Statistical inference of probabilistic O-D demand using day-to-day traffic data

Sean Qian
Assistant Professor, CEE & Heinz, CMU
Northwestern, May. 31, 2018
Outline

1. Massive data: opportunities and challenges

2. Statistical Origin-Destination Demand Estimation

3. Mobility Data Analytics Center (big MAC)
Smart decision making?

- Incident management
- Infrastructure retrofit
- Ride-sourcing impact/regulation
- Parking pricing
- ...

Massive data OD Estimation Mobility Data Analytics Center
Massive data: useful but challenging

Unexplored space

Feature \( i \)

Feature \( j \)

Recurrent traffic

Non-recurrent traffic

Desired recurrent traffic
A possible solution: data + physics
A generic infrastructure network

- An infrastructure node
- **Users**: human beings
- **Goods**: water, energy, vehicle...
- **△** A sensor
A generic infrastructure network

1. Sensing in sampled locations/time
2. Infer features of users, goods and infrastructure
3. Predict spatio-temporal distributions and system performance
4. Make decisions: manage supply and demand

Final goals: evaluation and intervention
Sensing-Learning-Managing

1. Sensing
   - Supply: network features, planned and unplanned incidents, weather, etc.
   - Passengers and vehicles: roadway, parking, transit, bikes, pedestrian, etc.

2. Learning
   - Behavior: choices of time, routes, modes and parking
   - Data mining: best estimation and prediction

3. Decision making
   - Short-term control
   - Long-term planning
Sustainable mobility

- Minimal congestion
- Resilient
- Safe
- Environmentally friendly

Sensing
- Supply
- Demand
- Integration

Learning
- Estimation
- Prediction
- Behavior

Managing
- Control
- Planning
**Concepts...**

- \( x \): link flow (flux, density, speed...)
- \( f \): path flow (flux, density, speed...)
- \( c \): system states (cost, time, emissions...)

Given \( x^o, f^o, c^o \) and supply, learn \((x, f, c) = G(\text{supply, demand})\)
ODE: Behavioral model $G$

- Use OD demand $q$ to approximate demand
- Define user behavior $G$

$$G : (\text{supply}; q) \mapsto (x, f, c)$$

- Given $x^o, f^o, c^o$ and supply, estimate $q$
- Calibrate $G$, estimate/predict $(x, f, c)$
Basic Notations

Supply:

- Transportation network $N$
- $A$ links, finite flow capacity $C_a$ of link $a$
- $K$ routes, a route $k$ contains different set of links

Demand:

- O-D: origin destination demand $q_{rs}$, indicating the number of travelers from $r$ to $s$.
- Associate an OD with multiple routes, flow rate $f_{rs}^k$
- Behavior: route choice

Observation:

- Link flow counts ($x^o$)
- Link travel time ($c^o$)
Traffic Assignment

Traditional Model:

- $TA : (N; q) \mapsto (x, f, c)$

Challenge:

- Data Variation
  - Variance-covariance of observed data
  - Variance-covariance of $(x, f, c)$
Statistical Traffic Assignment

- Make the best use of data: mean and variance
- $(x, f, q) \rightarrow (X, F, Q)$
- Statistical equilibrium: a new behavioral model
Generalized Statistical Traffic Assignment (GESTA)

First, we work on $G$

$$G : (N; Q) \leftrightarrow (X, F, C)$$
Generalized Statistical Traffic Assignment (GESTA)

Probabilistic traffic demand: $Q \sim \mathcal{N}(q, \Sigma_q)$
GESTA - cont.

Stochastic routing: \( F \sim \mathcal{MN}(\tilde{p} Q, \Sigma_f) \)
**GESTA - cont.**

**Sensing:** \( X_m = X + \epsilon_e \)
GESTA - cont.

**System states:** $C = C(X, F)$

![Visual representation of system states](image)
Perception: $p = f(C)$
GESTA - cont.

**Perceiving:** $p = f(C)$

**Routing:** $F \sim \mathcal{MN}(\bar{p} Q, \Sigma_f)$

**Loading:** $C = C(X, F)$
GESTA - cont.

Level 1: \[ X_m = X + \epsilon_e \] (Unknown Error)
\[ \epsilon_e \sim \mathcal{N}(0, \Sigma_e) \]

Level 2: \[ X = \Delta F \]
\[ F \sim \mathcal{MN}(\tilde{p} Q, \Sigma_f) \] (Route choice variation)

Level 3: \[ Q \sim \mathcal{N}(q, \Sigma_q) \] (Demand variation)
GESTA - cont.

**GESTA features:**

- Daily traffic condition is not in equilibrium
- Statistical equilibrium is built in a probability space
- Link/path flow variance = demand variance + choice variance

Now we know $G : (N; Q) \mapsto (X, F, C)$

How can we learn $X, F, C, Q$ from $X^o, F^o, C^o$
### Deterministic O-D estimation problem

\[
\min_q L(x^o, Aq) \tag{1}
\]

where \( A \) is the assignment matrix, \( q \) is O-D demand, and \( x^o \) is the observed link flows.

**Estimation Methods:**

- Entropy maximizing models
- Generalized least square
- Maximum likelihood estimator
New challenges

Figure: The Washington D.C. Downtown network
New challenges - cont’d
Challenges

- Data Variation
  - Multi-day data
  - Variance-covariance of $Q, X, F, C$
- Scalability
- Observability
Data Variation

Idea:

- Estimate the probabilistic O-D demand

Probabilistic O-D estimation problem

$$\min_{\mathbf{q}, \Sigma_{\mathbf{q}}} R(\mathbf{Q}^o, \mathbf{A} \mathbf{Q}) \quad (2)$$

where $R$ is the risk function, $\mathbf{A}$ is the assignment matrix, $\mathbf{Q}$ is the random vector of O-D demand, and $\mathbf{X}^o$ is the random vector of observed link flows.
Scalability and Observability

Since the model gets more complicated:

**Scalability:**

- Is it still possible to scale to large networks?
Scalability and Observability

Since the model gets more complicated:

**Scalability:**

- Is it still possible to scale to large networks?
- Solution: approximate high-dimensional probability distribution using data, instead of Bayesian inference.
Scalability and Observability

*Since the model gets more complicated:*

**Scalability:**
- Is it still possible to scale to large networks?
- Solution: approximate high-dimensional probability distribution using data, instead of Bayesian inference.

**Observability:**
- Can we still estimate OD using a small fraction of observations?
Scalability and Observability

Since the model gets more complicated:

**Scalability:**
- Is it still possible to scale to large networks?
- Solution: approximate high-dimensional probability distribution using data, instead of Bayesian inference.

**Observability:**
- Can we still estimate OD using a small fraction of observations?
- Solution: sparsity analysis when highly underestimated
Objective

How to estimate:

- OD mean and cov: $q, \Sigma_q$
- flow mean and cov: $c, \Sigma_c, x, \Sigma_x, f, \Sigma_f$
- Route choice probability $p$

Such that GESTA:

- Best fits data collected over many years
- Scales easily
- Has fairly good observability
IGLS Framework

- Iterative Generalized Least Square: EM like algorithm
- Separates the probabilistic OD estimation problem into two sub-problems:
  - Estimate OD mean vector
  - Estimate OD variance/covariance matrix
- Newton-Raphson step
Estimate OD mean

Traditional? with new statistical insights

$$\min_{f} \quad n (\Delta^0 f - \hat{x}^o)^T \left( \hat{\Sigma}_x^o \right)^{-1} (\Delta^0 f - \hat{x}^o)$$

$$\text{s.t.} \quad f \in \Phi^+$$

Where $\Phi^+$ is the feasible set of $f$, such as Probit-based GESTA.

**Methods:**
- Single level method, Shen & Wynter (2012)
Sparse penalization:

\[
\min_{\Sigma_q} \| S_x^o - \Sigma_x^o \|_F^2 + \lambda \| \Sigma_q \|_1 \\
\text{s.t.} \quad \Sigma_x^o = \Delta^o \Sigma_f |_q (\Delta^o)^T + \Delta^o \tilde{p} \Sigma_q \tilde{p}^T (\Delta^o)^T \\
\Sigma_q \succeq 0
\]  \hfill (4)

Methods:

- Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Nesterov 2005)
Observervility

The statistical risk of the OD mean estimator under IGLS - the statistical risk of the deterministic OD is bounded, and declines w.r.t. sample size

A small example

- OD: 1 → 3, 2 → 3
- Observation: link 1 and link 3
- 500 days
A small example - cont.

**Figure:** Synthesized “true” link flow data for different correlation $\rho$
### Table: Results of probabilistic ODE on the three-link toy network (no historic O-D demand information is used)

<table>
<thead>
<tr>
<th>True $\rho$</th>
<th>Settings</th>
<th>$\hat{q}_{1\rightarrow3}$</th>
<th>$\hat{q}_{2\rightarrow3}$</th>
<th>$\hat{\sigma}^2_{1\rightarrow3}$</th>
<th>$\hat{\sigma}^2_{2\rightarrow3}$</th>
<th>$\hat{\rho}$</th>
<th>RMPSE</th>
<th>KL-distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True value</td>
<td>700</td>
<td>500</td>
<td>175</td>
<td>125</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>722.17</td>
<td>500.41</td>
<td>186.69</td>
<td>134.21</td>
<td>0.56</td>
<td>3.62%</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td>Logit - w/o Lasso</td>
<td>682.36</td>
<td>499.63</td>
<td>207.94</td>
<td>134.21</td>
<td>0.50</td>
<td>2.08%</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>Probit - w/o Lasso</td>
<td>699.50</td>
<td>499.63</td>
<td>200.94</td>
<td>134.21</td>
<td>0.52</td>
<td>0.07%</td>
<td>0.01</td>
</tr>
<tr>
<td>0</td>
<td>w/o EC - w/o Lasso</td>
<td>715.91</td>
<td>500.46</td>
<td>143.05</td>
<td>138.74</td>
<td>0.03</td>
<td>1.87%</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Logit - w/o Lasso</td>
<td>681.28</td>
<td>500.46</td>
<td>162.49</td>
<td>138.75</td>
<td>0.02</td>
<td>2.21%</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>Probit - w/o Lasso</td>
<td>700.30</td>
<td>500.46</td>
<td>152.15</td>
<td>138.75</td>
<td>0.03</td>
<td>0.06%</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Logit - w/ Lasso</td>
<td>681.28</td>
<td>500.46</td>
<td>144.52</td>
<td>128.75</td>
<td>0.00</td>
<td>2.21%</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>Probit - w/ Lasso</td>
<td>700.02</td>
<td>500.46</td>
<td>132.27</td>
<td>128.75</td>
<td>0.00</td>
<td>0.05%</td>
<td>0.004</td>
</tr>
<tr>
<td>−0.5</td>
<td>w/o EC - w/o Lasso</td>
<td>703.41</td>
<td>499.06</td>
<td>173.34</td>
<td>132.60</td>
<td>−0.41</td>
<td>0.43%</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Logit - w/o Lasso</td>
<td>681.05</td>
<td>499.06</td>
<td>184.13</td>
<td>132.60</td>
<td>−0.39</td>
<td>2.23%</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>Probit - w/o Lasso</td>
<td>701.71</td>
<td>499.06</td>
<td>174.19</td>
<td>132.60</td>
<td>−0.41</td>
<td>0.23%</td>
<td>0.02</td>
</tr>
</tbody>
</table>
A small example - cont.

![Diagram showing link flows and variance]

- Link 1
- Link 2
- Link 3

**Link number**

<table>
<thead>
<tr>
<th>Link number</th>
<th>Link flow variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55%</td>
</tr>
<tr>
<td>2</td>
<td>95%</td>
</tr>
<tr>
<td>3</td>
<td>61%</td>
</tr>
</tbody>
</table>

**O-D demand variation**

- 45%
- 39%

**Route choice variation**

- 55%
- 95%
SR-41 Corridor

- 2,413 links and 7,110 O-D pairs
- 10% of O-D pairs (randomly chosen) are mutually correlated with a correlation randomly drawn from 0 to 0.5
- Randomly choose 50% of the links on the network to be observed for 1,000 days
The entire process of 900 iterations takes 486 minutes, but the estimate is reasonably good within approximately 300 minutes.
Figure: Estimated and “true” OD demand (Left: mean; Right: covariance)
SR-41 - cont.

**Figure:** Estimated and “observed” link flow (Left: mean; Right: variance of the marginal distributions)
Washington D.C. Downtown network

- 984 road junctions
- 2,585 road segments
- 4,900 O-D pairs
Figure: Estimated and observed link flow during the morning peak (Left: mean; Right: variance/covariance)
What’s next

Unsupervised learning:
- Weekdays/weekends
- Seasonal behavior

Hypothesis test and variance analysis
- Recurrent-nonrecurrent pattern detection
- Real-time subgraph anomaly detection

Extensions:
- Prior for variance/covariance matrix
- Other data sets, e.g., speeds
- Dynamic OD demand
- Multi-modal
MAC data sets in Pittsburgh

GIS, demographics, economics, weather

Traffic counts
- Highways, major arterials

Traffic time/speed
- INRIX, HERE, Uber Movement, AVI, BT

Transit
- APC-AVL, Park-n-ride, incidents

Parking
- Transactions of on-street meters and occupancy of garage

Incidents
- RCRS/PD/911/311/PTC/PennDOT Crash/Road closures/Events

Social media (e.g., Twitter)
Mobility Data Analytics Center (big MAC)

Using data analytics, quantitative techniques, and domain knowledge to address real-world problems

- Twitter-based incident detection
- Off-line dynamic network analysis
- Real-time traffic operation
- Parking
- Public transit
Twitter-based incident detection
Twitter-based incident detection
Off-line dynamic network analysis
Off-line dynamic network analysis

![Map with data updates interface]
Off-line dynamic network analysis
Off-line dynamic network analysis
Real-time traffic operation: traffic prediction
Real-time traffic operation: traffic prediction
Real-time traffic operation: demand management
Real-time traffic operation: demand management
Public transport
Public transport
Parking
Team

CURRENT MEMBERS

Prof. Sean Qian
   Director

Xidong Pi
   PhD Candidate

Wei Ma
   PhD Candidate

Pinchao Zhang
   PhD Candidate

Shuguan Yang
   PhD Candidate

Matthew Battifarano
   PhD Student

Weiran Yao
   PhD Student

Rick Grahn
   PhD Student

ALUMNI

Dr. Yiming Gu
   Senior Scientist
   United Technologies Research Center

Zhangning Hu
   Software Engineer
   Google

Zach Sussman
   Software Engineer
   Apple

Juncheng Zhan
   Undergraduate
   SCS CMU
Thanks! Questions and comments?

Sean Qian, seanqian@cmu.edu
MAC: mac.heinz.cmu.edu