Big Data Applications for Travel Demand Modeling

Tennessee Model Users Group

Ted Reinhold, Senior Solutions Engineer
Julie Hirni, Director, Public Sales Sector - Southeast

November 17, 2021
What is “Big Data?”
Big Data for transportation

1. **Mobile Device LBS**
   - Anonymized “opt-in” data from mobile device apps

2. **Vehicle GPS**
   - GPS devices installed in personal and commercial vehicles

3. **Internet of Things**
   - Connected vehicles, smart fare cards, RFID, geo-fencing, and more

**Traditional Data Sources**
- Household & Intercept Surveys
- Aerial Photos & Videos
- Bluetooth & Other Sensors
- Assumption-Based Modeled Data
Location-Based Services (LBS) and GPS data

- Every month, we process over 100 billion anonymized location records from smart phones and GPS navigation devices in cars and trucks.

- Route Science® transforms them into contextualized, normalized and aggregated travel patterns.

MOBILE DEVICE DATA
from 110M+ devices of U.S. and Canadian adults

Example, San Bernardino, CA
Oct 8, 2017 24-hr snapshot

CONTEXT
Parcel Data
Digital Road Network Data
U.S. Census
Data Processing

Proprietary algorithms and machine learning turn the data into contextualized, aggregated, normalized travel patterns.
Validation

Analytics providers validate results using counters, sensors, and other sources.
Which source data is the best? A constantly evolving mix

Data is always changing…
- *Historically we used cell tower data*
- *Today we use a blend of data sources →*
- *Tomorrow we’ll add something new*

**Framework for assessing Big Data sources**

**Data sample:** size, representativeness, coverage (temporal/spatial), frequency of updates

**Privacy protections:** how does the process protect privacy of individuals?

**Validation and uses:** applicability to transportation-related use cases

10+ Terabytes of data processed monthly
Big Data Applications for Modeling
Benefits for Modeling

• Supplement traditional data sources
• Overcome modeling assumptions
• Continually refine and validate model
Model Development
Use Big Data O-D as a seed matrix to estimate full O-D trip table
Supplement gaps in traveler surveys (e.g., tourists)
Granular metrics for non-motorized and transit modes

Pedestrian activity near downtown

Bus trips to downtown core
Develop special modules for freight hubs, campus, airports, parks…
Internal-External demand (II / IE / EI / EE)
Develop sub-area models from regional travel demand model
O-D inputs to traffic simulation models, TIA, cut-through studies
Model Calibration & Validation
Total trip production and attraction

Vehicle trips destined for downtown, typical weekday AM Peak (6-10a)
Total trip production and attraction
Total trip production and attraction

Traffic By Household Income

Traffic By Trip Purpose
Travel time distribution by time of day
Travel time distribution by time of day

- **Early AM (12-6a)**
- **Peak AM (6-10a)**
- **Mid Day (10a-3p)**
- **Peak PM (3-7p)**
Calibrate network assignment – route choice
Calibrate network assignment – link delays/congestion
Calibrate network assignment – link volumes/speeds

**Volume Distribution**

- **Brainerd Road / 862575265 / 3**

**Speed Distribution**

- **Brainerd Road / 862575265 / 3**
Calibrate model sensitivity to network or land use changes

Paths during bridge closure

Typical paths
Summary

1. Big Data can complement current sources of input, calibration and validation data for models
2. Opportunities to refine overlooked model components - I/E, freight, tourism, seasonality
3. Monthly data refresh facilitates frequent model updates and monitoring
4. Big Data enables a wide range of model applications and use cases
STREETLIGHT DATA

ted.reinhold@streetlightdata.com
julie.hirni@streetlightdata.com
# Vetting Big Data Metrics Providers

A Framework for Questions to Ask and Best Practices to Look For

<table>
<thead>
<tr>
<th></th>
<th>Data Sets and Sources</th>
<th>Processing Methods</th>
<th>Privacy Protections</th>
<th>Validation and Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>• Where does the data come from?</td>
<td>• What algorithm and machine learning techniques are used?</td>
<td>• How is the data collected, processed, and shared?</td>
<td>• How are the metrics validated?</td>
</tr>
<tr>
<td></td>
<td>• How big is the sample size, and from how many providers?</td>
<td>• How granular are the metrics?</td>
<td>• How does the process protect privacy of individuals?</td>
<td>• How have the metrics been used in real-world applications?</td>
</tr>
<tr>
<td></td>
<td>• How frequently are the data sources evaluated and updated?</td>
<td>• What transportation modes are included?</td>
<td>• Where are privacy practices built into the process?</td>
<td>• How do customers access the metrics?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample representativeness
Sample representativeness

Penetration Rate by % White

Penetration Rate by % Black

Penetration Rate by % Asian

Penetration Rate by % Hispanic
Sample representativeness

<table>
<thead>
<tr>
<th>Income Bin</th>
<th>StreetLight Data</th>
<th>2017 NHTS Sample</th>
<th>StreetLight Data Bin compared to &lt; $20K Bin</th>
<th>NHTS Data Bin compared to &lt; $20K Bin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than $20,000</td>
<td>9.80%</td>
<td>0.07%</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$20-$35K</td>
<td>8.10%</td>
<td>0.08%</td>
<td>0.83</td>
<td>1.14</td>
</tr>
<tr>
<td>$35-$50K</td>
<td>9.60%</td>
<td>0.09%</td>
<td>0.98</td>
<td>1.29</td>
</tr>
<tr>
<td>$50-$75K</td>
<td>11.00%</td>
<td>0.09%</td>
<td>1.12</td>
<td>1.29</td>
</tr>
<tr>
<td>$75-$100K</td>
<td>11.40%</td>
<td>0.09%</td>
<td>1.16</td>
<td>1.29</td>
</tr>
<tr>
<td>$100-$125K</td>
<td>11.20%</td>
<td>0.09%</td>
<td>1.14</td>
<td>1.29</td>
</tr>
<tr>
<td>$125-$150K</td>
<td>10.90%</td>
<td>0.09%</td>
<td>1.11</td>
<td>1.29</td>
</tr>
<tr>
<td>$150-$200K</td>
<td>10.20%</td>
<td>0.10%</td>
<td>1.04</td>
<td>1.43</td>
</tr>
<tr>
<td>$200K+</td>
<td>10.40%</td>
<td>0.09%</td>
<td>1.06</td>
<td>1.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Density</th>
<th>StreetLight Data</th>
<th>2017 NHTS Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>10.2%</td>
<td>0.08%</td>
</tr>
<tr>
<td>Rural</td>
<td>12.2%</td>
<td>0.11%</td>
</tr>
</tbody>
</table>
Is LBS data “unrepresentative”? Not if you work with excellent suppliers.

- **Share of person trips that involve a purchase (NHTS 2017):**
  - % of trips that involve a purchase
  - % of trips that involve a trip

- **Penetration Rate by Household Income:**
  - Biased trip purpose - hard to normalize
  - Well distributed - straightforward to normalize
Was LBS data impacted by the iOS14 updates?
Not if you work with excellent suppliers.

Index (Feb 2019 = 1)

- Normal supplier increases in trips
- Our key partners implement “Opt In” in advance of any OS changes, with clear explanations and processes. We weed out those individuals who don’t want to be in the sample.
- COVID crisis – actual decrease in trips reflected in data
- COVID 3rd wave lockdowns
- Short term spike with one supplier (app based)
- Actual activity and our supply continue to increase post iOS14.5

iOS 14.5 launched April 26
What is YOUR sample size?

110M
Usable devices in our current sample

1.5B
trips analyzed in January 2020*

Your project sample size
110 Million Usable Devices in U.S. + Canada

What are these devices?

- A mix of LBS smart phones, connected trucks
- We work with suppliers to increase presence of **highly active devices** that are making consistent trips throughout the year.
- We continually **de-duplicate devices**, and **remove devices with infrequent pings**
- These create over **half a trillion “pings”** each year
1.5+ billion usable trips analyzed per month

How do we do it?

• We continually vet the quality of apps and devices to achieve maximum number of high-quality trips.

• We work with suppliers to increase presence of highly active devices that are making consistent trips throughout the year.

• We continually de-duplicate devices, and remove “inaccurate”, unusable trips.

1.5B trips analyzed in January 2020*

*Last typical, pre-COVID month
Transparent Sample Size Information for All Projects

- Sample size **varies** by time, data period, etc.
- For every StreetLight project you run, we’ll share the **sample of devices and trips** that contributed to your analysis.
- Where feasible, we’ll share **confidence intervals** too
Different approaches to sample size: Big Data vs. Traditional

Traditional (tube counter, etc.)

<table>
<thead>
<tr>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>Sa</th>
<th>Su</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All people for a few days

VS.

Big Data

Many people for all days
Different approaches to sample size: Big Data vs. Traditional

Location of permanent counter near Berlin, Ohio
2019 AADT = 10,028 (3.66M trips/year)
Different approaches to sample size: Big Data vs. Traditional

<table>
<thead>
<tr>
<th>Traditional (tube counter, etc.)</th>
<th>Bluetooth, surveys, etc.</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>All people for a few days</td>
<td>Some people for a few days</td>
<td>Many people for all hours for all days</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>Sa</th>
<th>Su</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How we make our sample representative

A SIMPLIFIED SUMMARY

Population Adjustments
Each device carries an individual adjustment factor based on how many StreetLight devices live on its home block (versus actual population on that block).

IoT Counter Normalization
We normalize across space/time by comparing the number of trips we sense to 10,000+ permanent loop counters embedded in the infrastructure across the US.

And More…
We check for major shifts in supplier data feeds, we use inputs about weather, road type, season and much more to finalize our normalization.

Repeat and update every month
How accurate are the metrics?
AADT accuracy by size of road (link to detailed validation paper)

- Avg percent error decreases on higher volume roads
- Errors in each band are better than AADT estimates from temporary counts, modeled counts

Mean Average Percent Errors by AADT Volume Cluster

- +/- 8.25%
- +/- 9.4%
- +/- 10.2%
- +/- 24.9%
- +/- 16.3%
- +/- 13.2%
How accurate are the metrics?
Bike and pedestrian Counts have extremely high correlation (link to detailed validation paper)

Correlation results comparing StreetLight to permanent counters for bicycle and pedestrian modes for weekdays show $R^2$ of 0.84 for bicycles and $R^2$ of 0.78 for pedestrians, which means the actual and estimate values are highlight correlated

These comparisons are based on permanent bicycle counters in San Francisco (SFMTA) and the Delaware Valley (DVRPC) and permanent pedestrian counters in Washington D.C. and the Delaware Valley (DVRPC) from 2018.
How accurate are the metrics?
Turning Movement ratios and demonstrate high accuracy correlation

Correlation between Hennepin turning movement ratios and StreetLight Volume ratios demonstrate $R^2$ value of 0.98, indicating a very high correlation.
Big Data for Modeling

- **Trip Based Models**
  - Trip rates can be customized beyond basing them on household size, household size and income level.
  - Gravity models and destination choice models can be refined. Different trip distribution models can be estimated since big data can be used to calibrate and validate them, without which validation data is lacking.
  - Routing information from GPS and LBS data can be used to improve trip assignment algorithms.
Big Data for Modeling

• Tour-Based Models
  – Trip stops along a tour can be better understood – where they are and how often
  – Mode choice for tours and trips within them

• Activity Based Models
  – Reliable (anonymized) devices can be studied to better understand individual travel patterns to develop travel profiles
  – Long term (employment, school) and short term (daily routine) decisions impacting travel can be estimated
Trip and tour based models start with Production-Attractions and convert them into O-D to get the actual direction of trip before trip assignment.

Production-attraction format of trips expresses the directions going from home-end of the trip (production) to non-home end of the trip (attraction). That does not reflect the real directions from origin to destination.

StreetLight creates trips in O-D format. Knowing the home location (zone) of the device making the trip, the trips can be converted to P-A format.
Step 1: Pick the Right Data

- Cellular
- LBS
- Ad-derived Data
- Active Mode App
- GPS-Survey
- Counters
- Traditional Surveys
Step 2: Machine Learning to Recognize Modes at the Ping Level

**Training a Random Forest Classifier – Data Sets by Source**

<table>
<thead>
<tr>
<th>Source</th>
<th>Tagged Points</th>
<th>Harvested Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caltrans Travel Survey (NREL**)</td>
<td>26M</td>
<td>~500K</td>
</tr>
<tr>
<td>Atlanta Regional Travel Survey (NREL)</td>
<td>2.4M</td>
<td>~75K</td>
</tr>
<tr>
<td>Mid-Region Travel Survey – Albuquerque (NREL)</td>
<td>3M</td>
<td>~92K</td>
</tr>
<tr>
<td>Southern Nevada Household Travel Survey (NREL)</td>
<td>4.2M</td>
<td>~133K</td>
</tr>
<tr>
<td>Capital Bikeshare</td>
<td>334K</td>
<td>~48K</td>
</tr>
<tr>
<td>Beijing Pedestrian (Microsoft)</td>
<td>5.5M</td>
<td>~27K</td>
</tr>
<tr>
<td>Total Number of Points in Training Data Set</td>
<td></td>
<td>~900K</td>
</tr>
</tbody>
</table>
3 METHODOLOGY + VALIDATION

RAW Data

Build ML Algorithm

LEARN

Train ML Algorithm

EVALUATE

Evaluate training results

Modify ML Algorithm

ITERATE

Launch
Step 3: Group “Pings” into Mode-Assigned Trips

- Apply intelligence from machine learning process to infer probability of mode choice for each ping
- Stationary is a “mode”
Step 4: “Lock” to Allowable Networks

• No geometry subtracted from car Open Street Map (OSM), only added

• Implications for Pedestrians
  – Complexities with pedestrians: jaywalking can be missed
  – Very large and spread out buildings, like conferences centers, can be confusing
Inferred home and work locations

Identifying Inferred Locations

For home locations, we look at an entire calendar month, and identify the top five neighborhoods where a device pings during evening and night-time hours. Each of the five neighborhoods is assigned a probability weighting.

For example, a device may have a home location with 75% probability in Philadelphia and 25% in Cape May, N.J., distributed across three neighborhoods in Philadelphia and two neighborhoods in Cape May, N.J. Then, combining devices with a home neighborhood in Philadelphia and Cape May with their probability weighting would give us the Philadelphia/Cape May based composite travelers.

Similarly, for work locations, we look at the top five neighborhoods where a device pings during work hours only.

Contextualizing and Aggregating

StreetLight adds context by incorporating demographic information based on the 2010 U.S. Census and aggregates to 1km x 1km grids. Metrics are provided through StreetLight InSight® and can be viewed in our interactive visualizations.

StreetLight can also provide output at other geographic unit levels such as Census Block Groups, Zip Codes, metropolitan areas, and states.
StreetLight AADT is trained on a Set of Permanent Counters

- StreetLight 2020 AADT data is trained and validated using 3,000+ unique permanent counter locations (6,600 permanent counts) across 25 U.S. states.

- The $R^2$ between StreetLight AADT and 2020 AADT values from permanent counters is 0.98, indicating a very strong relationship.
Continued year-over-year improvements to StreetLight’s AADT Metric

AADT 2020 improves results for small traffic volume roads through large traffic volume roads, due to updated machine-learning model methodology

<table>
<thead>
<tr>
<th>AADT Volume Range</th>
<th>StreetLight 2018 AADT MAPE (%)</th>
<th>StreetLight 2019 AADT MAPE (%)</th>
<th>StreetLight 2020 AADT MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: &lt;= 499</td>
<td>840.4</td>
<td>42.51</td>
<td>74.81</td>
</tr>
<tr>
<td>B: 500 - 1,999</td>
<td>27.29</td>
<td>27.76</td>
<td>25.48</td>
</tr>
<tr>
<td>D: 5,000 - 9,999</td>
<td>17.47</td>
<td>14.72</td>
<td>12.52</td>
</tr>
<tr>
<td>E: 10,000 - 19,999</td>
<td>16.21</td>
<td>12.7</td>
<td>11.71</td>
</tr>
<tr>
<td>F: 20,000 - 34,999</td>
<td>13.72</td>
<td>10.53</td>
<td>8.84</td>
</tr>
<tr>
<td>G: 35,000 - 54,999</td>
<td>13.98</td>
<td>9.36</td>
<td>7.32</td>
</tr>
<tr>
<td>H: 55,000 - 84,999</td>
<td>11.54</td>
<td>8.38</td>
<td>7.55</td>
</tr>
<tr>
<td>I: 85,000 - 124,999</td>
<td>11.58</td>
<td>8.63</td>
<td>7.09</td>
</tr>
<tr>
<td>J: &gt; 125,000+</td>
<td>11.19</td>
<td>8.41</td>
<td>8.99</td>
</tr>
</tbody>
</table>

Comparison of cross validation results of mean percent error (MAPE) metric among StreetLight’s AADT models for 2018, 2019, and 2020 (U.S.).

Download the full white paper to go deeper on the AADT methodology and validation. ✅ StreetLightData.com /AADT2020
# Accuracy exceeds industry standards compared to temporary counts

For medium and large roads, the AADT 2020 model performs competitively with 48-hour same-year temporary counts across most accuracy metrics and consistently better than the typical situations with no counts.

<table>
<thead>
<tr>
<th>AADT Volume Range</th>
<th>Method (n)</th>
<th>95% TCE Error Range (%)</th>
<th>Median Bias (%)</th>
<th>MAPE (%)</th>
<th>NRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-499 (very low)</td>
<td>Same Year Temporary Counts (SY-TC)</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>500 - 4,999 (low)</td>
<td>SY-TC</td>
<td>34</td>
<td>-0.1</td>
<td>10.2</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>Typ-NC</td>
<td>n/a</td>
<td>n/a</td>
<td>50</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>AADT 2020 K-fold</td>
<td>40.50</td>
<td>4.3</td>
<td>19.0</td>
<td>24.0</td>
</tr>
<tr>
<td>5,000 – 54,999 (medium)</td>
<td>SY-TC</td>
<td>28</td>
<td>1.1</td>
<td>8.6</td>
<td>14.2</td>
</tr>
<tr>
<td></td>
<td>Typ-NC</td>
<td>n/a</td>
<td>2</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>AADT 2020 K-fold</td>
<td>27.88</td>
<td>-0.7</td>
<td>10.6</td>
<td>15.4</td>
</tr>
<tr>
<td>55,000+ (high)</td>
<td>SY-TC</td>
<td>24</td>
<td>1.4</td>
<td>5.3</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>Typ-NC</td>
<td>n/a</td>
<td>1.5</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>AADT 2020 K-fold</td>
<td>15.65</td>
<td>-0.9</td>
<td>8.1</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Note: there are no targets available for roads with AADT under 500.


StreetLight’s AADT 2020 cross-validated results compared to 48-hour same year temporary count expansion as represented by Krile et al (2015) and typical situations with no counts for key statistical indicators.
Validation of AADT counts in Tennessee

StL AADT Cross-Validation vs TDOT Permanent Counts

$R^2 = 0.9894$