“Big Data” in Transportation Planning: User Perspectives, Challenges and Successes

NYSAMPO 2015 Conference
June 23, 2015
Introduction

A brief review of the state of the practice in transportation planning and travel modeling with respect to using “big data” sources of travel data. Are there practical uses of these datasets to improve travel models, reduce the expense of developing and updating them, and to conduct analyses that might not have been possible prior to the availability of these data?

- Types of big data
- User perspectives
  - What type of analysis needs are big data supporting?
  - What challenges do users face when trying to use “big” data?
  - What is known about data quality?
- Case studies – successful and less successful applications
- Lessons learned and comments
Big Data: Source Types

Earlier in the session, Nikola Ivanov from UMD CATT Lab introduced big data concepts. Here we look at data sources and types in use in the practice and discuss applications.

Vehicle Based GPS Data
- Examples: ATRI, INRIX Insights Trips and Volume, NAVTEQ, TomTom, Strava (bicycles)

Cell phone based GPS/cell location data
- Examples: AirSage, Smartphone apps for household travel surveys, e.g. rMOVE

Roadside/in road sensor data
- Bluetooth survey data, PeMS (performance monitoring system) data
Big Data: Use Types

From a use perspective we are interested in what the data describe about travel behavior and the transportation system.

System Performance: Volumes, Speeds
Examples: PeMS, INRIX Insights Volume, NAVTEQ, TomTom, Strava (bicycles)

Origin to Destination Data: People
Examples: AirSage, Smartphone apps for household travel surveys, e.g. rMOVE

Origin to Destination Data: Vehicles
Examples: Bluetooth survey data, ATRI, INRIX Insights Trips, Strava (bicycles)
What Types of Analysis Needs are Big Data Supporting?

Focus is on better understanding current travel behavior and system performance: whenever we need a more complete and up to date observations of what is happening NOW (or to some extent, in the recent past), these data have a role

- Current origin – destination travel patterns
  - Base year travel model development (direct trip table development)
  - Model calibration/validation (ODs, trip length frequencies)
  - All modes (person travel) or some modes (e.g., trucks)
- Current roadway volumes and speeds
  - NPMRDS data (earlier presentation) is a great example; these data were extremely difficult and expensive to collect previously
  - Support model calibration, corridor analysis, congestion management
- Other modal planning studies
  - Health and recreation studies and non-motorized planning, e.g., bicycle usage using Strava or Cycletracks
Big Data Issues and Limitations

The data hold a lot of promise but as with all new information sources there are issues and limitations that may not be immediately clear to the user

• Big data don’t tell us anything about the “why” which is important if we are going to understand change – in the future and/or when a system change is made

• Vendors have commercial sensitivities around their analytical process
  - How they summarize and analyze data, and its resulting margins of error is often not available
  - Results in user not having a clear understanding of the analytical process used to create the data or its reliability

• Data users often have significant problems making use of the data due to technology barriers
  - Large and/or difficult data sizes beyond in house agency capabilities (e.g. NPMRDS data)
  - Fast changing industry results in low data continuity
Case Study 1 – Idaho Statewide Model
AirSage data to make OD trip table
Statewide Travel Demand Model

- Statewide Travel Demand Model (STDM) project started in 2013
- As part of ITD’s data-driven, performance-based needs analysis approach
- Key model requirement is to forecast auto and truck traffic and network LOS
Cell Phone OD Data

- AirSage converts cell phone time and location data into trip OD data
- Has exclusive agreement with Verizon and others to aggregate and sell the cell phone location data
- Extracts the time and location of the cell phone every time it talks to the network - email, texts, phone calls, location, etc.
- Identifies cell device usual home and work location based on the cluster of points identifying where the phone “sleeps” at night and “works” during the day
Cell Phone OD Data

- Trips are coded with respect to the home and work anchor locations:
  - Home-based work
  - Home-based other
  - Non-home-based
  - Resident versus visitor

- AirSage expanded the sampled trips to better match the population using various Census data sets

Idaho Cell Coverage
Cell Phone OD Data Purchase

- **Calendar**: Average weekday for the month of September 2013
- **Markets**: Resident HBW, HBO, NHB; Visitor NHB
- **Time period**: Daily
- **Zones**: 750 x 750 super zones to reduce cost
- **Price**: Reasonable
- **License**: Data licensed only for the project; derivative products can be used for other purposes though
Disaggregation to Model Zones and Initial Network Assignment

• Disaggregated the 750 zones to 4000+ zones using each model zone’s share of super zone (pop + 2 * emp)

• Results in daily raw cell phone flows between model zones for four markets

• Assigned cell phone “trips” through the network using free flow travel time

• Compared available trip length data to check results
Cell Phone HBW and Census JTW Trip Lengths

- Similar results within each District as well
Cell Phone OD and COMPASS MPO Survey Trip Lengths

HBW

Avg. Trip Length (COMPASS) - 8.65 Miles
Avg. Trip Length (AirSage) - 8.84 Miles
Coincidence Ratio - 0.72

HBO

Avg. Trip Length (COMPASS) - 4.94 Miles
Avg. Trip Length (AirSage) - 5.33 Miles
Coincidence Ratio - 0.91

NHB

Avg. Trip Length (COMPASS) - 4.19 Miles
Avg. Trip Length (AirSage) - 6.24 Miles
Coincidence Ratio - 0.68

Coincidence Ratio and Average Trip Length Difference by Trip Category

<table>
<thead>
<tr>
<th>Trip Category</th>
<th>Coincidence Ratio</th>
<th>Average Trip Length (Miles)</th>
<th>Avg. Trip Length Difference</th>
<th>% Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COMPASS</td>
<td>AirSage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBW</td>
<td>0.72</td>
<td>8.65</td>
<td>8.84</td>
<td>0.19</td>
</tr>
<tr>
<td>HBO</td>
<td>0.91</td>
<td>4.94</td>
<td>5.33</td>
<td>0.39</td>
</tr>
<tr>
<td>NHB</td>
<td>0.68</td>
<td>4.19</td>
<td>6.24</td>
<td>2.05</td>
</tr>
</tbody>
</table>
Observations from Initial Assignment

• Reasonable goodness-of-fit between the cell phone TLFs and the Census JTW and Boise MPO travel survey data sets

• Significant differences for NHB trips, especially short distance trips

• **Why?** Most likely classification issues:
  – Survey NHB trips are just household-based non-home-based trips, where as the AirSage trips are everything else, including commercial vehicles
  – Very short trips in terms of distance and time may drop out of the AirSage data set as well
  – Difficult to identify non-home or work-based trips
  – Simplified procedure to disaggregate super zone flows to model zones likely creating differences for some OD pairs

• Remember we’re comparing processed cell phone movements to person reported travel
Discussion

• ITD wanted a data driven model in order to get components of the system up and running as early as possible and to estimate current year auto and truck volumes
• The cell phone OD data is a reasonable starting point for generating statewide trip matrices
• Used in conjunction with existing travel modeling tools and techniques, the cell phone OD data has a promising future in our industry
• Need to better understand how synthesized cell phone flows are different than traditional travel modeling data sets
Case Study 2 – SMTC Travel Model
AirSage for Model Validation
Case Study 2: SMTC

• Use AirSage data to validate SMTC model origin-destination patterns and trip lengths
• Use AirSage data to understand origin-destination movements that use I-81 viaduct
SMTC Travel Model Coverage Area

- 1,185 I & E Model Zones
- 788 Zones for Cellular Data
- 45 Miles North-South
- 35 Miles East-West
- Four-step Travel Model built in TransCAD
Comparing Trip Purpose Shares

- AirSage appears capable of identifying Home-Based Other (HBO) trip purposes
- AirSage appears less capable of differentiating between HBW and NHB trip purposes
Dividing the Region into 26 Medium Districts Municipalities
Town-to-Town Trip Flows
\[ 26 \times 26 \] = 676 OD pairs plotted

Model vs. Airsage

Correlation = 0.945

AirSage Trips (Log)

Model Trips (Log)
Town-to-Town HBW Trip Flows
[26 x 26] = 676 OD pairs plotted

Model vs. Airsage

Correlation = 0.947

HBW
Town-to-Town HBO Trip Flows
[26 x 26] = 676 OD pairs plotted

Model vs. Airsage

Correlation = 0.931

AirSage Trips (Log)

Model Trips (Log)
Town-to-Town NHB Trip Flows
[ 26 x 26 ] = 676 OD pairs plotted

Model vs. Airsage

Correlation = 0.919
Trip Length Frequency Distribution
Total Trips (2-mile bins)

AirSage and Model trip length distributions for ALL trips are consistent and compare favorably.
Trip Length Frequency Distribution
HBW w LEHD (2-mile bins)

- Census Worker-Flows (LEHD) trip length frequency distribution lies in between the Model and AirSage distributions.

Model = 316k total trips
LEHD = 374k total trips
AirSage = 484k total trips
Trip Length Frequency Distribution
NHB Trips (2-mile bins)

AirSage has fewer short non-home based trips which are less than 6-miles in length
The AirSage HBW and NHB trip length distributions don’t really match our “mental model”

Given the open-question regarding the % split of HBW and NHB trips - we have combined the two trip length distributions

The resulting TLFD is much more consistent and compares favorably
Key Takeaways: Travel Model vs. AirSage

1. Aggregation is good; disaggregation is bad
   - Agencies should think about the level of lowest level of resolution that they can be happy with and develop zones accordingly. Will be economical too.

2. Get creative with external zone boundaries
   - Don’t make them too small so you don’t miss out on external trips.
   - Don’t make them too large so that you include distant mid to large cities that you don’t care about.

3. Think about what trip purposes you really need and why
   - Home and Work based trip purposes should be pretty good unless there are significant number of students or shift workers in your region

4. Select link analysis should only be done on long links and with care
Lessons Learned

- Don’t be afraid to try out the data – the pricing is generally competitive and certainly less than traditional methods (OD surveys, travel time surveys)
- Negotiate on prices – generally the vendors can be haggled with! Talk to your colleagues at other agencies and find out what they were able to get
- Don’t attempt to push the data too far – aggregate to get more reliable data points
- Remember that the data become modeled data as soon as the processing includes rules and inferences (e.g., trip purpose) and it has generally not been proven that any of the algorithms for this are highly reliable
- Ask the vendor to process and summarize the data if you are not able to do it in-house, or look to resources such as University Transportation Centers for support
- Share your experiences with your colleagues!