A technical memorandum to

Mobility Measurement in Urban Transportation (MMUT)

Pooled Fund Study

Submitted by the

TEXAS A&M TRANSPORTATION INSTITUTE

TECHNICAL MEMORANDUM

Tools and Best Practices for Using Passive Origin-Destination Data

Task 4 – Syntheses Development (Synthesis 3 of 3)

FY 2019

Authors:
Michael Martin
Byron Chigoy
Ed Hard

Texas A&M Transportation Institute

August 2019

[This page intentionally left blank.]

EXECUTIVE SUMMARY

The recent application of passive origin-destination (O-D) data for travel studies has empowered transportation professionals with the ability to quickly and efficiently observe population movements between areas and along roadways. As such, the passive O-D data consumer market continues to grow in terms of product options, analysis tools, and providers. As the supply and demand of passive data continues to increase, it is important for transportation experts and officials to stay abreast of the facts to be educated consumers.

This synthesis covers the key data technology characteristics and offers a comparison of technologies to elucidate what the data represent, what are the available products and tools, and offer suitable applications with best practice examples from several departments of transportation (DOTs) and metropolitan planning organizations (MPOs) from around the country.

The first step toward understanding passive O-D data is to know their source technologies. These fall into three main categories:

- **Global Positioning System (GPS)**: positional information gathered through GPS satellites by mobile device applications, vehicle navigation units, or commercial fleet-tracking and telematics.
- **Location-Based Services (LBS):** positional information derived from mobile device application (app) software development kits (SDKs) that rely upon a variety of technologies with various levels of precision and accuracy to locate the devices.
- **Cellular:** data derived from Call Detail Records (CDR) that are generated through the interaction of mobile devices with a cellular network.

Data passively gathered from the public using these technologies is considered crowdsourced, as opposed to point-to-point sensor technologies, such as stationary Bluetooth and Wi-Fi detectors, which more limited in scope.

At this point, confusion can quickly ensue since some technologies fall into both groupings depending upon their application. For example, LBS relies on a combination of technologies like GPS, Bluetooth, Wi-Fi, cellular triangulation, etc. to locate devices as best as it can. If a smart device cannot acquire a GPS signal because it is located deep within a building, then the application SDK will try the next best option, Wi-Fi combined with cellular triangulation to create the best guess of where the device is located.

What sets LBS apart as a separate passive O-D technology category is that it is what powers the location services of smart device applications. LBS was originally intended for advertisement and marketing purposes within smart device applications. Though it uses GPS technology, the data are vastly different from traditional passive GPS O-D data, which are sourced from vehicle navigation units and/or commercial fleet tracking systems. The same goes for cellular data,

which are traditionally sourced from cellular network carriers and pre-processed by data vendors for the consumer.

These nuances are important to consider when applying passive O-D data to travel studies. Some technologies directly represent vehicles moving through time and space (e.g., GPS), whereas others are smart devices that are being located intermittently based on its user or the software application (e.g., cellular or LBS).

Although these sources offer a great variety of data in vast quantities, the data still only represent a sample of the population. This sample comes with biases that must be controlled for if the data are used to produce generalized results about the population. Biases are introduced at the source and that source must be understood to have an idea what the data represent. Once biases are known and controlled for, the sample still represents a subset of the population and in some instances needs to be expanded. For transportation purposes, relevant traffic counts are the known measures to which the samples should be expanded.

As the passive data industry grows and new technologies, products, and tools are introduced to the market, the current data sources and their applicable uses will also change. To know whether a product or tool can be useful, potential users should be skeptical but also curious to understand the data and their true capabilities.

KEY POINTS

- Patterns may be more useful than specific numbers.
- Passive data offer a large sample that is not necessarily representative of the population—biases still exist that must be controlled for by weighting and expansion.
- The best selection of passive O-D data technology for a study is dependent upon study size, type, and objectives.
- Sample penetrations defer by passive O-D technology: GPS (0.5–2.0 percent), LBS (12–20 percent), and Cellular (15–25 percent)
- LBS data location accuracy and precision depends upon the technology that the app used to acquire the device’s location.
- Cellular data have fallen out of favor due to the advent of LBS, although a few products are still offered.
- LBS and cellular can distinguish between residents and visitors, whereas GPS cannot.
- GPS data can distinguish between passenger and commercial modes, whereas only the general mode of a trip (bike, pedestrian, and vehicle) can be inferred with LBS.
- GPS has a low penetration rate for passenger vehicles.
- GPS data provide the greatest spatial and temporal granularity for routing and travel time and speed estimation.
- LBS has a relatively high sample penetration but low sample frequency.
- Point-to-point sensor data are suited for corridor analyses of travel time, speed, and O-D matrices between sensors.

INTRODUCTION

In recent years, the mass adoption of smart mobile devices, connected vehicles, and the ever-expanding internet-of-things has provided the opportunity to passively observe the public's travel behavior. In other words, data collection can occur without the direct interaction of an individual or relying on respondents to take time out of their busy schedules to reveal details about their trips.

By eliminating respondent burden, origin-destination (O-D) travel data, which serve as an essential component of many transportation planning and operational studies, can be purchased and applied at a much lower cost in terms of money, time, and resources to studies like travel demand model development, corridor, traffic and revenue, and sub-area studies. Such studies provide essential information to assess transportation network performance and improvement alternatives.

Methods and practices used to collect and estimate these data are rapidly evolving through the pervasive collection and use of new technology passive O-D data. This new technology O-D data is entirely replacing traditional O-D data collection efforts such as field studies using surveys and video license plate capture.

The technologies that currently provide the bulk of passive O-D data for travel studies include:

- GPS.
- LBS.
- Cellular.

Each of the passive O-D data technologies has its own unique pros and cons depending upon the transportation planning or operational application. As of yet, there is not a one-size-fits-all panacea for transportation agencies to turn to for their data needs.

This synthesis provides an overview of the following key characteristics for each passively crowdsourced O-D data technology that will help agencies make an informed decision when purchasing and applying passive O-D data:

- What the data represent.
- Trip definitions.
- Positional accuracy and frequency.
- Sample penetration.
- Available sources and tools.
- Best practices examples from several DOTs and MPOs from around the country.

PASSIVE O-D DATA SOURCES

Third-party providers offer pre-processed GPS, LBS, and cellular O-D data obtained passively from mobile phones, in-vehicle navigation equipment, or apps that log device location information. These data are considered **crowdsourced** from the general public and fleet/connected vehicles having widespread, ubiquitous coverage. Whereas, Bluetooth and Wi-Fi technologies produce what are typically called **point-to-point sensor**, or vehicle/device re-identification, data because they are collected via two or more sensor devices that identify and match signals that passed both locations. The use of these technologies is ideal for corridor analyses related to travel time, speed, and development of sensor-to-sensor O-D matrices. Data collection is limited to sensor locations and is impractical for the collection of widespread, large-scale O-D data needed to identify location-to-location trips ends and trip purposes, but can offer ground-truth and benchmarking data for comparison to crowdsourced passive data.

The following are general definitions for each crowdsourced passive O-D data source:

- **GPS** data are positional information gathered through GPS satellite trilateration. It provides data on a device's location over time via time-stamped longitude/latitude coordinates.
- **LBS** data are positional information derived from mobile device app SDKs, bid stream/advertisement exchange data, retail beacons, and wireless networks.
- **Cellular** data are derived from the interaction of mobile devices with a cellular network, as known as CDR.

Although O-D data can be derived passively from GPS, LBS, and cellular technologies, the data from each technology have different characteristics and attributes that lend themselves to different uses. Because of their unique characteristics, the technologies have different capabilities and limitations in estimating various aspects of short- and long-distance travel. Table 1 provides an overview of the key characteristics of each passive data source.

Table 1. Overview of Crowdsourced Passive O-D Data Technologies.

Technology	Popular Data Providers	Popular Tools/Products	Data Summary	Relative Sample Size (Relative Unit)
GPS	- INRIX - ATRI - HERE - TomTom - Wejo	- StreetLight - Moonshadow - Independent Expert Analyses	- GPS ping; time-stamped coordinates of vehicle waypoints and trip ends - GPS begin and end points of vehicle trip - Can distinguish cars and trucks - Can determine travel speed and time - Daily/Monthly GPS pings of a unique vehicle (ATRI)	- Small (Passenger Vehicle) - Small (Service Trucks) - Medium (Heavy Trucks)

Technology	Popular Data Providers	Popular Tools/Products	Data Summary	Relative Sample Size (Relative Unit)
LBS	- Cuebiq - SafeGraph - Quadrant - Telenav - Skyhook - Groundtruth - Foursquare - PlaceIQ - Factual	- AirSage - StreetLight - Independent Expert Analyses	- Varies depending on device application - GPS points or locational reference points based on interaction with Wi-Fi, cellular towers, or geographic location where app is used - Data are usually provided as trip matrices for geographic zones provided by purchaser - Raw LBS data can also be obtained for custom analysis - Cannot distinguish cars and trucks - Can determine approximate travel speed and time	- Large (Total Population)
Cellular	- Wireless Carriers	- Verizon Traffic Data Services - Teralytics - Kido - Cellint	- Raw data are disaggregated and are collected from cell tower based on CDR - Data may represent the location of the cell tower that interacted with device or the devices position relative to cell tower coverage - Data are usually provided as trips matrices for geographic polygons provided by the purchaser - Cannot distinguish cars and trucks - Cannot determine travel speed and time	- Large (Total Population)

The following sections provide more detailed information of the characteristics, tools, and products for GPS, LBS, and cellular data technologies.

GPS

GPS data are positional information gathered through GPS satellite trilateration. It provides data on a device's location over time via time-stamped longitude/latitude coordinates. Third-party providers offer GPS data for purchase as unprocessed waypoints and trip ends, aggregated O-Ds, aggregated O-Ds with waypoints, or trip matrices for a requested traffic analysis zone (TAZ) structure. Groupings of GPS-based data include non-commercial and commercial categories based on the available data sources from which the data are aggregated.

GPS data sources include:

- **Mobile app GPS**, which is various LBS data, from non-navigation application SDKs such as random, ad-hoc locations based on device use or location;
- **Navigation GPS** data from mobile applications (e.g., TomTom and Waze) and in-vehicle navigation systems (for both fleet and non-fleet vehicles); and
- **Commercial fleet-tracking and telematics** data, used for tracking the movement and safety of fleet vehicles (both commercial and non-commercial), as well as for automobile insurance where vehicles are tracked to assess driving habits and provide discounts for good driving.

Current sources of GPS data collected for navigation and/or fleet tracking have a bias toward commercial vehicles, since the tracking of commercial fleets provides much of the data. On the other hand, this bias makes fleet tracking and navigation-GPS valuable for applications dependent on truck O-D data.

Non-commercial categories primarily reflect consumer vehicles, while commercial categories mostly reflect fleet vehicles such as long-haul trucking and local delivery or service fleets/vehicles. Current providers do not categorize GPS data into various segments including resident and non-resident travel for a study area, or by trip purpose, unlike cellular or LBS data.

Trip Definition

With GPS data, trips are determined using estimated stops isolated in the GPS data stream. These stops are identified based on dwell times when a device is stationary. Trip ends are the last point of a stop and the first point of the subsequent stop. Once trip ends are isolated, the GPS data stream points occurring during the stop are filtered out and the points occurring between trip ends provide waypoint information that coincides with the trip route.

Positional Accuracy and Frequency

The positional accuracy of GPS data is generally in the range of 1 to 10 meters in ideal conditions. However, phenomena such urban canyon effect, which distorts the GPS signal, or tunnels, which blocks the GPS signal, can greatly reduce the positional accuracy. Sample frequency is typically in terms of seconds or minutes between waypoints delivered. Collection of GPS data occurs either as a direct stream when connected to a Wi-Fi network or cellular network, or cached on a device when no network is present and later uploaded once reconnected.

Sample Penetration

GPS O-D data for passenger travel have a sample penetration rate in the range of 0.5–2.0 percent of average annual daily traffic (AADT), but the rate should be continually improving as GPS data become more widely captured. GPS O-D data for truck/freight travel have an estimated sample penetration rate of about 11 percent truck AADT. However, studies have found that in less dense traffic areas, especially at the boundaries of urban areas where external studies are conducted, the penetration rate for trucks may reach 40 percent of truck AADT.[1] The fact that these data represent a convenience sample drawn from devices and vehicles that possess such technology means that biases exist that must be controlled for. For example, newer, higher-end model vehicles and larger trucking companies are more likely to have GPS technology, thus possibly over-representing higher income households and long-haul routes such as interstates, respectively.[2]

Products and Tools

Providers of GPS waypoint and trip end data for O-D studies include INRIX, which provides mobile app, navigation GPS, and fleet-tracking GPS data and the American Transportation Research Institute (ATRI), which provides fleet tracking for commercial vehicles. Other sources of GPS data include TomTom O/D Analysis Trip Dynamics and HERE Traffic Analytics Trip Data. These two sources offer preprocessed O-D trip matrix products.

ATRI and INRIX Insights™ Trips Reports are comprised of GPS waypoints, but only INRIX provides unique trip IDs between origins and destinations. These GPS waypoint and trip end data sets require significant post-processing by the data user. This processing requires large computing resources and data management to extract and impute additional details of the data. These details include TAZ attribution, map matching to roadways, study area determinations of trip ends for trip type attribution (e.g., external-external [E-E]), and trip ends for ATRI only. Additional metrics that need to be calculated are speed and time between GPS waypoints. Despite the additional labor and computations involved in the analysis of GPS data, they offer a rich source of routing data and flexibility as compared to aggregated zonal data, such as cellular or point sensor technologies.

DB4IoT from Moonshadow is a time-series geospatial NoSQL database engine and analytics platform that can generate O-D matrices from INRIX Trips data. The DB4IoT platform facilitates very fast data visualizations and queries of very large data sets in terms of hundreds of millions of waypoint records within seconds.[3] As such, it eliminates a key hurdle for most data analysts—storing and querying big data sets. Due diligence is still necessary since the underlying data are a sample, and therefore, an incomplete picture of the real world.

StreetLight InSight® is an online platform tool that relies upon INRIX GPS, Cuebiq LBS, and other contextual data to produce on-demand analyses. The web-based tool allows users to perform customizable queries by data periods, day parts and types, and trip type. Analyses include: zonal origin-destination (volumes, average travel time, and average trip length), select link (pass-through zone), and zone activity analysis (total pass-through, originating, and terminating

---

[1] TxDOT Travel Survey Program external studies for Abilene, Amarillo, Austin, Dallas-Fort Worth, El Paso, Houston-Galveston, Killeen-Temple, San Antonio, Tyler, Waco, and Wichita Falls.

[2] Greg Giaimo, P.E., Ohio DOT. *Use of StreetLight OD Data for Travel Demand Models User Guide*. version 2, February 2018.

[3] Moonshadow DB4IoT, http://db4iot.com/

trips). Since StreetLight incorporates GPS and LBS data, the analysis results are derived from a combination. Currently, volume estimates are not expanded to traffic counts.[4] StreetLight also offers premium packages for trip and traveler attributes, such as: medium- and heavy-duty commercial trucks; trip purpose, duration, speed, length, and circuity; Census demographics (income, race, education, family status); and tourist and their home metro area/state. With this option, the purchaser only receives access to the online tool platform to perform the analyses listed above and is unable to access or retain the underlying passive O-D data.

LBS

LBS data are positional information derived from mobile device software app SDKs, bid stream/advertisement exchange data, retail beacons, and wireless networks. Various types of apps on mobile devices serve as data sources of location data, such as: weather, navigation, retailers, shopping, social connections, and others. Figure 1 illustrates an example sequence of events for gathering LBS location information from a mobile application SDK.

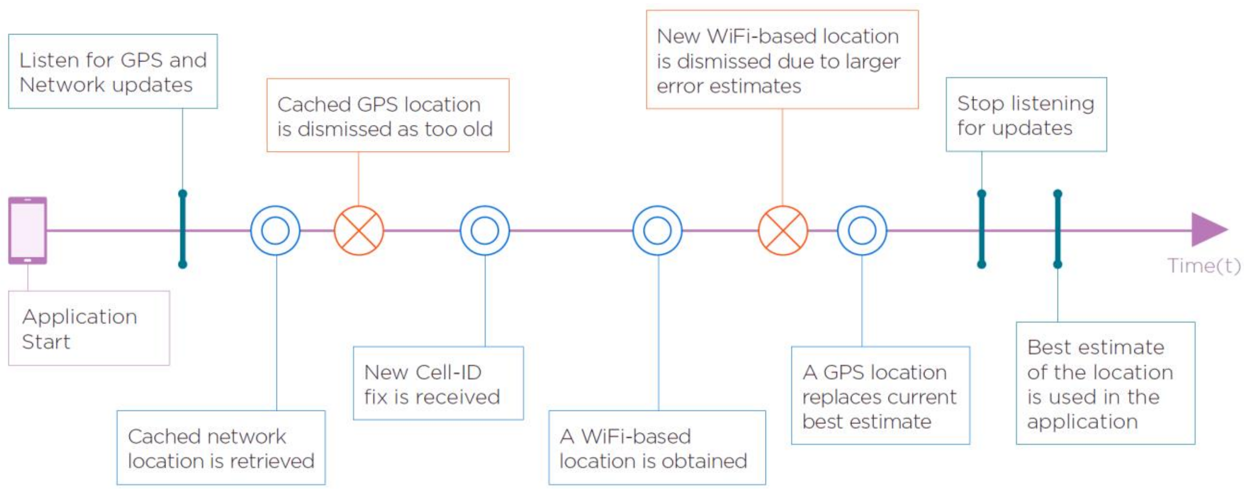

Figure 1. LBS Device Location Update Example Scenario. [5]

Once an application starts, it begins trying to locate the device through various technology options. In the example above, the device is located by multiple sources but at different times. Thus, positional accuracy and precision is dependent upon which technology is used and when the device was located last and is optimal when devices are stationary for periods of time, creating more opportunity for device sightings. There are tradeoffs the app developers must work around, such as GPS offering the most accurate and precise location data but requiring a high level of battery power to do so. Therefore, the developer may choose to use GPS location only once instead of constantly updating the device's location. For advertisement marketing purposes, this is likely a tradeoff that is acceptable. However, for transportation analysis

[4] StreetLight Data Methodology and Data Sources, https://www.streetlightdata.com/wp-content/uploads/2018/10/StreetLight-Data_Methodology-and-Data-Sources_181008.pdf
[5] Mobile Marketing Association, https://www.mmaglobal.com/documents/demystifying-location-data-accuracy

purposes, the infrequent location data will leave a lot left unknown, such as trip path or actual trip ends. Figure 2 provides an illustrative comparison between LBS (purple) and GPS (green) data in a spatial context. The GPS points are very precise, whereas the LBS buffers indicate the possible area in which a device is located.

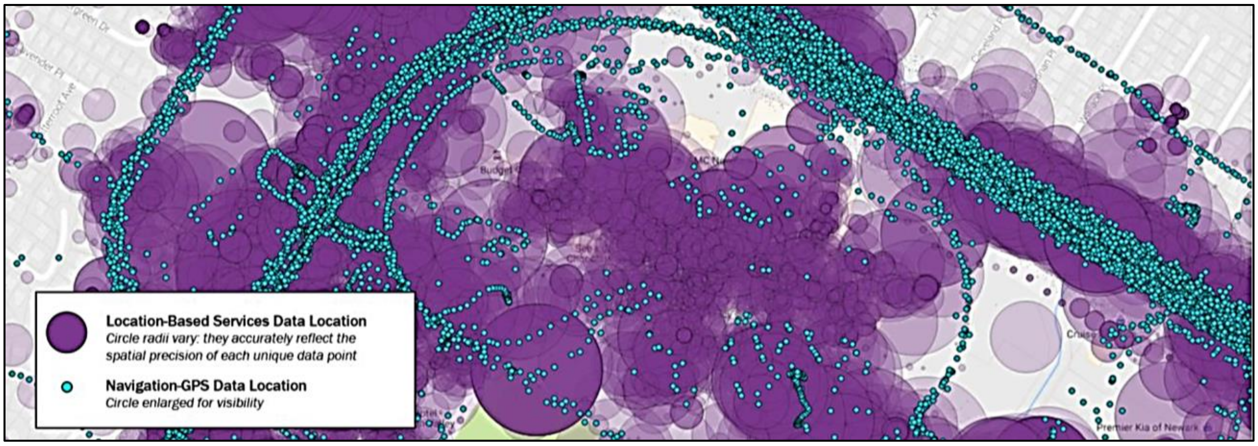

Figure 2. StreetLight GPS and LBS Data Example. [6]

Although in general LBS data do include trips from modes other than just vehicles, the data cannot discern the travel mode without complex machine learning techniques. Mode-specific LBS data can be obtained from apps that are dedicated to active transportation modes, such as apps that track biking or running metrics via a smartphone or smartwatch; however, these sources tend to be biased toward avid runners and cyclists.[7] Users can use LBS data to determine basic trip purposes, whether travelers are residents of or visitors to a study area since the device ID is persistent, and type of place can be inferred based on the time of day (e.g., home at night and work during the day). However, LBS data are derived from various mobile devices that may or may not travel with their users (e.g., a personal electronic tablet that remains home for nighttime reading versus a smartphone that travels with its owner and is used throughout the day for multiple purposes). These different use cases present challenges for analysts to accurately infer activities due to the variety of assumptions that must be made with incomplete information on device user behaviors and purposes.

Exactly how the LBS location data are collected depends on the type of app and the technology used by the app to determine the location of the device. Table 2 offers a list of LBS location data technologies, their data traits, and originally intended marketing use cases.

[6] StreetLight Data Methodology and Data Sources, https://www.streetlightdata.com/wp-content/uploads/2018/10/StreetLight-Data_Methodology-and-Data-Sources_181008.pdf

[7] StreetLight Active Mode Metrics, https://www.streetlightdata.com/wp-content/uploads/2019/01/StreetLight_Active-Mode_Methodology-and-Validation_190108.pdf

Table 2. LBS Data Technologies. [8]

Technologies	Location Data Traits	Marketing Use Cases
Beacons	- Very accurate and precise data, capable of identifying user location to within a matter of feet. - Currently lacks scale for broad marketing opportunities.	- In-store targeting by department or aisle, verification of store visits.
GPS	- Very accurate and precise data under the right conditions, capable of identifying user location to within 1–10 meters (in most ad platform scenarios). - Large scale but dependent on publisher and user enablement of GPS. - The quality of a GPS signal degrades significantly indoors or in locations that do not have an unobstructed view of multiple GPS satellites.	- Numerous targeting opportunities based on user location and context.
Wi-Fi	- Very accurate and precise data, particularly in signal-dense environments (e.g., buildings and malls). - Capable of identifying user location to within 10–100 meters when Wi-Fi signals are present.	- Numerous targeting opportunities based on user location and context.
Cell Triangulation	- Accurate but less precise data. - Typically capable of identifying user location within a zip or neighborhood area.	- Targeting over a broader geo-area such as a city or zip code.
IP Addresses	- Not reliably accurate or precise. - Typically based on IP address of app server, which can vary significantly from true location of mobile user.	- Not relevant for real-time or behavioral targeting.
Registration Data	- Not relevant for real-time or behavioral targeting. - Typically relevant only for a mobile user's home zip code at time of registration.	- Targeting at the zip code or marketing zone.

Cellular network connectivity is the primary method of transmitting LBS data. In some cases, mobile apps may cache the LBS data during periods without cellular network connectivity and

[8] Adapted from: Mobile Marketing Association, https://www.mmaglobal.com/documents/demystifying-location-data-accuracy

upload once reconnected. In-vehicle navigation GPS traces are not impacted by cellular network connectivity but represent a small sample in LBS data.

With the ubiquity of smart, connected devices, LBS has now effectively replaced cellular data in transportation studies due to its better location accuracy while still representing non-commercial personal trips despite a smaller sample penetration rate. Conversely, LBS locates devices less frequently than GPS with lower quality positional accuracy and precision depending on the technology that the supplying apps use. By pairing LBS data with GPS data from a provider like INRIX, ATRI, HERE or TomTom, a more complete picture of travel patterns can be viewed.

Trip Definition

Similar to GPS, trips are estimated using the LBS data stream and identifying trip ends based on dwell times when a device is stationary for some duration. Typically, a cluster of pings will be sighted at origins and destinations and can be clustered to indicate that the device is probably at a specific location for the user, for example home or work depending on the time stamps and land use, if available. Different proprietary algorithms exist depending upon the LBS analysis tool.

Positional Accuracy and Frequency

The accuracy and frequency of LBS data varies depending upon the technology chosen by the app to locate the device. Typically, GPS, Wi-Fi, or BT beacons are preferred choices since these technologies offer the highest levels of positional accuracy and precision, but Wi-Fi and beacons may not always in present. Based on information from StreetLight and AirSage, two main LBS data consumers, LBS accuracy can range from a few meters to up to about 100 meters. Since LBS device locations are sighted in a happenstance manner, the data may better represent longer trips with more exposer time as opposed to shorter ones. This bias can manifest into producing higher vehicle-miles traveled (VMT) estimates.[9]

Sample Penetration

The sample size of LBS data is less than that of cellular data, but still higher than GPS sources, especially for personal travel via passenger vehicles. StreetLight indicates their sample size for LBS data is approximately 12 percent of the U.S. population and 1–5 percent of daily trips on any specific day,[10] while AirSage indicates that their LBS sample size is similar to their older cellular data technology, which was approximately 15–20 percent of the U.S. population. However, it is suspected that sample size estimates are reflective of urban regions and this sample size for LBS will be less in rural and remote areas due mainly to the lack of network connectivity. Additionally, sample size information of LBS data is often based on the number of

---

[9] Greg Giaimo, P.E., Ohio DOT. *Use of StreetLight OD Data for Travel Demand Models User Guide*. version 2, February 2018.

[10] StreetLight Data Methodology and Data Sources, https://www.streetlightdata.com/wp-content/uploads/2018/10/StreetLight-Data_Methodology-and-Data-Sources_181008.pdf

applications that can provide these data, and in reality, not all installed applications are active on any given day. Additionally, some users limit or disable location services and/or tracking on their devices. Since smart devices (phones, watches, tablets, etc.) are so ubiquitous nowadays, LBS data are not hindered by a larger income bias, which has a negligible effect on trip making, but instead may suffer from large age and gender biases.[11]

Products and Tools

LBS data are offered in several forms from various vendors. To begin with, there are LBS data aggregators like Cuebiq, SafeGraph, and Quadrant that collect primary LBS data from mobile apps and retail anonymized data products to customers. Customers of these vendors may include transportation agencies with the need for a disaggregated O-D matrix product or Software as a Service (Saas) companies like StreetLight, Replica (by Sidewalk Labs), and Citilabs that build LBS data into a web-based analysis tool for on-demand analyses.

In 2017, AirSage enhanced its data insights platform, formerly built upon cellular carrier CDR, by integrating GPS and other LBS data sources from 120 million devices.[12] The sources comprise a panel of smartphone application SDKs, fleet, and navigation systems. The primary LBS data are used to produce pre-processed and expanded O-D matrix products segmented by time of day, travel purpose, and traveler type (resident or visitor) for an average weekday/weekend in a calendar month delivered in CSV format. These products include: regional trip matrix, long distance trip matrix, sub matrix (i.e., select-link analysis), and demographic trip matrix.

As mentioned in the GPS section, the StreetLight InSight® product uses a combination of GPS, LBS, and other contextual data to produce on-demand analyses with its web-based tool, StreetLight InSight®. StreetLight limits the spatial precision of its LBS data to 20 meters that it sources from Cuebiq, an LBS data aggregator.[13] The InSight® tool offers zonal O-D analysis (volumes, average travel time, and average trip length), and zone activity analysis (total passthrough, originating, and terminating trips). Currently, select link analysis (passthrough zone) is only offered with the use of GPS data and not LBS due to its spotty nature.[11]

Replica by Sidewalk Labs[14] is also a web-based, on-demand analysis tool that is powered by smartphone LBS data from approximately 5 percent of the population.[15] Replica uses LBS data along with various Census data sets to produce a synthetic population that statistically represents the full population through an opensource process called Doppelgänger. The goal is

---

[11] Greg Giaimo, P.E., Ohio DOT. *Use of StreetLight OD Data for Travel Demand Models User Guide*. version 2, February 2018.
[12] Airsage Sept. 13, 2017 press release, https://www.airsage.com/static/download/gps.pdf.
[13] StreetLight Data Methodology and Data Sources, https://www.streetlightdata.com/wp-content/uploads/2018/10/StreetLight-Data_Methodology-and-Data-Sources_181008.pdf.
[14] Sidewalk Labs – Replica, https://replica.sidewalklabs.com/.
[15] Replica – Introduction, https://medium.com/sidewalk-talk/introducing-replica-a-next-generation-urban-planning-tool-1b7425222e9e.

"to create a complete representation of the community, which allows her or him to quickly and easily summarize the community along any relevant dimension at any relevant geographic scale."[16] The synthetic population is used in partnership with LBS data to develop an agent- or activity-based model that is updated every three months to offer information on segment-specific traffic values by mode (car, transit, bike, or foot) and trip purpose. By updating to the model with current LBS data on a quarterly basis, users can better understand more current travel patterns as compared to using traditional survey data that are collected every 10 years. Currently, only travel behavior metrics can be produced. Future developments will incorporate the ability to explore intervention impacts on travel behavior.

Streetlytics™ developed by Citilabs is a proprietary optimization process software that uses a combination of LBS data, logical travel behavior estimates, and local traffic counts.[17] These three data inputs are used to produce three independent views of population movement. The software begins by weighting and correcting LBS data to produce a scaled estimate of the total population. Second, predictive modeling algorithms based upon traditional travel surveys and household/employment data are used to develop logical travel behavior estimates: number of trips, time of day, and mode of travel. Third, local traffic counts are gathered to provide an independent view of traffic volumes to reflect local conditions. Lastly, Streetlytics assigns a weight to each view according to their characteristics and data quality to produce an optimized view of the total population movement for an area.

Cellular

Cellular data are derived from the interaction of mobile devices with CDR. They are a measure of estimated movements of a cell device between pre-defined zones over a period of weeks or months. Cellular data do not represent actual trips, but rather estimated trips derived from analysis of the mobile device's movements and patterns over time. Figure 3 illustrates how devices are located by using the cellular network and difference in sighting frequency between on-call and off-call events. Devices that are actively making a call will be sighted more frequently than devices that are powered on but off-call. Traveling on-call devices are handed off between each cellular tower coverage areas to maintain network connection, whereas off-call devices interact less frequently with network towers.

[16] Doppelgänger – population synthesizer, https://medium.com/sidewalk-talk/a-first-step-toward-creating-a-digital-planning-laboratory-is-populating-it-beeb87d485f1.

[17] Citilabs – Streetlytics™, http://citilabs-website-resources.s3.amazonaws.com/resources/Streetlytics-Understanding_Movement.pdf.

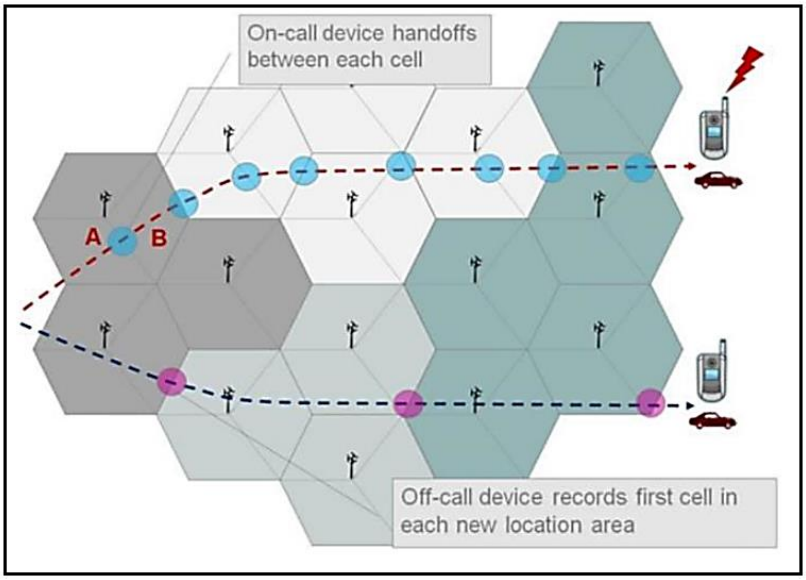

Figure 3. Illustration of CDR On-Call and Off-Call Sighting. [18]

Trip Definition

Like GPS and LBS, cellular data can infer trip ends based on dwell times when a device is stationary for some duration. For example, clustered device sightings around the same location during evening hours could be assumed to be the home location of the device. Usually pattern recognition and statistical clustering algorithms are used to better determine activity locations, such as home, work, etc., while trip imputation algorithms are used to determine trip ends.[18]

Positional Accuracy and Frequency

AirSage indicates that, on average, the accuracy of its data is (was) within about 300 meters, but varies depending on factors such as device type, device quality, and density of the cellular tower network. The smallest time interval over which cellular data can be reported is three hours because of the frequency of cellular data collection. In 2017, Airsage switched its platform from cellular to LBS data. They no longer offer cell-based data products.

Sample Penetration

However, cellular data's sample penetration range of 15–25 percent is its best characteristic due to the ubiquitous use and high market penetration of cellular devices. The penetration rate is largely dependent upon market penetration of cellular carriers with whom AirSage had an agreement. Cellular data allow identification of selected trip purposes, commuters, and a device's general home location (e.g., by city or county). However, they cannot discern between vehicle types (i.e., passenger versus freight). Cellular data also are biased toward to passenger vehicles, though commercial vehicle travel is present in the data.

---

[18] TMIP, FHWA-HEP-16-083. *Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data*, 2016. [https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/origin-destination/fhwahep16083.pdf.](https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/origin-destination/fhwahep16083.pdf)

Products and Tools

For many years, AirSage was the sole provider of pre-processed cellular O-D travel data in the United States. The emergence of LBS data as a quality O-D source has led to it effectively serving as an improved replacement for cellular data. This change, and Airsage's discontinuation of cellular-based O-D data, is an advancement and part of the evolution of passive data.

Despite the move toward LBS, there are a few companies that still offer preprocessed, aggregated cellular data products: Kido Dynamics and Teralytics. Both Switzerland-based startup companies have U.S. offices but are mostly focused on other international markets.

Verizon Traffic Data Services (TDS), a partnership between Verizon and Cellint, offers a passive O-D tool that is based on CDR trips georeferenced to road segments between their origins and destinations. TDS goes beyond using triangulation to locate mobile devices by using a reference location database of cellular signal signatures that are attributed to roadways every 250 meters in urban environments to then locate devices and their trajectories as they interact with cellular network towers.[19] According to Cellint, their software, TrafficSense, can determine exact routes a device travels while distinguishing between parallel, close-by roads, and separated lanes. Travel times and speeds can be estimated at the road segment level since the reference locations are every 250 meters in urban areas.[20] The Verizon TDS O-D tool appears to be better suited for smaller scale or individual corridor O-D studies. Using the tool for regional or areawide O-D studies would require driving out the entire roadway network to establish baselines, which would be impractical.

BEST PRACTICES

O-D data from GPS, LBS, and cellular technologies have different characteristics and attributes that lend themselves to different uses. Because of their unique characteristics, the technologies have different capabilities and limitations in estimating various aspects of travel.

Table 3 summarizes the strengths and weaknesses of each passive O-D technology. Ultimately, to determine the best passive technology completely depends upon the study size, type, and objectives.

---

[19] Verizon Traffic Data Services: reporting and data analytics using cellular data. https://enterprise.verizon.com/resources/reports/2018/traffic_data_services_congestion_analysis_reports.pdf.

[20] Cellint, How it works. http://www.cellint.com/trafficsense/#!/HowItWorks.

Table 3. Strengths and Weaknesses of Passive O-D Data Technology.[21]

Technology	Strengths	Weaknesses
GPS	- Can distinguish between passenger and truck/freight vehicles - Is well suited for truck/freight studies	- Unable to distinguish between resident and non-resident travel - Is biased toward truck/freight vehicles - Has low penetration rate for passenger travel
LBS	- LBS data have better positional accuracy than cellular - Good sampling rate for passenger vehicles (i.e., devices) - Good source for external-internal (E-I)/internal-external (I-E) data - Can distinguish between some trip purposes - Can distinguish between non-resident, commuters, and residents	- Cannot discern vehicle type - Travel routes cannot be easily ground truthed to an urban-scale - Provides no direct link between LBS data and subscriber demographics
Cellular	- Same as LBS strengths, except that sample size is larger	- Same as LBS, except for relatively low positional accuracy

Table 4 offers a quick reference for the suitability of passive O-D data technology by study type and use along with a point-to-point sensor suitability for comparison purposes. The level of technical expertise required for each study type and use will vary by technology. Some of the uses listed are still being researched and verified.

Please refer to the appendix for best practice examples for each data type from DOTs and MPOs from around the country. Each example offers details on the data application and the tools and products used for analysis.

[21] Adapted from: Arizona DOT. SPR-744, *Optimizing Technology for Collecting Arizona Long-Distance Travel Data*: https://www.azdot.gov/docs/default-source/research-reports/spr744.

Table 4. Suitability of Passive O-D Data Technology by Study Type and Use. [22]

Suitability by Study Type and Use		GPS	LBS	Cellular	Point-to-Point Sensor
Travel Demand Modelling	Four-step model calibration	✘	✓	✓	✘
	Activity-based model calibration	✘	✓	✓	✘
	Destination choice model	✓	✓	✓	✘
	Truck model	✓	✘	✘	✘
	Network calibration	✓	✓	✘	✘
	ODME	✓	✓	✓	✘
	Statewide O-D	✓	✓	✓	✘
	Regional O-D	✓	✓	✓	✘
	Demographic information	✘	✓	✓	✘
External Study	E-E trips	✓	✓	✓	✓
	E-I/I-E trips	✓	✓	✓	✘
	Trip purpose	✘	✓	✓	✘
	Commuter information	✘	✓	✓	✘
	Residency	✘	✓	✓	✘
	Commercial/Freight	✓	✓	✘	✘
	Route information	✓	✓	✘	✘
	Ability to apply travel time constraint	✓	✘	✘	✓
Corridor Study	Within urban areas (operational)	✓	✓	✘	✓
	Within urban areas (select-link analysis)	✓	✓	✓	✓
	County to county	✓	✓	✓	✘
	Multicounty metro regions	✓	✓	✓	✘
	Between major metro areas	✓	✓	✓	✘
Time-Period	Hourly or peak hour	✓	✘	✘	✓
	Peak period	✓	✓	✓	✓
	15-minute bins	✓	✓	✘	✓
	Average weekday, weekend, etc.	✓	✓	✓	✓
Miscellaneous	Population/human activity movements	✘	✓	✓	✘
	Travel time	✓	✓	✘	✓
	Travel speed	✓	✓	✘	✓
	Traffic operations studies	✓	✓	✘	✓
	Freight studies	✓	✓	✘	✘
	Bottleneck study	✓	✓	✘	✓
	Land use study	✘	✓	✓	✘
	Market analysis & transit study	✘	✓	✓	✘
	ITS & network monitoring	✓	✓	✘	✓

[22] Adapted from: TMIP, FHWA-HEP-16-083. *Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data*, 2016. https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/origin-destination/fhwahep16083.pdf.

CONCLUSIONS

These technologies—GPS, LBS, and cellular—are ubiquitous in today's smart, connected world, and with their continuous feed of information, travel behaviors can now be passively observed. By not having to physically measure or recruit participants, transportation agencies are able to reduce costs, estimate measures otherwise impossible, and have more up-to-date information of transportation conditions along road segments, at intersections, and within/between areas. Passive data have entered the market during a time when the populous is becoming more-and-more disinterested in taking time to provide input and with ever-increasing budgetary constraints. Although passive data offer relief in these areas, they are not without some drawbacks.

Passive data typically draw upon big data sources that offer an enormous volume of data and at a tremendous velocity. Although these sources offer a great variety of data in vast quantities, the data still only represent a sample of the population. This sample comes with biases that must be controlled for if the data are used to produce generalized results about the population. Biases are introduced at the source and that source must be understood to have an idea what the data represent. For example, GPS typically has a commercial vehicle bias due to one of its sources being fleet-tracking vehicles. But, what are the fleets comprised of: light, medium, or heavy trucks? Are the fleets local delivery vehicles or long-haul trucks? These questions must be explored by the end user of the data.

Once biases are known and controlled for, the sample still represents a subset of the population and in some instances needs to be expanded. For transportation purposes, relevant traffic counts are the known measures to which the samples should be expanded. Some passive O-D products and tools do this for their customers (e.g., Airsage LBS expanded trip matrix based upon Census tract population). However, others do not (e.g., INRIX GPS trips waypoints represent unique trips) and those will need to be expanded to traffic counts at particular locations.

There is also inherent uncertainty with each passive O-D technology and how the data are processed. For example, home and work locations are imputed in many instances based upon assumptions. These assumptions affect the veracity of the resulting passive data product and should be considered by the end user.

As the passive data industry grows and new technologies, products, and tools are introduced to the market, the current data sources and their applicable uses will also change. To know whether a product or tool can be useful, potential users should to be skeptical but also curious to understand the data and their true capabilities.

APPENDIX

GPS Best Practice Examples

Minnesota DOT (MnDOT) Rethinking I-94 Study (Minneapolis, MN): SRF employed INRIX GPS data to develop O-D matrices for the I-94 study corridor through Minneapolis and St. Paul. The objective was to inform design decisions for spot improvement locations, toll access locations, and left-side access ramp usage in terms of safety and operational issues. The O-D matrix created from the INRIX GPS data informed the select-link analysis for specific road segments to identify travel patterns, such as neighborhood usage, route choice, and trip length. Figure 4 provides an example of INRIX O-D flow band map between downtown Minneapolis and St. Paul.

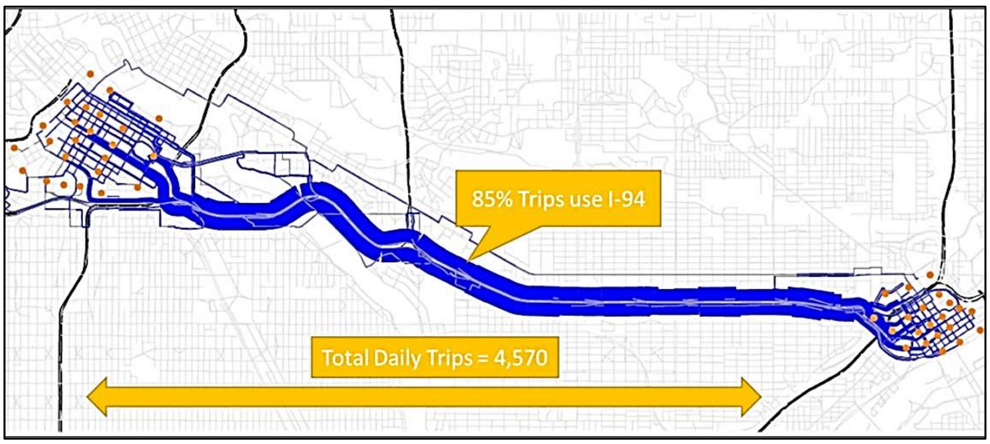

Figure 4. INRIX O-D Flow Band Map for the MnDOT Rethinking I-94 Study. [23]

TxDOT Permian Basin Impact Study: The booming energy sector in West Texas is testing the limits of the transportation infrastructure. The amount of travel demand is equivalent to that of a major urban area but located in a very remote rural setting. This presents the challenge of needing good, up-to-date traffic and road network data for a physical environment that is everchanging and that is so heavily dependent upon one thing: the price of oil.

The Texas Department of Transportation (TxDOT) and the Texas A&M Transportation Institute (TTI) turned to using GPS data from INRIX to help locate pop-up intersections where driveways essentially become busy intersections, in terms traffic volumes, overnight. INRIX waypoint density along US 285 north of Pecos was used to locate points along the roadways with turning traffic. In any many instances, the locations where not reflected on commonly available road network layers or satellite imagery. Figure 5 illustrates the travel shed for INRIX GPS trips from January–June 2018 that traveled along US 285 North. Knowing where these pop-up intersections are helps inform the current safety and operational improvement designs along the major roadway.

[23] MnDOT, https://www.dot.state.mn.us/I-94minneapolis-stpaul/vision.html.

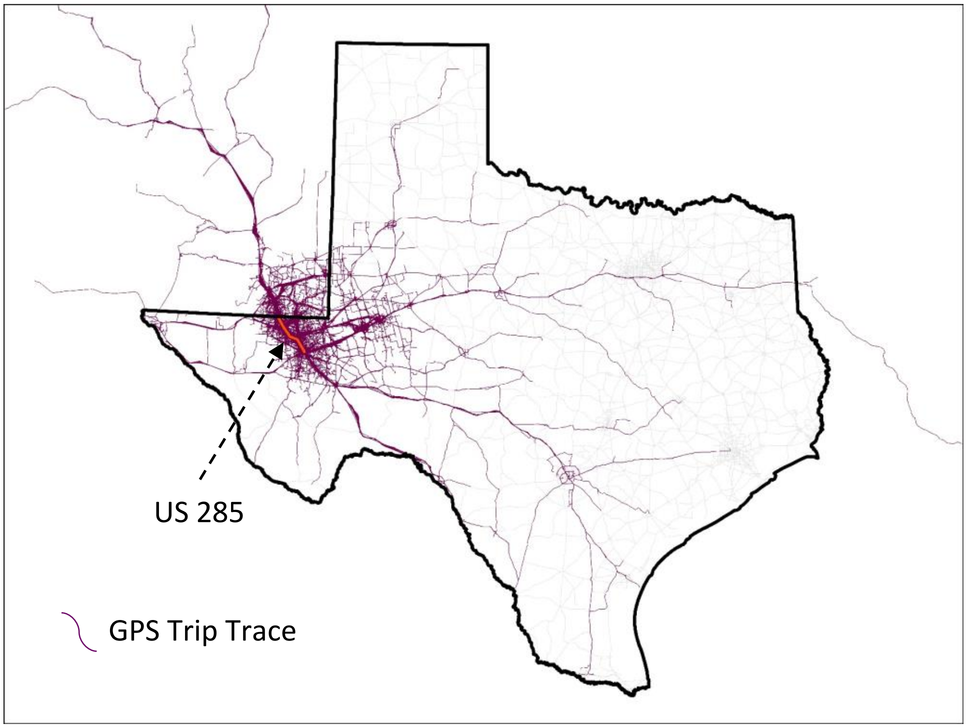

Figure 5. INRIX GPS Trip Traces for US 285 North of Pecos, TX. [24]

Maricopa Association of Governments (MAG) External Study (Phoenix, AZ): In 2018–2019, MAG is conducted an external study using two years of ATRI GPS and one month of AirSage LBS data to determine external trip volumes. Similar to the TxDOT external study process, the ATRI GPS waypoint data are processed by determining trip ends and filtering for external trips then applying a map-matching technique to assign roadway attributes to the GPS points in the proximity of each external roadway of interest. The resulting external volumes are then expanded to relevant traffic classification counts to represent the total population of external trips.[25] This project was completed in summer 2019.

TxDOT External Studies: Over the past five years, TxDOT has integrated passive O-D data into its formal Travel Survey Program to conduct external travel studies for 23 MPOs around the state with TTI's assistance. A combination of three months of INRIX GPS and one month of AirSage LBS data is used to assess the external travel into, out of, and through study areas. This combination of passive technologies helps to mitigate the biases of the individual data sources.

Unlike cellular data, the GPS data require significant post-processing by the data user. Due to the large amount GPS data and its raw form, the processing requires technical expertise, large computing resources, and efficient data management practices to extract and impute additional details of the data. Such details include TAZ attribution, map matching to roadways,

[24] TxDOT Odessa District.
[25] MAG 2019 external study.

and study area determinations of trip ends for E-E, E-I, and I-E attribution. Additional metrics that need to be calculated are speed and travel time between GPS waypoints. Finally, the GPS volumes per external station are expanded by growing the sample data to match a known population, such as traffic counts, using iterative proportional fitting. Despite the additional labor and computations involved in the analysis of GPS data, they offer a rich, disaggregate source of trip routing data and flexibility as compared to zone aggregated cellular data.

CATT Lab Trajectory Analytics Suite: The CATT Laboratory at the University of Maryland has developed web-based analytics suite of tool for INRIX GPS waypoint and trip end data shown in Figure 6. The suite currently contains data for the state of Maryland and Washington, D.C., proper and its metropolitan statistical area. It can also integrate agency transit ridership swipe data and activity-based model outputs. The O-D matrix, segment analysis, and route analysis tools allow users to visually explore travel times, origin and destinations of a selected link, determine specific trip routes, and filter for trip types and time ranges for the sample data.

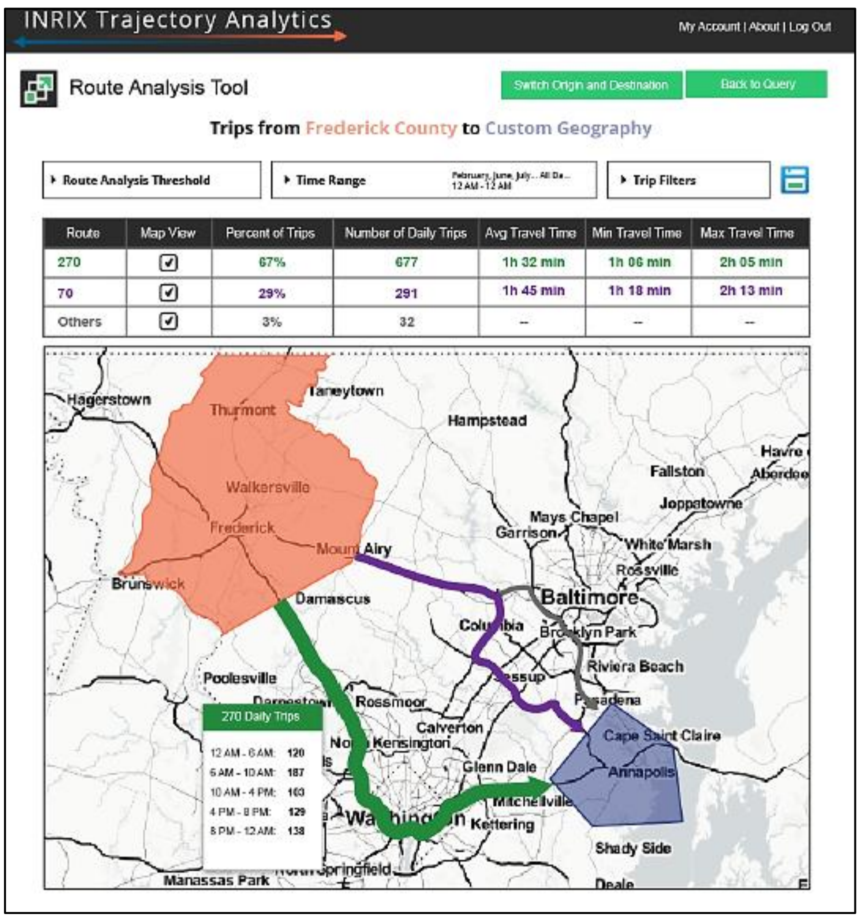

Figure 6. INRIX Trajectory Analytics Route Analysis Tool.[26]

[26] I-95 Corridor Coalition, https://i95coalition.org/wp-content/uploads/2018/12/I-95CC-O-D_Webinar-Dec2018-post.pdf?x70560.

LBS Best Practice Examples

MAG External Study (Phoenix, AZ): As previously mentioned in the GPS section, MAG recently conducted an external study using ATRI GPS and AirSage LBS data to determine external trip volumes.[27] The study uses TransCAD to perform a select-link analysis that assigns the AirSage LBS O-D matrix to a routable network. The matrix zone structure was developed to contain both internal and external zones. Since AirSage LBS data use only GPS-based sources, the high level of locational precision allowed for relatively small internal zones as compared to the external zones, which serve as the travel sheds around the study area. The routable network was developed by stitching together the local travel demand model network and nationwide Federal Highway Administration (FHWA) Freight Analysis Framework network.[28] The select-link analysis results are presented in terms of E-E, E-I, and I-E trips per station along the boundary. These totals are expanded results since AirSage pre-processes the LBS data to represent the total population per device home Census tract. Figure 7 shows an example of the E-E trip volumes visualized as flow bands between external stations located around the study area boundary.

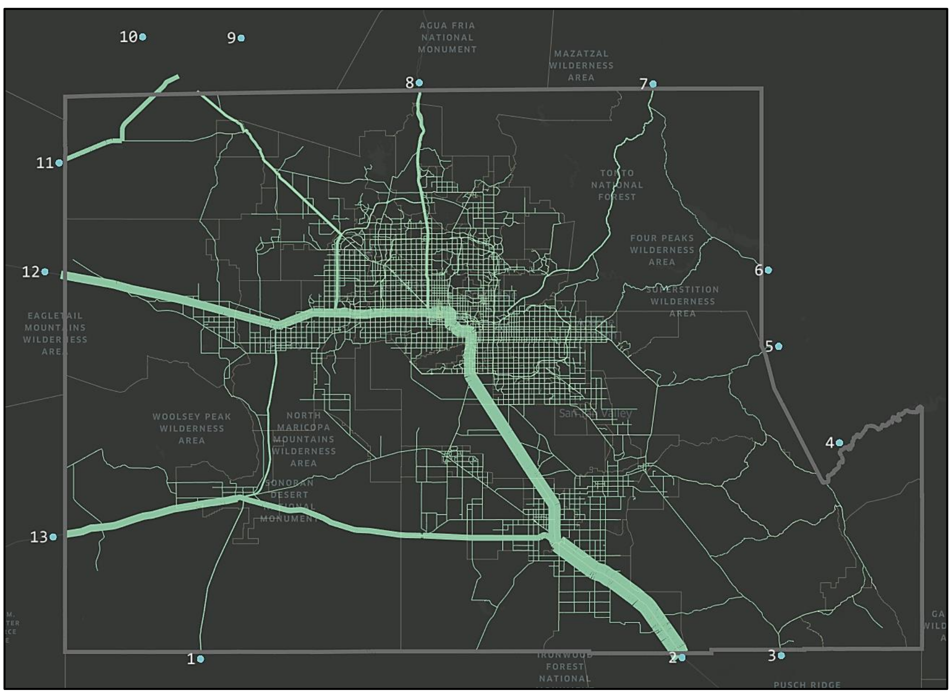

Figure 7. MAG External Analysis Using AirSage LBS Data.[27]

[27] MAG 2019 external study.
[28] FHWA Freight Analysis Framework, https://ops.fhwa.dot.gov/FREIGHT/freight_analysis/faf/index.htm.

Ohio DOT (ODOT) Model Updating and Validation: ODOT has developed procedures to use StreetLight InSight® O-D data for travel demand model updating and validation. A user guide (available upon request) was developed to cover common processing and data conversion steps that covers how to produce trip end summaries, trip length frequency distributions, coincidence ratios, trip assignment to model network, and factoring.[29] The user guide illustrates that passive data, in this case StreetLight, must be carefully examined and applied to avoid introducing bias error. For example, GPS and LBS passive data samples tend to generate higher VMT estimates as compared to model estimates. ODOT found that their model was generating too few short trips while the passive data are biased toward longer trips due to the nature of how it is collected. Figure 8 provides an example of model trip type estimates compared to StreetLight estimates showcasing that passive data E-E movements between major routes match well with external survey data use in the model, whereas local internal trips (I-I) between zones do not.

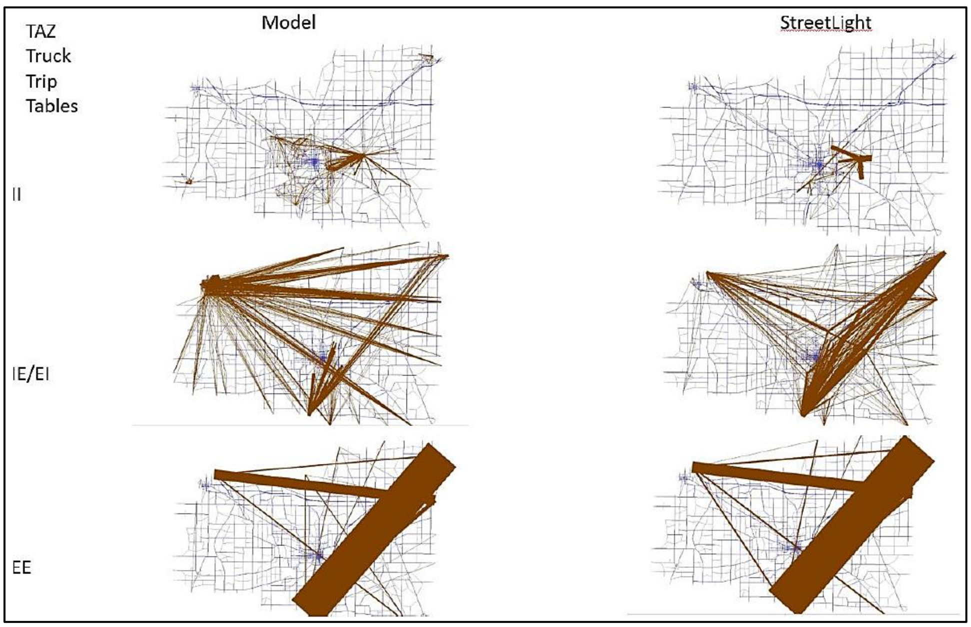

Figure 8. ODOT Travel Demand Model O-D Updates. [30]

Atlanta Regional Commission (ARC) External Model (Atlanta, GA): ARC used a combination of AirSage LBS and CDR data to develop an external model to predict trips for passenger cars and trucks entering and exiting the regional study area. The model estimates attractions for the following four passenger car purposes: work trips on interstates, non-work trips on interstates, work trips on non-interstates, and non-work trips on non-interstates. ARC developed two zone structures with identical internal zones for AirSage to use to aggregate

[29] Greg Giaimo, P.E., Ohio DOT. *Use of StreetLight OD Data for Travel Demand Models User Guide*. version 2, February 2018.
[30] Greg Giaimo, P.E., Ohio DOT. *Maintaining and Sustaining Travel Behavior Data Programs in a Changing World: State Department of Transportation Perspective*. TRB presentation, January 13, 2019.

passive data. The CDR data required larger external zone structures due to the data's low positional accuracy. The LBS data allowed for smaller zones to represent each external station in effect acting as a selected-link to capture trips and scaled use average daily traffic (ADTs) counts. The CDR data were disaggregated and scaled using ADTs for all stations in the catchment area (i.e., external zone). Both passive data sets were delivered and expanded to the zonal populations. The AirSage data were also used to determine the time of day shares to convert the daily estimates in the time periods used in the highway assignments in the model.[31]

City of Portland (Oregon), Metro, TriMet Replica Model Pilot: Currently, a Replica model for Portland, OR, is being developed through a collaboration between Sidewalk Labs and three local public agencies, each responsible for validating a different component as part of an intergovernmental agreement:

- City of Portland – bicycle and pedestrian volumes and movements;
- Metro – vehicle volumes; and
- TriMet – transit ridership.

Collectively, these agencies are working with Sidewalk Labs under their Replica Charter Customer Program to pilot test its capabilities for a year with the data being updated quarterly. The benefit to the agencies of paying $0.20 per resident during the initial service period (i.e., 1-year pilot) is a reduced cost of $0.12 per resident on an individual-agency basis for a subsequent year following the pilot test.[32] This study is on-going and is scheduled to be completed by late 2019 or early 2020.

Mid-America Regional Council (MARC) KC Regional Transportation Data Pilot – "Replica" Model (Kansas City, KS): MARC entered into an agreement with Sidewalk Labs to test and develop a Replica model for the greater Kansas City region and its public transportation agencies. The pilot is a 1-year project that involves building a Replica model as a data analysis tool to better predict travel patterns and mode choice.[33] MARC, in partnership with KDOT and MoDOT, have contracted with Westat to conduct a regional household travel survey with an approximate sample size of 4,000 starting in spring 2019. A primary goal of the survey is to compare respondent's revealed travel data with results produced by the Replica synthetic population/LBS data model to determine if the tool can produce detailed household travel diary information comparable to the survey and suitable for travel demand modeling and forecasting purposes.[34] This study is on-going and is scheduled to be completed by late 2019.

---

[31] ARC External Model draft report, received on March 1, 2018.
[32] Replica Charter Customer Program Terms (Metro), http://opb-imgserve-production.s3-website-us-west-2.amazonaws.com/original/replica_portland_term_sheet_1544652344627.pdf.
[33] MARC Total Transportation Policy Meeting minutes from October 16, 2018, http://www.marc.org/Transportation/Committees/agendas/TTPC/TTPCMeetingPacket_10162018.aspx.
[34] MARC Regional Household Travel Survey RFP, http://marc.org/Requests-for-Proposals/RFP/HHSRFPFinal.aspx.

Cellular Best Practice Examples

<u>TxDOT External Studies:</u> TxDOT has integrated passive O-D data into its formal Travel Survey Program to conduct external travel studies for 23 MPOs around the state with TTI's assistance. From 2014 to 2017, a combination of three months of INRIX GPS and one month of AirSage cellular data were used to assess the external travel into, out of, and through eight MPO study areas in Texas. This combination of passive technologies was used to better assess travel and trip making of consumer/passenger vehicles and commercial/truck traffic.

The cellular data component of the external travel studies was based on select-link analysis for each of the specified external stations around the study area's boundary. The analysis used a combined network comprised of the local model and the statewide model networks stitched together. This was necessary since the local model network does not extend past the study area boundary in the external travel shed. The Airsage cellular data were used for consumer/passenger travel and delivered as a matrix. External trip types were developed based on the O-D pairs and their respective zones' location with respect to being internal or external. The local external (E-I and I-E) and external pass-through (E-E) trips were then assigned to the combined network as part of the select-link analysis. The total trips and their corresponding types provided were provided per station along with a link-flow table. Figure 9 depicts the cellular E-E external trip assignment flow bands developed from the link flow table.

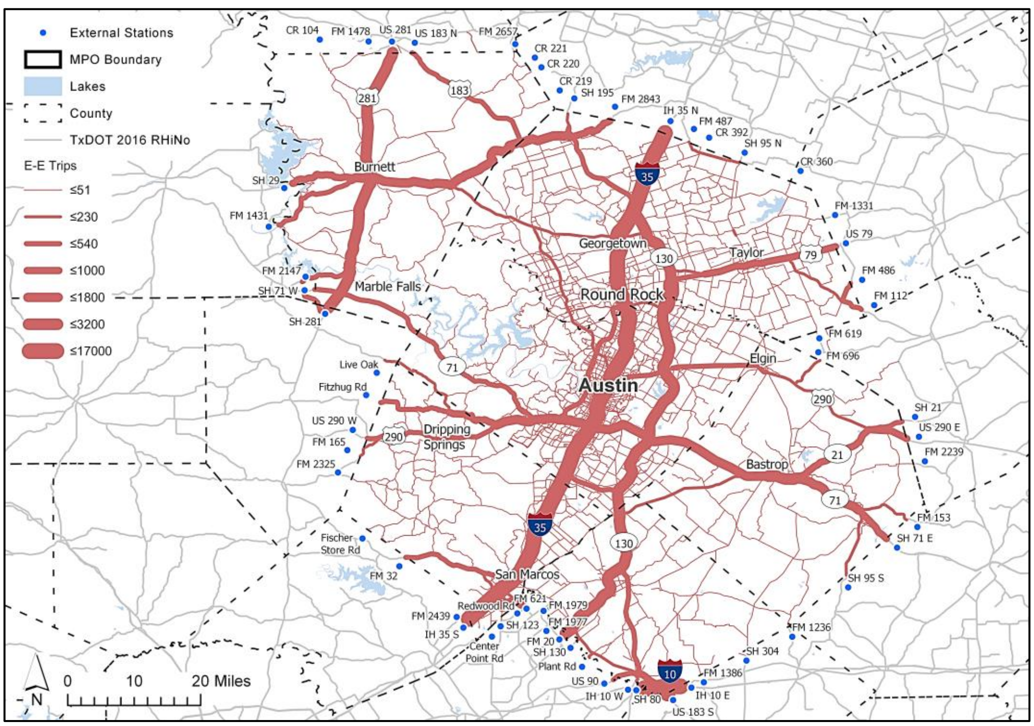

Figure 9. TxDOT Austin, TX, Cellular External Trip (E-E) Assignment Flow Bands. [35]

[35] TxDOT Travel Survey Program external studies for Abilene, Amarillo, Austin, Dallas-Fort Worth, El Paso, Houston-Galveston, Killeen-Temple, San Antonio, Tyler, Waco, and Wichita Falls.