STREETLIGHT INSIGHT

StreetLight Data Sources and Methodology White Paper

Updated December 2022

STREETLIGHT

Table of Contents

StreetLight Data, Inc. (“StreetLight”) pioneered the use of Big Data analytics to shed light on how people, goods, and services move, empowering smarter, data-driven transportation decisions. StreetLight's proprietary data processing engine, Route Science® algorithmically transforms its vast data resources to measure travel patterns of vehicles, bicycles and pedestrians, accessible as analytics on the StreetLight InSight® SaaS platform. StreetLight provides innovative digital solutions to help communities reduce congestion, improve safe and equitable transportation, and maximize the positive impact of infrastructure investment.

This white paper describes the data sources and methodology employed by StreetLight to develop travel pattern Metrics. This document is relevant for all StreetLight InSight® Metrics, whether they are available via the StreetLight InSight® platform, data API, or custom delivery.

Data Sources

Since 2011, we have harnessed hundreds of data sources that contribute to our RouteScience® engine, developing unmatched transportation data processing capabilities and a deep, empirical understanding of how North America’s roads, sidewalks and transit interact. Our market-leading repository continues to expand to take advantage of new data sources, new transportation modes, and changing travel patterns.

StreetLight’s Metrics are primarily derived from the following list; the relative share of each source has changed over time as a result of changes in availability (as shown in Figure 1, below):

- Connected Vehicle Data (CVD)
- GPS data
- Commercial truck data for a range of weight classes
- Location-based services (LBS) mobility data
- Thousands of vehicular, bicycle and pedestrian sensors
- Land use data, parcel data
- Census characteristics (e.g., demographics, vehicle ownership, housing density)
- Road network and characteristics from OpenStreetMap (OSM)

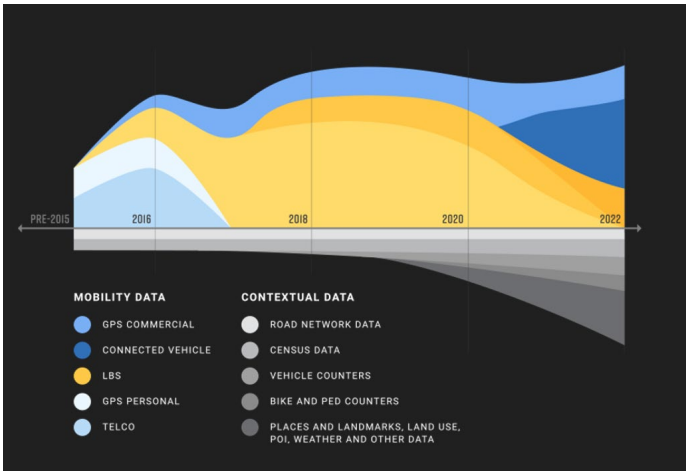

Figure 1: This chart demonstrates the evolution of StreetLight’s robust historical data foundation.

Data Processing Methodology

The following section contains an overview of the fundamental methodology that StreetLight uses to develop all metrics. Each StreetLight InSight® Metric has specific methodological details which can be shared with clients as needed on request.

Step 1 – ETL (Extract, Transform, and Load)

First, we pull data in bulk batches from our suppliers' secure cloud environments. This occurs daily, weekly, or monthly, depending on the supplier.

The ETL process retrieves the data securely from one environment to another, eliminates corrupted or spurious points, reorganizes data, and indexes it for faster retrieval and more efficient storage.

Step 2 – Data Cleaning and Quality Assurance

After the ETL process, StreetLight runs several automated quality assurance tests to establish key parameters of the data. To give a few examples, StreetLight conducts tests to:

- Verify that the volume of data has not changed unexpectedly.
- Ensure the data is properly geolocated.
- Confirm the data shares similar patterns to the previous batch of data from that specific supplier.

In addition, StreetLight staff visually and manually review key statistics about each data set. If anomalies or flaws are found, the data is reviewed by StreetLight in detail. Any concerns are escalated to our suppliers for further discussion.

Step 3 – Implement Mode Tagging Algorithms

While we know our navigation-GPS data comes from commercial trucks and that CVD comes from personal vehicles, we do not have labeled modes in our LBS data. Since LBS data is derived from smartphones rather than vehicles themselves, we assume that all modes of travel could be included in our LBS data set.

To address this, we designed a way to ingest mode-agnostic LBS data and identify likely vehicle, bus, rail, bike, and walking trips. We built a multi-pass algorithm to identify each mode of travel using a combination of heuristic and probabilistic algorithms, including machine learning.

After much research, we developed the following approach to infer probable mode of travel. We first isolate air and boat/ferry trips (although not currently offered in the StreetLight InSight® platform, they are available for custom analyses via our Services team), and then subsequently isolate the remaining modes as follows:

1. Isolate rail trips based on a series of heuristic rules, given the fact that rail trips occur along a uniquely identifiable rail network.
2. Identify and group points into walking or stationary trips. As an example, a device continuously pinging at rest has a unique signature in comparison to modes of travel.
3. The remaining, unassigned LBS pings are assigned mode probabilities for vehicle, bike, and bus, based on the movement patterns and spatial features described below.

Expanding on step three outlined above, in order to assign mode probabilities to every ingested LBS ping, we relied on machine learning techniques, specifically a random forest model to isolate vehicle, bike, and bus trips from one another.To do this, our model utilizes various sources of training data, including a combination of NREL's Transportation Secure Data Center, GPS data from transit agencies, data derived from bus lines in service, and vehicular data from navigation-GPS devices. We complement these data sources with a proprietary set of mobility data developed in-house by StreetLight.

In the end, we evaluated a subset of the features (attributes or explanatory variables in the model) in order to assess what was most impactful in the training of the random forest classifier. This included evaluating many features inherent to the pings themselves, such as time, distance, speed, acceleration, circuity, and angular velocity for each ping as well as for its preceding and subsequent pings, day of week, hour of day, and so on. We also considered contextual features, such as road classification, road network density, presence of bike and bus lanes, and proximity to parks.

We took a probabilistic approach to consume the results of the mode classifier. When we run our algorithm, vehicle, bike, and bus pings are not assigned a single mode. Instead, they are assigned a mode probability distribution (0–1).

Step 4 – Create Trips

For any type of data supply, the next step is to group the data into key patterns. For example, for navigation-GPS data and CVD, a series of data points whose first timestamp is early in the morning, travels at reasonable speeds for a number of minutes, and then stands still for several minutes could be grouped into a probable trip.

With LBS data, this approach to trip creation is more complex given the presence of multiple modes. We analyze the pings ordered by their timestamp and assign a stream of points to either a rail, stationary, or walking trip—or leave them unassigned. For remaining pings not assigned to rail, stationary, or walking, we predict travel mode using the machine-learning, model-based classifier (as described in Step 3). The stream of mode-tagged pings is then sent to trip creation algorithms.

When we observe a sequence where the mode probability changes, we end the current trip and start the new trip with the new mode. For example, if we detect a string of 10 pings that are all high vehicle probability, followed by a string of five high-probability bike pings, then we create two trips: a vehicle trip followed by a bike trip. Table 1 below illustrates this process. We may also end a trip when a device pings at rest for an extended period of time. We refer to this

period of non-movement as a stationary trip, as mentioned in Step 3. Stationary trips are identified when we detect pings clustered within a very small area, moving at very low speeds for 5 or more minutes. This approach aims to end trips when minimal movement is detected, but to continue trips in scenarios where vehicles might be stalled at a traffic light or moving slowly in heavy congestion.

As a final check, we verify the correctness and feasibility of each trip for the mode assigned to it. If we find a trip that appears to be missing its end or beginning—for example, a bus trip that stops in the middle of the highway—we eliminate it. If we find a trip that appears to have erroneous data—for example, goes from California to Ethiopia and back in four seconds—we eliminate it.

Ping Timestamp	Bus	Vehicle	Bike	Prediction
3/28/21 10:51	0.036	0.939	0.025	Vehicle Trip
3/28/21 10:53	0.087	0.872	0.041	Vehicle Trip
3/28/21 10:55	0.109	0.752	0.139	Vehicle Trip
3/28/21 10:57	0.261	0.137	0.602	Bike Trip
3/28/21 10:59	0.215	0.320	0.465	Bike Trip
3/28/21 11:01	0.162	0.165	0.673	Bike Trip

Table 1: The table highlights an example of how we approach mode inference for bus, bike, and vehicle, by assigning mode probabilities and grouping pings into trips. Based on the mode with the highest probability, we predict the mode for each ping and string pings into trips.

Step 5 – Trip Locking

A trip from LBS, CVD, or GPS is defined as a series of connected pings. Since these pings do not necessarily lie on the network (road or otherwise), we need to tie them to the links on the network. This process is called Trip Locking (aka map matching). If the traveler turns a corner but the device is pinging only every 10 seconds, then that intersection might be missed when all the device's pings are connected to form a trip. For all data sources and modes, StreetLight utilizes network information from OpenStreetMap (OSM), including route types, speed limits, and directionality to lock the trip to the network. This locking process ensures that the complete route of the vehicle, bus, train, etc., is represented, even though discrepancies in ping frequency may occasionally occur. Figure 2, below, illustrates this process.

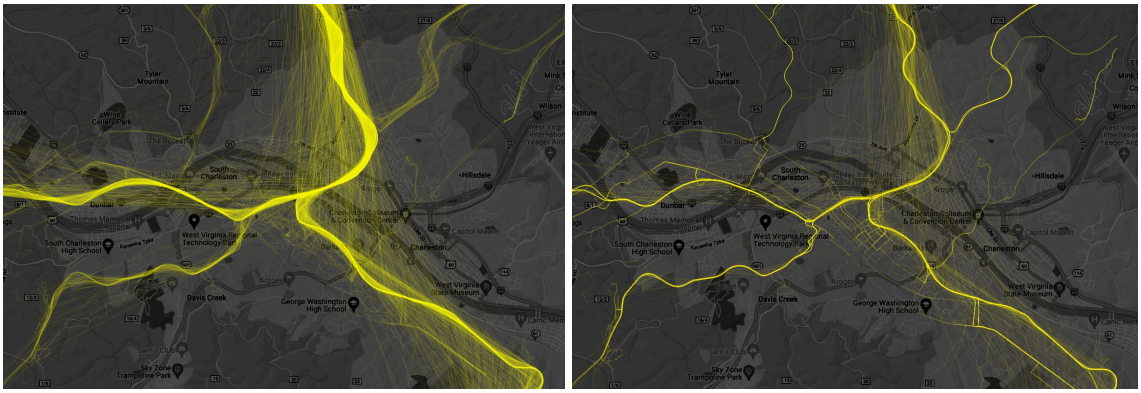

Figure 2: Unlocked trips (left) become locked trips (right).

With our LBS data, each individual mode is locked to its respective OSM network, meaning there may be some OSM layers that are not accessible to all modes. For example, we should not lock vehicle, bus, or rail trips to a recreational trail or path; those should only be accessible to bike and walk trips. Similarly, we should not lock bike or walk trips to highways, those should only be accessible to vehicle and bus trips. Rail trips will be locked only to rail lines and will never lock to the vehicular road network, though there may be some locations where rail and vehicular OSM lines share the same road (think: streetcars). We also honor one-way roads if they are identified as such in OSM.

Step 6 – Contextualize

Next, StreetLight integrates other contextual data sets to add richness and improve accuracy. These include road network information like speed limits and directionality, and geospatial features, such as the presence of bus routes and bike lanes, proximity to parks, and road network density. We also rely on land use data, parcel data, census data, and more.

Our demographic data sources for the U.S. are the 2010 and 2020 Census, as well as the 2019 American Community Surveys. In Canada, our source is Manifold Data. These contextual sources allow us to normalize the LBS sample to the population when needed and to add rich insights to analytics of travelers, such as trip purpose and demographics.

Step 7 – More Quality Assurance

After patterns and context are established, additional automatic quality assurance tests are conducted to flag patterns that appear suspicious or unusual. For example, if a trip appears to start at 50 miles per hour in the middle of a four-lane highway, that start is flagged as bad. Flagged trips and activities are not deleted from databases altogether, but they are filtered out from StreetLight InSight® queries and Metrics.

Given the nature of our mode-tagging algorithms, mode-based quality assurance is more involved and includes the following additional checks:

1. **Mode classification of individual pings:** To test the classifier, we employed techniques called folding that are typically used in testing machine learning algorithms. We trained the model on 80% of the training data and tested it against the 20% we had held back and not exposed to the model.
2. **Unit-level testing for trips:** For testing the start and end of the trips, we hand-picked 100+ unit test trips against which we verified the trip boundaries, the overall trip travel mode, and the mode of the individual pings as determined by the system.
3. **Meso-level testing for trips:** We looked at certain regions, primarily metropolitan areas, to make sure trip distributions looked reasonable. We focused on areas with well-known behavior—like bus and bike lanes—to verify that our trips mirror the real world.
4. **System-level testing for trips:** We performed tests on the overall trips generated in each month of data, including visual checks, statistical checks, spatial and spatiotemporal checks, and real-world data comparisons.

Step 8 – Normalization and Expansion to Estimated Trip Counts

Normalizing and expanding the data is necessary to derive a StreetLight Volume estimate, or a StreetLight Index value. Since we expect fluctuation in our suppliers' data samples month over month, we need to design approaches to normalize our sample data and make it comparable over time. For example, a 20% increase in our sample size one month may not mean that traffic has actually increased by 20%.

For our CVD trips, as well as LBS trips tagged as vehicle, bike, and walk, we can expand our sample to estimate the actual flow of travel. This process involves implementing machine learning models that produce monthly estimates for each respective mode at a given location. These models rely on an extensive set of permanent loop counters for both training and validation. Counters are aggregated on a monthly basis for up-to-date benchmarks of real-world activity. For more details on how those models are trained and tested, as well as expected errors (see our All Vehicles Volume, Bicycle Volume, and Pedestrian Volume white papers).

For commercial trips (navigation-GPS), we have two distinct approaches to normalization. First, StreetLight uses a modeling approach similar to the one described in the previous paragraph in order to produce a Truck Volume estimate accessible for certain analysis types (see our Truck Volume white paper for more details). Second, we rely on a set of permanent loop counters at certain highway locations to measure the change in trip activity each month. Then we compare the total monthly navigation-GPS trips to the monthly recordings from the permanent loop counters and normalize appropriately, resulting in a Metric called the StreetLight Index.

For our bus and rail trips (LBS), StreetLight uses a set of ridership metrics provided by transit agencies to measure the change in trip activity each month. Using a set of bus and rail station polygons, we quantify the number of bus and rail trips that start at each station, compare those to the agency-reported values, and use this ratio to normalize appropriately. Due to limited availability of agency data and the high frequency of monthly updates in StreetLight's data processing pipeline, we determined that creating monthly expansion ratios based on agency counts would be risky due to the uncertainty availability of timely truth data. Instead, we

calculate expansion ratios for each mode in a seed month in 2019, then use variation in our LBS vehicular penetration rates to adjust scaling values month over month. As a result, the StreetLight sample trips for bus and rail are normalized to the StreetLight Index to adjust for change in our sample size.

Step 9 – Store Clean Data in Secure Data Repository

After being made into patterns, checked for quality assurance, normalized, expanded, and contextualized, the data is stored in a proprietary format. This enables extremely efficient responses to queries via the StreetLight InSight® platform. By the time the data reaches this step, it takes up less than 5% of the initial space of the data before ETL.

Step 10 – Aggregate in Response to Queries

Whenever a user runs an analysis via StreetLight InSight®, the platform automatically pulls the processed data from the data repository and aggregates the results. For example, if a user wants to know the share of trips from origin zone A to destination zone B versus destination zone C during September 2021, they specify these parameters in StreetLight InSight®. Trips that originated in origin zone A and ended in either destination zone B or destination C during September 2021 will be pulled from the data repositories, aggregated and expanded appropriately, and organized into the desired metrics.

Results always describe aggregate travel patterns, not the travel of individuals.

Step 11 – Final Metric Quality Assurance

Before delivering results to the user, final metric quality assurance steps are automatically performed. First, StreetLight InSight® determines if the analysis zones are appropriate. If they are nonviable polygon shapes, outside of the coverage area (for example, in an ocean), or too small (for example, analyzing trips that end at a single household), the zone will be flagged for review. If a Metric returns a result with too few trips or activities to be statistically valid or to protect privacy, the result will be flagged. When results are flagged, StreetLight's support team reviews the results to determine if they are sufficient to deliver from a statistical or privacy protection perspective. The support team then discusses the best next steps with the user.

In general, StreetLight InSight® response time varies according to the size and complexity of the user's analysis. Some analysis runs take seconds or minutes, while others may take several hours. Users receive email notifications when projects are complete, and they can also monitor progress within StreetLight InSight®. Results can be viewed as interactive maps and charts within the platform or downloaded as CSV and shapefiles to be used in other tools.

Measuring Sample Size

We have 1 billion+ monthly trips with our CVD data sources, which is comparable to LBS sources from prior months. Sample size and penetration rate for a given analysis depend on the specific parameters used in the study. The available sample varies across time and space and

mode. In addition, some data are useful for certain analyses, but not for others. Efficiently identifying the data that is useful for a particular analysis plays a critical role in the data science value that distinguishes StreetLight. Because penetration rates vary, sample sizes are automatically provided for almost all StreetLight InSight® analyses[1]. This allows users to calculate penetration rates in order to better evaluate the representativeness of the sample. Sample size values also are useful to clients who wish to explore StreetLight InSight® results through additional statistical analysis.

For LBS analyses, sample size is currently provided as the number of unique devices and/or number of trips, depending on the type of analysis. For CVD analyses, sample size is currently provided as the number of trips per project. Sample size is provided as number of trips for navigation-GPS analyses. These values may be thought of as conceptually similar to vehicle trips.

For commercial navigation-GPS data analyses, commercial trucks that rely on up-to-date fleet management tools are more likely to be included in StreetLight's navigation-GPS data set than fleets that lag in adoption of such tools.

[1] Sample sizes are not automatically provided for AADT analyses. They are available by request. These analyses use a very large volume of location data, so providing sample sizes automatically via StreetLight InSight® would negatively impact data processing speeds.

STREETLIGHT DATA