Exploring Google’s Passive Data for Origin-Destination Demand Estimation

2018 SF Bay Area ITE/ITS CA Joint Transportation Workshop

Bhargava Sana
Background: San Francisco Freeway Corridor Management Study (FCMS)

- Potential strategies
  - TDM, managed lanes
  - Ramp metering
- OD data could help impute
  - Socio-demographics (age, income, auto ownership etc.)
  - Willingness To Pay (WTP)
Background: OD Data Collection

- Traditional OD data collection
  - License plate surveys
  - Roadside interview surveys
- Emerging passive OD data sources
  - Bluetooth
  - Cellphone Call Detail Records (CDR)
  - GPS-based
Aggregated and Anonymized Trip (AAT) Data

- Google’s Better Cities Program
  - Minimize congestion, improve safety and reduce infrastructure spending

- Users opt-in to share location
  - GPS, cell-phone towers, Wi-Fi detectors

- Aggregated and Anonymized Trip (AAT) information from location reports
  - Extract data from moving users
  - Clean data and snap to road network
  - Aggregate OD trip counts
  - Apply differential privacy filters and minimum trips threshold
San Francisco AAT Dataset

- 85-district system: combination of Tract and County boundaries
- Hourly AAT data for six months (Apr-Jun and Sep-Nov 2015)
AAT Dataset for FCMS

“via polygons” for 4 segments around US-101 and I-280 interchange in San Francisco
- Flow data provided as relative trips as opposed to absolute counts
- Trips needed for planning purposes
- Convert relative flows to trips to obtain
  1. Facility-specific OD matrices
  2. Regional dynamic OD matrices
Facility-specific OD Matrices

- **Observed dataset: Caltrans’ Performance Measurement System (PeMS)**
- **PeMS counts available for same time and location as AAT**
- **PeMS vehicle counts -> person trips**
  - Use observed occupancy on corridor **(1.16 – 1.24)**

![Graph (e)](image-url) Corridor AAT Flow on Typical Weekday (Tu,W,Th)

![Graph (f)](image-url) Corridor PeMS Flow on Typical Weekday (Tu,W,Th)
Relative Flow Conversion Model

- AAT relative flows vary by
  - Hour of day: added indicator variables
  - Roadway segment location: separate models estimated
  - Several model forms and transformations explored

\[ PT_t = K + \beta_f \log(RF_t) + \beta_t + \epsilon; \quad 0 \leq t \leq 23 \]

- \( PT_t \): Person trips on roadway facility for hour-of-day \( t \)
- \( K \): Regression constant/Intercept
- \( \beta_f \): Coefficient of AAT relative flow
- \( RF_t \): AAT relative flow for hour-of-day \( t \)
- \( \beta_t \): Constant for hour-of-day \( t \)
- \( \epsilon \): Error term
Model Validation

Google AAT Data: US-101 Segments

PeMS Counts and Model Predictions: US-101 Segments

Google AAT Data: I-280 Segments

PeMS Counts and Model Predictions: I-280 Segments
Comparison of Facility-Specific OD Matrices with SF-CHAMP Select Link

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Segment</th>
<th>Correlation Coefficient</th>
<th>County level</th>
<th>District level</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM Peak</td>
<td>US-101 north of interchange</td>
<td>0.948</td>
<td></td>
<td>0.440</td>
</tr>
<tr>
<td></td>
<td>US-101 south of interchange</td>
<td>0.751</td>
<td></td>
<td>0.390</td>
</tr>
<tr>
<td></td>
<td>I-280 north of interchange</td>
<td>0.947</td>
<td></td>
<td>0.279</td>
</tr>
<tr>
<td></td>
<td>I-280 south of interchange</td>
<td>0.881</td>
<td></td>
<td>0.342</td>
</tr>
<tr>
<td>PM Peak</td>
<td>US-101 north of interchange</td>
<td>0.964</td>
<td></td>
<td>0.422</td>
</tr>
<tr>
<td></td>
<td>US-101 south of interchange</td>
<td>0.685</td>
<td></td>
<td>0.334</td>
</tr>
<tr>
<td></td>
<td>I-280 north of interchange</td>
<td>0.940</td>
<td></td>
<td>0.253</td>
</tr>
<tr>
<td></td>
<td>I-280 south of interchange</td>
<td>0.846</td>
<td></td>
<td>0.312</td>
</tr>
</tbody>
</table>

Pearson’s correlation = 0.811

Pearson’s correlation = 0.809
Regional OD Matrices

- Observed data: California Household Travel Survey (CHTS)
- CHTS 2012 Bay Area Sample
  - ~8,000 households
  - ~70,000 trips
- Expanded survey data compared to AAT’s “average weekday” weights

<table>
<thead>
<tr>
<th>Data</th>
<th>Count</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google AAT hourly OD</td>
<td>145,758</td>
<td>0.0002</td>
<td>0.9257</td>
<td>0.0007</td>
<td>0.0118</td>
</tr>
<tr>
<td>CHTS hourly OD</td>
<td>9,969</td>
<td>52</td>
<td>404,951</td>
<td>2,311</td>
<td>18,256</td>
</tr>
<tr>
<td>Google AAT daily OD</td>
<td>7,225</td>
<td>0.0028</td>
<td>13.3745</td>
<td>0.0137</td>
<td>0.2292</td>
</tr>
<tr>
<td>CHTS daily OD</td>
<td>2,996</td>
<td>61</td>
<td>4,676,065</td>
<td>7,692</td>
<td>129,297</td>
</tr>
</tbody>
</table>

Observed data: California Household Travel Survey (CHTS)

- CHTS 2012 Bay Area Sample
  - ~8,000 households
  - ~70,000 trips
- Expanded survey data compared to AAT’s “average weekday” weights
Several supervised machine learning techniques were explored

Learning and application done using “scikit-learn” Python library

<table>
<thead>
<tr>
<th>Model</th>
<th>Optimal Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-nearest Neighbors Regression</td>
<td>5 nearest neighbors distance-weighted mean</td>
</tr>
<tr>
<td>Random Forest Regression</td>
<td>10 trees</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>Regularization parameter = 0.1</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>Regularization parameter = 0.001</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>RBF kernel function (gamma = 0.1) penalty parameter = 1e7</td>
</tr>
<tr>
<td>Neural Networks (Multi-layer Perceptron)</td>
<td>2 hidden layers with 10 neurons each</td>
</tr>
</tbody>
</table>
## ML Results: Performance Metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>R-Sqrd test</th>
<th>R-Sqrd train</th>
<th>RMSE test</th>
<th>RMSE train</th>
<th>WAPE test</th>
<th>WAPE train</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-nearest Neighbors Regression</td>
<td>0.7897</td>
<td>1</td>
<td>7,602.81</td>
<td>0</td>
<td>62.31%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Random Forest Regression</td>
<td>0.8071</td>
<td>0.9608</td>
<td>7,280.02</td>
<td>3,763.11</td>
<td>61.88%</td>
<td>24.52%</td>
</tr>
<tr>
<td>OLS Regression</td>
<td>0.9541</td>
<td>0.9778</td>
<td>3,549.93</td>
<td>2,831.72</td>
<td>41.22%</td>
<td>35.97%</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>0.9539</td>
<td>0.9778</td>
<td>3,559.19</td>
<td>2,834.67</td>
<td>41.67%</td>
<td>36.19%</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>0.9541</td>
<td>0.9778</td>
<td>3,550.75</td>
<td>2,831.81</td>
<td>41.28%</td>
<td>36.01%</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>0.9438</td>
<td>0.9742</td>
<td>3,929.32</td>
<td>3,055.74</td>
<td>40.52%</td>
<td>33.13%</td>
</tr>
<tr>
<td>Neural Networks (Multi-layer Perceptron)</td>
<td>0.9594</td>
<td>0.9786</td>
<td>3,339.80</td>
<td>2,780.82</td>
<td>39.85%</td>
<td>36.23%</td>
</tr>
</tbody>
</table>

*R-sqrd = Coefficient of determination, 1 - (sum of squared residuals/total sum of squares)*

*RMSE = Root mean squared error*

*WAPE = Weighted absolute percent error*
ML Results: Hourly Demand Comparison

Regional Demand

Intra San Francisco

East Bay to San Francisco

San Francisco to East Bay

Trips (000s)

Hour of Day

Trips (000s)

Hour of Day

CHTS
K-nearest Neighbors Regression
Support Vector Regression
Neural Networks (Multi-layer Perceptron)
OLS Regression
Lasso Regression
Ridge Regression
Random Forest Regression
ML Application: Scaled AAT inputs

Regional Demand: Growth Factor = 1

Regional Demand: Growth Factor = 2

Regional Demand: Growth Factor = 4

Regional Demand: Growth Factor = 5

Graphs showing the relationship between CHTS OD Size (in thousands) and Total predicted trips (in thousands) for different growth factors. The graphs compare various regression models including CHTS, K-nearest Neighbors Regression, Support Vector Regression, Neural Networks (Multi-layer Perceptron), OLS Regression, Lasso Regression, Ridge Regression, and Random Forest Regression.
ML Application: Scaled AAT inputs

Intra San Francisco County Demand (Total Trips: 2,354,447)

Mission Bay/Potrero to SF Core Demand (Total Trips: 10,316)
Conclusions

- Reasonable temporal patterns observed in AAT data at aggregate levels
- Conversion of relative flows to trips possible with statistical models
  - Minimum trips threshold to protect privacy is a limitation
- Model sensitivities hard to interpret due to non-linear nature
- Comparison with other data sources might provide insights into penetration and accuracy
- Further segmentations based on mode and purpose might be useful
Thank You