Our Methodology and Data Sources

Updated October 2018

StreetLight InSight® Metrics: Our Methodology and Data Sources

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

Table of Contents

Locational Data Sources and Probe Technologies

StreetLight Data's Metrics are currently derived from two types of locational "Big Data": navigation-GPS data and Location-Based Services (LBS) data. StreetLight has incorporated and evaluated several other types of mobile data supply in the past, including cellular tower and ad-network derived data.

As the mobile data supply landscape has evolved and matured over time, we have determined that a combination of navigation-GPS data and LBS data is best suited to meet the needs of transportation planners. Our team phased out the use of cellular tower data because its low spatial precision and infrequent pinging frequency did not meet our standards for use in corridor studies, routing analyses, and many other Metrics. LBS data is suitable for these studies and offers a comparable sample size to cellular tower data.

As of July 2018, StreetLight's data repositories process analytics for about 65M devices, or ~23% of the adult US and Canadian population, and about 12% of commercial truck trips. As detailed later in this report, sample size varies regionally, historically and by type of analysis conducted.

Our data supply grows each month as updated data sets are provided by suppliers. We currently use one major navigation-GPS data supplier, INRIX, and one LBS data supplier, Cuebiq. See Table 1, below, for more details on the different locational data sources StreetLight Data has recently evaluated.

Table 1 – Overview of Big Data supply options for transportation analytics. StreetLight recommends and uses a mix-and-match approach currently focused on navigation-GPS and LBS data types.

Type	Pros	Cons	Notes
Cellular Tower: Derived from cellular tower "triangulation" and/or "multi-lateration" (100-2000m spatial precision)	• Large sample size - Most telecom providers have over 30M devices • Ability to infer home and work locations	• Very poor spatial precision (average of several hundred meters) • Infrequent pings for some suppliers • High cost • Consumers typically opt-out of data collection (vs. opt-in) • No differentiation of personal and commercial trips • Poor coverage in rural areas • No capture of short trips • No ability to reliably infer active modes of transportation	We haven't seen the US cellular industry making investments to improve these weaknesses.
In-Vehicle Navigation-GPS: From connected cars and trucks (3-5m spatial precision)	• Excellent spatial precision • Very frequent pings • Separates personal and commercial trips • Opt-in for consumers	• Usually lower sample size • Difficulties inferring home/work (depending on supplier practices) • No non-vehicular modes	This data has been traditionally used for speed products.
Location-Based Services: Mix of navigation-GPS, aGPS, and sensor proximity data from apps that "foreground" and "background" with locational data collection (5-25m spatial precision)	• Very good spatial precision • Frequent ping rate • Superior ability to infer trip purpose and trip chains • Ability to infer modes (walk/bike/transit/Gig Driving) accurately • Large and growing sample size • Opt-in for consumers	• Less mature suppliers • Variation in sample size and characteristics across suppliers requires more sophisticated data processing	Several players are emerging in this new market with very large sample sizes, opening up the possibility of a healthy, competitive supply base.
Ad-Network Derived Data: When user sees an ad on their phone, their location is recorded by the ad-network	• Large sample size of individuals	• Few pings per month mean inference of travel patterns is not feasible	This source should not be used until significant changes are made.

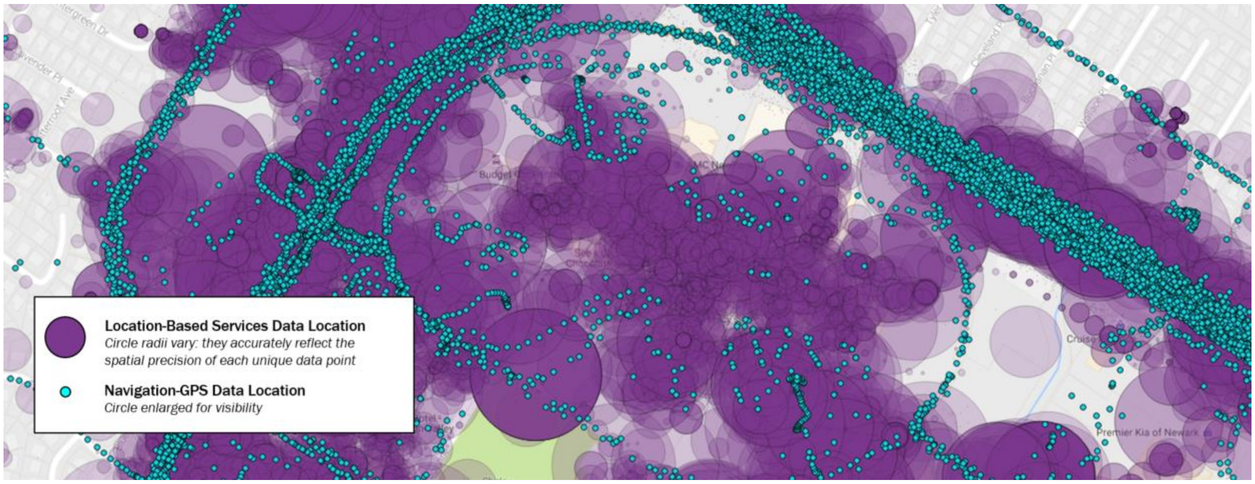
Our Navigation-GPS and LBS Data Sources

In this section, we will explain why access to two different Big Data sources is uniquely beneficial for transportation professionals. First, it is important to note that *StreetLight InSight* is:

- The first and only on-demand platform for planners to process Big Data into customized transportation analytics to their unique specifications, including the type of Big Data they would like to use.
- The first and only online platform that automatically provides comprehensive sample size information for analyses. (See more information on sample size on page 8 of this report.)

We selected navigation-GPS and LBS data because they are complementary resources that provide unique and valuable travel pattern information for transportation planning. See Figure 1 below for a visualization of these data sources.

Figure 1 – Filtered visualization of a subset of unprocessed navigation-GPS and LBS data near a mall in Fremont, California.


Location-Based Services Data Location
Circle radii vary: they accurately reflect the spatial precision of each unique data point
Navigation-GPS Data Location
Circle enlarged for visibility
Premier Kia of Newark

Location-Based Services (LBS) Data

LBS data can be processed into personal travel patterns at a comprehensive scale. Its fairly high spatial precision and regular ping rate allow for capturing trips as well as activity patterns (i.e.: home and work locations), trip purpose, and demographics. This makes it an ideal alternative to data derived from cellular towers, which also has a large sample size but unfortunately lacks spatial precision and pings infrequently.

Cuebiq, our LBS data supplier, provides pieces of software (called SDKs) to developers of mobile apps to facilitate Location-Based Services. These smartphone apps include couponing, dating, weather, tourism, productivity, locating nearby services (i.e.: finding the closest restaurants, banks, or gas stations), and many more apps, all of which utilize their users' location in the physical world as part of their value. The apps collect anonymous user locations when they are operating in the foreground. In addition, these apps may collect anonymous user locations when operating in the background. This "background" data collection occurs when the device is moving. LBS software collects data with WiFi proximity, a-GPS and several other technologies. In fact, locations may be collected when devices are without cell coverage or in airplane mode. Additionally, all the data that StreetLight uses has better than 20-meter spatial precision. (Similarly, our partner INRIX collects some LBS data from navigation-oriented smart phone apps).

Navigation-GPS Data

Navigation-GPS data has a smaller sample size than LBS data, but it does differentiate commercial truck trips from personal vehicle trips. This makes navigation-GPS data ideal for commercial travel pattern analyses. Navigation-GPS data is also suitable for very fine resolution personal vehicle travel analyses (e.g.: speed along a very short road segment) because of its extremely high spatial precision and very frequent ping rate.

INRIX, our navigation-GPS data supplier, provides data that comes from commercial fleet navigation systems, navigation-GPS devices in personal vehicles, and turn-by-turn navigation smartphone apps. (These apps produce data that are like the LBS data described above). Segmented analytics for medium-duty and heavy-duty commercial trucks are available. For commercial trucks, if the vehicle's on-board fleet management system is within INRIX's partner system, INRIX (and thus StreetLight) will collect a ping every one to three minutes whenever the vehicle is on, even if the driver is not actively using navigation.

For personal vehicles, if the vehicle is in INRIX's partner system and has a navigation console, INRIX (and thus StreetLight) will collect a "ping" every few seconds whenever the vehicle is on, even if the driver is not actively using the navigation system. This provides a very complete picture of vehicles' travel patterns and certainty that the trips are in vehicles.

Data Processing Methodology

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

Step 1 – ETL (Extract Transform and Load)

First, we pull data in bulk batches from our suppliers' secure cloud environments. This can occur daily, weekly, or monthly, depending on the supplier. The data do not contain any personally identifying information. They have been de-identified by suppliers before they are obtained by StreetLight. StreetLight Data does not possess data that contains any personally identifying information.

The ETL process not only pulls the data from one environment securely to another, but also 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, we run several automated, rigorous quality assurance tests to establish key parameters of the data. To give a few examples, we conduct 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 particular supplier.

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

Step 3 – Create Trips and Activities

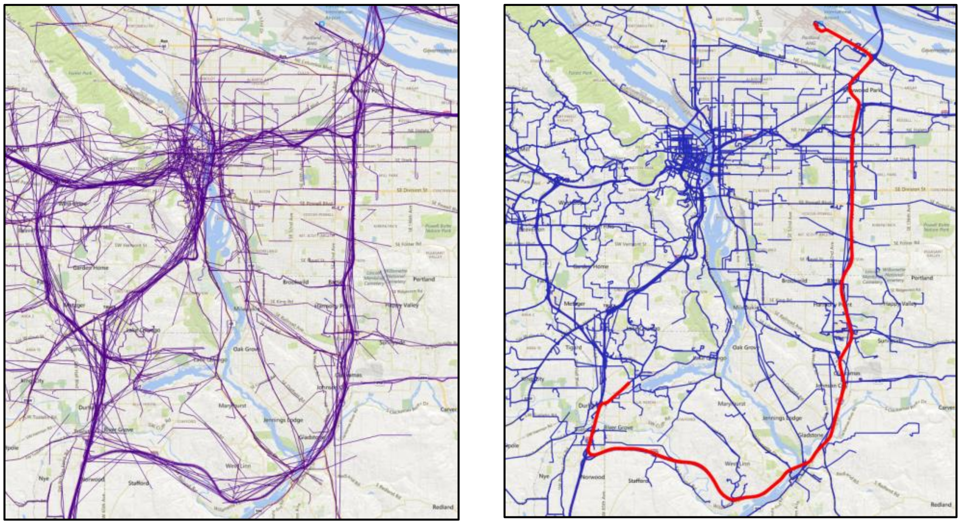
For any type of data supply, the next step is to group the data into key patterns. For example, for navigation-GPS data, a series of data points whose first time stamp is early in the morning, travels at reasonable speeds for a number of minutes, and then stands still for several minutes, could be grouped into a probable “trip.” For LBS data, we follow a similar approach. However, since LBS data continues to ping while the device is at the destination, we see clusters of pings in close proximity at the beginnings and ends of trips.

Step 4 – Contextualize

Next, StreetLight integrates other “contextual” data sets to add richness and improve accuracy of the mobile data. These include road networks and information like speed limits and directionality, land use data, parcel data, and census data, and more.

For example, a “trip” from a navigation-GPS or LBS device is a series of connected dots. If the traveler turns a corner but the device is only pinging every 10 seconds, then that intersection might be “missed” when all the device’s pings are connected to form a complete trip. StreetLight utilizes road network information including speed limits and directionality, to “lock” the trip to the road network. This “locking” process ensures that the complete route of the vehicle is represented, even though discrepancies in ping frequency may occur. Figure 2, below, illustrates this process.

Figure 2: “Unlocked” Trips becoming locked trips.



As another example, if a device that creates LBS data regularly pings on a block with residential land use, and those pings often occur overnight, there is a high probability that the owner of the device owner lives on that block/block group. This allows us to associate “home-based” trips and a “likely home location” to that device. In addition, we can append distribution of income and other demographics for residents of that census block to that device. That device can then “carry” that distribution everywhere else it goes. (Our demographic data sources for the US are the Census and American Community Surveys. In Canada, our source is Manifold Data.) This allows us to normalize the LBS sample to the population, and to add richness to analytics of travelers such as trip purpose and demographics.

Step 5 – 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.

Step 6 – Normalize

Next, the data is normalized along several different parameters to create the StreetLight Index. As all data suppliers change their sample size regularly (usually increasing it), monthly normalization occurs.

For LBS devices, we perform a population-level normalization for each month of data. For each census block, StreetLight measures the number of devices in that sample that appear to live there, and makes a ratio to the total population that are reported to live there. A device from a census block that has 1,000 residents and 200 StreetLight devices will be scaled differently everywhere in comparison to a device from a census block that has 1,000 residents and 500 StreetLight devices. Thus, the StreetLight Index for LBS data is normalized to adjust for any population sampling bias. It is not yet "expanded" to estimate the actual flow of travel.

For navigation-GPS trips, StreetLight uses a set of public loop counters at certain highway locations to measure the change in trip activity each month. Then it compares this ratio to the ratio of trips at the location, and normalizes appropriately. In addition, StreetLight systemically performs adjustments to best estimate total, normalized trips based on external calibration points. Such calibration points include public, high-quality vehicle count sensors (for example, those in PEMs systems, or the TMAS repository) as well as reports from surveys and other externally validated sources. Thus, the StreetLight Index for GPS data is normalized to adjust for change in our sample size. It is not normalized for population sampling bias (because we cannot infer home blocks for GPS data). This is one of the reasons we recommend LBS data for all personal travel analytics. The StreetLight Index for GPS data is not yet "expanded" to estimate the actual flow of travel.

Step 7 – Store Clean Data in Secure Data Repository

After being made into patterns, checked for quality assurance, normalized, 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. However, no information has been lost, and contextual richness has been added.

Step 8 – Aggregate in Response to Queries

Whenever a user runs a Metric query via *StreetLight InSight*, our platform automatically pulls the relevant trips 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 vs. Destination Zone C during September 2017, 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 2017 will be pulled from the data repositories, aggregated appropriately, and organized into the desired Metrics.

Results always describe aggregate behavior, never the behavior of individuals.

Step 9 – 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 personally reviews the results to determine if they are appropriate to deliver from a statistical/privacy perspective. The support team then personally 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 query. Some runs take two seconds. Some take two minutes. Some take several hours. Users receive email notifications when longer 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

StreetLight's Big Data resources include about 65M devices in the US and Canada, which covers approximately 23% of these countries' combined adult population. However, clients should not expect a 23% penetration rate for all *StreetLight InSight* analyses they run. Penetration rates for individual analyses can range from as small as 1% to as large as 35%.

As is the case with any Big Data provider, sample size and penetration rate for a given analysis depend on the specific parameters used in the study. The reason is that some data are useful for certain analyses, but are not useful for others. For example, a device may deliver high-quality, clean location data for one study, but messy, unusable location data – or no data at all – for another. Efficiently identifying the data that are "useful" for a particular analysis is a critical component of the data science value that differentiates StreetLight Data. Because penetration rates vary, sample sizes are automatically provided for almost all *StreetLight InSight* analyses[1]. This allows users to calculate penetration rates and to better evaluate the representativeness of the sample. Sample size values also are

[1] Sample sizes are not automatically provided for Visitor Home-Work, AADT, or Traffic Diagnostics Projects. 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.

useful to clients who wish to normalize *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 for LBS analyses, depending on the type of analysis. These values should be thought of as most similar to "person trips." Including both the number of devices and trips for all LBS analyses is in our product roadmap. Sample size is provided as number of trips for navigation-GPS analyses. These should be thought of as "vehicle trips."

In general, though not always, the trip sample size for commercial navigation-GPS data will be higher than the device (truck) sample size. Commercial trucks that are in active use typically take many trips per week that are often on set routes; thus, they are more likely to have up-to-date fleet management tools, and that means they are more likely to be included in StreetLight's navigation-GPS data set. Trucks that are more rarely used are less likely to be included in the data set.

In general, though not always, the trip sample size for LBS data will be lower than the device (person) sample size. The reason is that not all devices in StreetLight's database capture every single trip perfectly. To illustrate, consider this hypothetical example:

- 8:00AM: Device creates location data at expected home location
- 2:00PM: Device creates location data at sports arena

This device has created useful information for analyzing the home locations of visitors to the arena. However, since the device didn't create any location data on the trip to arena, perhaps because it was off, then the route taken and the travel time cannot be calculated with certainty. As result, it could not be used in an analysis of road activity on an arterial near the arena.

As another example, consider a device that generates regular pings for each trip taken over 10 days. However, the user deletes the smart phone app that created that data, and it stops pinging. That device then disappears for the last 20 days of the month. The device's data can still be used, but the trip penetration for the month is only 33% of this person's trips, not 100%.

Typical daily trip penetration rates are between 1 and 5% of all trips on any one specific day. StreetLight's pricing and data structure encourage looking at many days of data. The costs are the same for analyzing an average day across three months and analyzing a single day. Thus, we encourage clients to evaluate the total sample across the entire study period instead of focusing on per-day penetration rates.