation, route choice variation, unknown error
- Variance and covariance in O-D demand
- Random selection of the route
- Unknown error caused by measurement
Statistical Traffic Assignment

How to generate traffic?

- Travelers’ Information Structure: Know the distribution of path costs
- Travelers’ Route Choice: Probability based

Number of travelers → Route to choose → Road Network Conditions

Multivariate Normal Distribution → Multinomial Distribution

http://www.nationalpower.info/winning-probability/
Statistical Traffic Assignment

- **OD**
  - MVN: \( Q \sim N(q, \Sigma_q) \)
  - Scalable, stable and consistent

- **Route choice**
  - \( p \) is deterministic for a recurrent traffic network
  - A generalized probability function: \( p = \psi(C; \Theta) \)

- **Path flow**
  - Path flow follows multinomial distribution
  - \( F_{rs} | Q_{rs} \sim MN(Q_{rs}, p_{rs}) \)
  - Normal approximation

- **Link flow**
  - \( X = \Delta F \), where \( \Delta \) is the path/link incidence matrix

- **Path cost**
  - A generalized path cost function: \( C \sim t(X; \Theta) \)
  - Normal approximation
Model: Hierarchy

Level 1:
\[ X_m \sim N(X + e, \Sigma_x + \Sigma_e) \]
Unknown error

Level 2:
\[ X \sim N(\Delta p Q, \Sigma_x) \]
Route choice variation
\[ F \sim N(p Q, \Sigma_f) \]

Level 3:
\[ Q \sim N(q, \Sigma_q) \]
OD variation

- Level i is conditional on level i+1
- Each level reflects one single source of the link flow variation
Model: Property

- Existence and consistence of the solution
  - Fixed point problem
  - If the path cost function is continuous, the solution \((x, \Sigma_x, f, \Sigma_f, p)\) exists.
  - Once the solution \((x, \Sigma_x, f, \Sigma_f, p)\) is determined, then it’s consistent.

- Reducible
  - Model can be reduced to existing statistical assignment models

- Data driven
  - Friendly to model learning techniques for large scale networks
  - Dimension reduction, sparse regularization, model selection
Model: Application

- Marginal distribution

  Marginal distribution of $X_m$ can be written as:
  \[
  X_m \sim N(x_m, \Delta p \Sigma q p^T \Delta^T + \Delta \Sigma_{f|q} \Delta^T + \Sigma_e)
  \]

  Three matrices represent the variance from OD, route choice and measurement error separately.

- Variance ratio

  - To measure the portion of each source of variance
  - To help to reduce the system variance
  - Trace norm, nuclear norm
Experiment

- **Settings**

- OD: 1 -> 3, $q = 1000$, $\sigma^2 = 10000$
- Route choice model: Probit + Multinomial
- Link cost function: BPR function
- Measurement error: $e = 0$, $\sigma_e^2 = 100$
Experiment

- Basic results

Network condition

Confidence interval
Large Scale Network

- Settings
  - SR41 freeway network: 2413 links and 7110 OD pairs
  - OD demands were carefully calibrated
  - Assume the O-D demand variance is 20% of its mean

- Results
  - Terminated in 9 iterations and 467 seconds
• Using Rectangle CI approximation.
• Red represents volume/capacity > 1, and green represents volume/capacity = 0, other colors are smoothly transitioned from green to red as volume/capacity increases from 0 to 1)
Background

Look the problem as a regression model

\[ Y = f(q, s) + e \]

(speed, flow) = Travelers’ behavior(O-D demand, network supply) + error
Probabilistic O-D Estimation

- Objective: estimate the mean and variance/covariance matrix of O-D demand from day-to-day traffic data
- Data: partial observations $X^o_m$, 10% of the total road segments, speed on major roads
Probabilistic O-D Estimation

Why this is difficult?

• Hierarchical statistical model

• For Pittsburgh area: number of traffic flow data: ~2300
  number of traffic speed data: ~1000
  number of O-D pair: ~86000

• Under-determined problem, multiple solutions for O-D demand

• Number of parameters in variance/covariance matrix for O-D demand: 86000 * 43000!
Probabilistic O-D Estimation

Equilibrium:

- Rational travelers in recurrent traffic conditions
- If link 1 and link 2 are exactly the same, then the traffic flow on both links should be the same
- Multiple player equilibrium: Nash Equilibrium
Probabilistic O-D Estimation

General Framework:

• Iterative estimation framework (IGLS).
• EM algorithm in ML
Probabilistic O-D Estimation

**Estimate OD mean:**

GLS, equilibrium constraint, single level relaxation

$$\min \limits_f n \left( \Delta^o f - \hat{x}^o \right)^T \Sigma_x^{o-1} \left( \Delta^o f - \hat{x}^o \right) + (q^H - M f)^T \Sigma_q^{H-1} (q^H - M f)$$

s.t. $f \in \Phi^+$

**Estimate OD variance:**

MLE, convex relaxation, LASSO model selection.

$$\min \limits_{\Sigma_q} \| S^o_x - \Sigma_x^o \|_F^2 + \lambda \| \Sigma_q \|_1$$

s.t. $\Sigma_x^o = \Delta^o \Sigma f_{|q} \Delta^o T + \Delta^o \hat{p} \Sigma q \hat{p}^T \Delta^o T$

$\Sigma_q \in$ semidefinite($\mathbb{R}^{|K_q| \times |K_q|}$)
Probabilistic O-D Estimation

Model Observability:

• Non-unique solution
• No worse than deterministic O-D estimation

Goodness of fit:

• Distribution based
• Hellinger distance or Kullback-Leibler distance

\[ D_H((\mu_1, \Sigma_1)^T, (\mu_2, \Sigma_2)^T) = 1 - \frac{|\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}}}{\left(\frac{1}{2} \Sigma_1 + \frac{1}{2} \Sigma_2\right)^{\frac{1}{2}}} \exp \left( -\frac{1}{8} (\mu_2 - \mu_1)^T \left( \frac{1}{2} \Sigma_1 + \frac{1}{2} \Sigma_2 \right)^{-1} (\mu_2 - \mu_1) \right) \] (27)

\[ D_{KL}((\mu_1, \Sigma_1)^T, (\mu_2, \Sigma_2)^T) = \frac{1}{2} \left( \log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{tr} (\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right) \] (28)
# Probabilistic O-D Estimation

Link 1 and Link 3 Observed, 500 samples

<table>
<thead>
<tr>
<th>True $\rho$</th>
<th>Method</th>
<th>$\hat{q}_{1\rightarrow2}$</th>
<th>$\hat{q}_{3\rightarrow2}$</th>
<th>$\hat{\sigma}_{1\rightarrow2}$</th>
<th>$\hat{\sigma}_{3\rightarrow2}$</th>
<th>$\hat{\rho}$</th>
<th>RMPSE</th>
<th>KL-distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>True</td>
<td>700</td>
<td>500</td>
<td>13.23</td>
<td>11.18</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>0.5</td>
<td>w/o EC - w/o Lasso</td>
<td>611.02</td>
<td>588.01</td>
<td>14.18</td>
<td>11.46</td>
<td>0.37</td>
<td>14.64%</td>
<td>106.80</td>
</tr>
<tr>
<td></td>
<td>Logit - w/o Lasso</td>
<td>728.55</td>
<td>588.85</td>
<td>12.81</td>
<td>11.07</td>
<td>0.55</td>
<td>11.00%</td>
<td>33.78</td>
</tr>
<tr>
<td></td>
<td>Probit - w/o Lasso</td>
<td>618.63</td>
<td>590.17</td>
<td>15.34</td>
<td>11.36</td>
<td>0.45</td>
<td>14.31%</td>
<td>101.69</td>
</tr>
<tr>
<td>0</td>
<td>w/o EC - w/o Lasso</td>
<td>765.43</td>
<td>588.78</td>
<td>11.95</td>
<td>10.25</td>
<td>0.03</td>
<td>12.90%</td>
<td>43.08</td>
</tr>
<tr>
<td></td>
<td>Logit - w/o Lasso</td>
<td>727.94</td>
<td>588.05</td>
<td>13.05</td>
<td>11.00</td>
<td>0.05</td>
<td>10.89%</td>
<td>33.24</td>
</tr>
<tr>
<td></td>
<td>Probit - w/o Lasso</td>
<td>618.58</td>
<td>587.79</td>
<td>13.79</td>
<td>11.16</td>
<td>0.07</td>
<td>14.02%</td>
<td>49.07</td>
</tr>
<tr>
<td></td>
<td>Logit - w/ Lasso</td>
<td>728.27</td>
<td>588.48</td>
<td>12.87</td>
<td>11.53</td>
<td>0.00</td>
<td>10.95%</td>
<td>33.60</td>
</tr>
<tr>
<td></td>
<td>Probit - w/ Lasso</td>
<td>621.39</td>
<td>588.56</td>
<td>13.71</td>
<td>11.67</td>
<td>0.00</td>
<td>13.94%</td>
<td>49.04</td>
</tr>
<tr>
<td>−0.5</td>
<td>w/o EC - w/o Lasso</td>
<td>780.12</td>
<td>586.04</td>
<td>7.80</td>
<td>11.18</td>
<td>−0.58</td>
<td>13.85%</td>
<td>95.26</td>
</tr>
<tr>
<td></td>
<td>Logit - w/o Lasso</td>
<td>726.41</td>
<td>586.08</td>
<td>12.99</td>
<td>11.12</td>
<td>−0.58</td>
<td>10.61%</td>
<td>52.44</td>
</tr>
<tr>
<td></td>
<td>Probit - w/o Lasso</td>
<td>621.40</td>
<td>588.66</td>
<td>13.29</td>
<td>10.68</td>
<td>−0.37</td>
<td>13.79%</td>
<td>32.90</td>
</tr>
</tbody>
</table>

w/o EC: without equilibrium constraint; Logit/Probit: using Logit/Probit based SUE constraint; w/o Lasso: without Lasso regularization; w/ Lasso: using Lasso regularization

---

Carnegie Mellon University
Civil and Environmental Engineering
Probabilistic O-D Estimation

- Logit - w/ Lasso - w/ History O-D method on a desktop computer (Inter(R) Core 529 i5-4460 3.20 GHz 2, RAM 8 GB)
- Average computation time for each IGLS iteration is 233.05s.
- Converge in 20 iterations.
<table>
<thead>
<tr>
<th>Model complexity</th>
<th>Model reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic</td>
<td>Deterministic</td>
</tr>
<tr>
<td>Static</td>
<td>Stochastic</td>
</tr>
<tr>
<td>Dynamic</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Static</td>
<td>Static</td>
</tr>
</tbody>
</table>
Moving to Dynamic Network

• Static O-D demand --- Dynamic O-D demand
• Travelers’ behavior --- **Dynamic network loading**
• Static traffic condition --- Network flow evolution

Time varying speed, flow = DNL(dynamic O-D demand, network supply)
Dynamic Network Loading

**MAC-POSTS:**

- Mobility Data Analytics Center - Prediction, optimization, and simulation toolkit for transportation systems
- [https://github.com/Lemma1/MAC-POSTS](https://github.com/Lemma1/MAC-POSTS)
One lane closed
Speed limit drops from 60 to 40 miles/h

Philadelphia, DVRPC area
Dynamic network loading

PLAY

2:00 PM – 3:00 PM
Real-time traffic management

Objective:
• real-time control of dynamic message sign on I-95 corridor to reduce network congestion

Data Input:
• Estimated dynamic O-D demand, calibrated behavior parameters
• Real-time traffic speed feeds
• Real-time accident report

Prediction Method:
• Dynamic network loading
Real-time traffic management

Dynamic message signs along the I-95 corridor in Philadelphia
1. **Acquire Speed**: acquire the real-time traffic speed in [8:00, 8:15]
2. **Estimate Previous**: estimate the route choice probability in [8:00, 8:15]
3. **Compliance ratio update**: update the DMS compliance ratio calculated in [7:45, 8:00]
4. **Optimize Next interval**: optimize the route choice probability in [8:15, 8:30]
5. **Generate DMS Message**: based on the optimized and estimated route choice, generate DMS message for [8:15, 8:30]
6. **Predict Future**: Predict traffic state in next one hour using updated compliance ratio
7. Move current time to 8:30, go back to step 1
Overview of real-time traffic management framework

Data input
- INRIX API
  Acquire real time speed data
- Historical OD
  Estimated by off-line DTA
- Capacity Drop report
  User can report capacity drop on browser

Server
- Path generator
  Generates path for real-time vehicle routing
- State Estimator
  Estimates travelers route choice in current interval
- State Optimizer
  Optimizes travelers route choice in next interval
- DMS message generator
  Generate the DMS messages according to the estimated and optimized route choice
- Feedback learner
  Update DMS compliance rate by feedback learning

Browser
- Animation
  Traffic condition prediction of next one hour
- DMS message
  Updated DMS message online and an API is also provided

Carnegie Mellon University
Civil and Environmental Engineering
Thanks!

Contact Info:
Carnegie Mellon University
Mobility Data Analytics Center
Wei Ma: weima@cmu.edu