Understanding Trip-Making with Big Data

Suppose there's a dispute over a busy commercial stretch in your town. The department of transportation, which is responsible for the road, says regional growth has caused more through-traffic and they must remove on-street parking to add lanes. Business owners and nearby residents, however, insist business is booming, they need the parking, and most of the traffic is local. Big data now lets us understand how the road's being used more easily than in that past.

Big data can sound intimidating, but the information it provides is becoming easier to access and interpret than ever before. Much in the same way that we can get real-time traffic updates from our smartphones, government officials and transportation professionals can now harness massive amounts of information about the way people move around to better understand how they should operate, manage, and invest in our transportation systems.

Traditionally, this sort of information comes from expensive studies that can take months to complete or from complex models that simulate real world conditions with varying degrees of accuracy. New data from in-vehicle GPS, cellphones, and mobile apps, however, are being collected all the time and turned into useful information that can often be purchased for a fraction of the cost of conventional sources. For transportation agencies and metropolitan planning organizations, big data can replace or complement more conventional data sources that may be less reliable or difficult to gather and compile.

In many cases, the main barrier to this data is a foggy understanding of what it is, where to get it, and how to use it. This brief provides guidance for practitioners regarding the different types of data available, their relative benefits and drawbacks, example applications, and lessons learned from a transit planning application in Sacramento, California – a project that involved several data providers, university researchers, philanthropic groups, and public agencies.

Key takeaways include:

- **Familiarize yourself with and consider a range of data sources.** Knowing different data providers and the features of each data source lets you choose the right one.
- **Ask a specific question to address a planning need.** Knowing this question helps point to a clear data need, a potential data provider, and a specific kind of analysis.
- **Enlist the right people to help interpret the data.** Interpreting the data might require basic GIS skills, basic data analysis, and knowledge of transportation planning principles.
- **Aggregate big data appropriately to leverage its full potential.** Reliable findings depend on large sample sizes, so working with the data may require trade-offs and generalizations.

This brief describes more detailed findings from the *Connecting Sacramento* study, led by SSTI, and offers guidance to those responsible for working with the data applying it in practice.

Introduction

Knowledge about where and when people travel, by what modes, and along which routes, is essential for planning, building, and operating efficient and useful transportation networks. Traditionally, this information has come from travel surveys, trip diaries, and traffic studies, but these methods are resource-intensive and not easily scalable. Newer technologies such as Bluetooth, mobile phones, and GPS-enabled devices all present new opportunities for understanding how people get around. The *Connecting Sacramento* study incorporates location-based trip-making data collected passively from two different sources – cellphones and mobile GPS devices – and as a proof-of-concept test, applies them to better understand travel patterns along transit corridors.

Data sources and types[1]

Bluetooth probe data are collected by setting up detectors that can identify Bluetooth devices by their unique addresses. The data are most commonly used in travel time studies, but they can also be used for trip distribution modeling and origin-destination studies. Several transportation agencies have used Bluetooth studies in place of license place matching and time lapse aerial photography (TLAP), due to its substantially lower cost. However, because the detectors are never deployed across an entire metro area, these studies do not provide full origin and destination data.

Location data from cellphones are based on their position relative to nearby cell towers. These data are readily available in urbanized areas at a coarse resolution. Like Bluetooth data, they have also been used in travel time studies and increasingly for origin-destination studies. The data are typically acquired through a third-party vendor such as Airsage in the U.S. or Teralytics internationally.

GPS location data offer higher resolution and more frequent sampling rates than cellular data, which makes them especially useful for route information, detailed trip characteristics, and modal recognition. Early GPS-based data, used to mainly augment or replace trip diaries, required specialized devices. Smartphones now let data be collected via apps like CycleTracks, developed by the San Francisco County Transportation Authority, and Strava, used primarily for fitness tracking. Many vehicles now come equipped with on-board GPS units, used for navigation, on-board assistance, and commercial vehicle tracking. One third-party provider, StreetLight Data, now compiles GPS data from a variety sources, including in-vehicle units and

---

[1] Richard J. Lee, Ipek N. Sener, and James A. Mullins, "An Evaluation of Emerging Data Collection Technologies for Travel Demand Modeling: From Research to Practice," *Transportation Letters* 8, no. 4 (2016): 181–93, doi:10.1080/19427867.2015.1106787.

mobile apps, to produce trip metrics for origin-destination studies, route mapping, and other uses.

Location-Based Services (LBS) let software developers incorporate user location information into mobile apps. The associated data is like cellular location data in that the sample size is large and a signal can be traced over long periods. The spatial precision, which depends on Wi-Fi and assisted GPS, is higher than cellular location data but usually not as high as GPS location data. StreetLight Data recently acquired this data, in addition to their navigation-GPS data.

Many data providers analyze the raw data and run their own proprietary algorithms to identify and classify trips and determine trip characteristics. The results are typically reported in tables with geographic identifiers that can be used in GIS. Some leading data providers are:

- AirSage, specializing in cellular data (U.S. only).
- Teralytics, specializing in cellular data (international).
- StreetLight Data, specializing in GPS and LBS data.

	Cellular	LBS	GPS
Spatial precision	~200-1,000 meters	~5-50 meters	~5 meters
Ping frequency	15+ minutes	Variable	<1 minute
Route tracking	Imprecise	Imprecise	Yes
Travel modes	Automobile; transit possible with secondary data		Automobile, walking, and biking
Sample size (varies by location)	~10-30%	>10%	~1-4% (personal) ~10-12% (commercial)
Providers	AirSage, Teralytics	StreetLight Data, Cuebiq	StreetLight Data, INRIX

The *Connecting Sacramento* study incorporates the following data:

- Transit trip characteristics from cellular location data combined with General Transit Specific Feed (GTFS), provided by Teralytics.
- Vehicle trip characteristics from GPS data, provided by StreetLight Data.
- Preliminary pedestrian trip characteristics from GPS data, provided by StreetLight Data.

Data validation

For this study, Teralytics data is reported as light rail transit (LRT) trips and non-LRT trips. The data represents 43,100 LRT trips during a typical weekday in March 2015, compared to 43,001 average daily trips during the same quarter, as reported by Sacramento Regional Transit.

Shown below is the distribution of Teralytics LRT trips by time of day during a typical weekday and weekend, which are consistent with our general understanding of transit trip-making patterns and with the LRT's hours of operation.
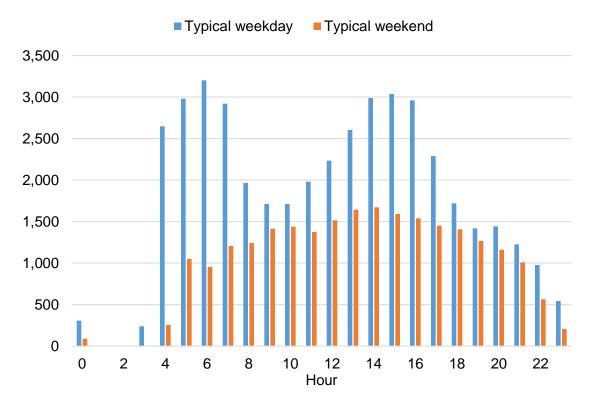
Figure 1. LRT trips by time of day for a typical week in March 2015 (Teralytics)

StreetLight Data reports trip volumes using a *StreetLight Index*, which is meant to be consistent across geographies and over time but does not represent actual trip numbers. For this study, those values are compared to average daily traffic (ADT) counts on 30 highway segments, provided by Caltrans. There is a strong relationship between ADT from non-trucks and the corresponding StreetLight Index for personal vehicles, as shown below. Based on this relationship, one StreetLight Index point is equivalent to approximately 0.85 actual trips.
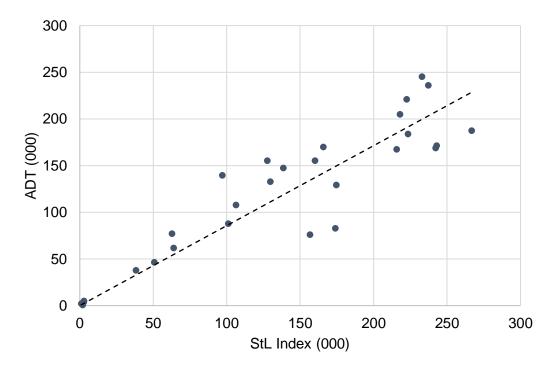
Figure 2. Relationship between observed traffic counts (ADT) and reported trip index values from StreetLight Data (StL Index) on 30 highway segments

The main reason this calibration isn't done automatically is the lack of reliable, nationwide traffic volume data. However, StreetLight Data has begun incorporating traffic counts into their data products to be able to report ADT values.

Applications in Sacramento

For the *Connecting Sacramento* study, the region was divided into 250 traffic analysis zones (TAZs). Zones near light rail stations are more granular (including small zones encompassing park and ride lots) and more distant zones are coarser. Most analyses focus on specific transit catchment areas along the light rail system. Trips contained within these catchment areas (Figure 3) are considered potential transit trips, particularly those made by personal vehicle. The aim of this study was to understand where those personal vehicle trips occur and identify opportunities for transit to capture some of them. Some influential factors include: the ability to access transit stations by walking or biking, the cost and availability of parking, trip purpose (inferred), and demographic characteristics of the origin zones.
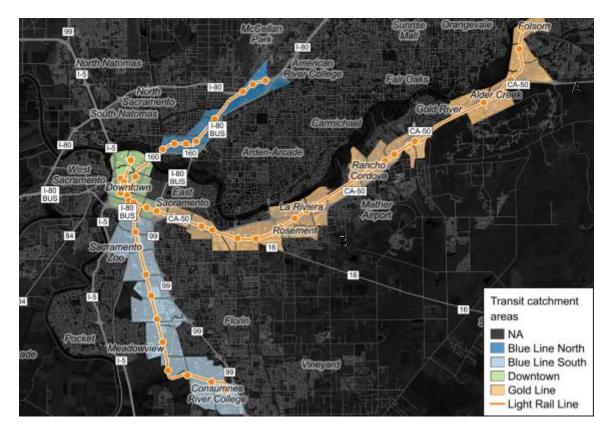
Figure 3. Transit catchment areas and case study locations

The study relied on several key metrics for each TAZ, derived from the various trip-making data:

- Vehicle trip generation (StreetLight Data).
- Vehicle trip destinations from a given TAZ (StreetLight Data).
- Light rail trip generation (Teralytics).
- Light rail trip destinations from a given TAZ (Teralytics).
- Light rail mode share (Teralytics).

Vehicle trip generation and light rail mode share are two key metrics used to identify opportunities to increase transit ridership. For example, on a typical weekday, the neighborhood near Iron Point station (A in Figure 3) generates roughly 7,400 vehicle trips that end somewhere else along light rail corridor – the highest trip generation outside of Downtown. In another example, only 1.4% of trips beginning just south of Butterfield station (B in Figure 3) are made by light rail, compared to 5% in other nearby areas.

Further investigation gives more insight. For example, of the 7,400 vehicle trips beginning near Iron Point station each weekday, roughly 2,800 end within two to three stations and another 1,700 end elsewhere along the Gold Line, including Downtown (Figure 4). Those short trips ending nearby might be more difficult to shift to transit unless service is frequent, access to the stations is excellent, and parking near the stations is managed appropriately.
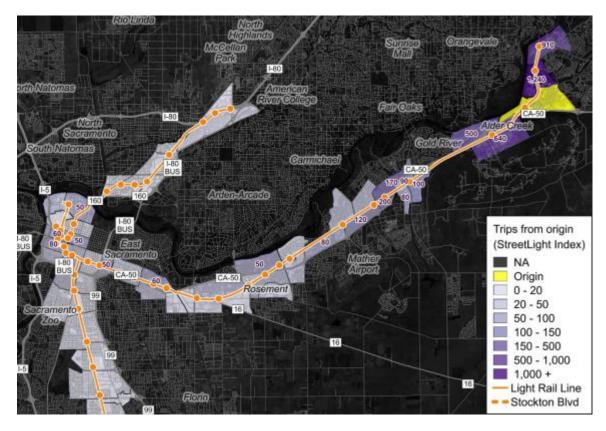
Figure 4. Vehicle trip destinations, beginning near Iron Point station (StreetLight Data)

The area with exceptionally low light rail use south of Butterfield station (Figure 5) includes several large office buildings served by abundant parking, which likely encourages driving and makes walking to the station less appealing.

Figure 5. Neighborhood south of Butterfield Station (1.4% light rail mode share)

First- and last-mile connections

Trip-making data can also help better understand how people access light rail stations – the so-called first- and last-mile connections to transit. By creating zones around park-and-ride lots, the data may reveal important trip-making patterns of people driving to and from the stations. For example, the Meadowview station was the southernmost stop on the Blue Line for much of the *Connecting Sacramento* study period. Analysis shows where many of the trips leaving its 700-space parking lot end (Figure 6). Some of those trips, southeast toward Cosumnes River College, are now better served by the Blue Line extension. Other trips ending one or two miles east of the station, however, could be made by other modes like walking and biking if reasonable connections were provided. Improved options like these, paired with the right parking price adjustments, could be helpful in managing parking demand, particularly if a lot is near capacity.
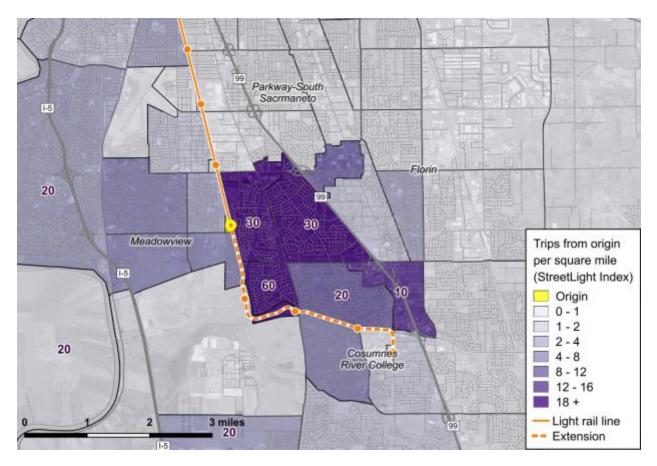
Figure 6. Vehicle trip destinations from Meadowview station (StreetLight Data)

StreetLight Data also provided pedestrian trip data for the *Connecting Sacramento* study, to analyze trip-making patterns of people walking to and from the stations. For example, analysis of pedestrian trips to and from the Zinfandel station platform reveal several things (Figure 7):

- Most walking trips (63% in total) come from areas just north and east of the station, where there are large shopping centers.
- 19% of trips come from residential areas to the south and west.
- Only 5% of trips come from residential areas to the northeast (more use Cordova Town Center station, according to the data).
- 15% of trips come from a large cluster of office buildings southeast of the station across the Lincoln Highway (US-50), via a single access point across the freeway at the Zinfandel Drive interchange (suggesting a need for more or improved crossings).
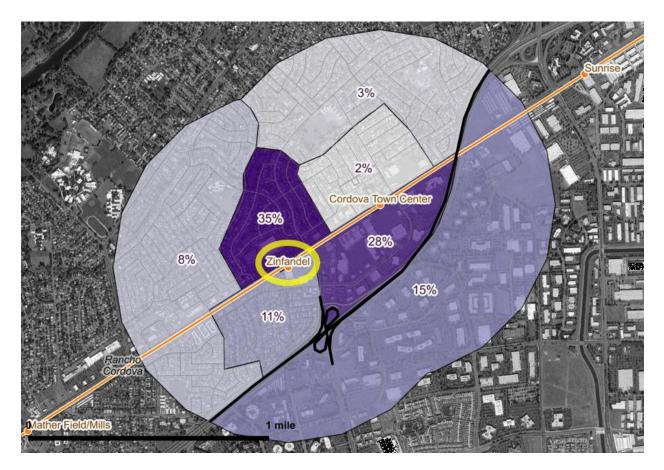
Figure 7. Origins and destinations of walking trips to and from Zinfandel station (StreetLight Data)

These pedestrian data, derived from GPS-enabled smartphone apps, are in a test phase, but the results are promising. StreetLight Data's modal recognition algorithms perform well, but need further validation. The above analysis also fails to include trips of less than 500 meters, which is a temporary artifact of the vehicle trip classification methods from which this approach is derived. StreetLight Data expects to improve and commercially launch pedestrian metrics in fall 2017.

Lessons learned

Trip-making data has many transportation-related applications extending well beyond those explored in *Connecting Sacramento*. Knowing where to start can be challenging. The following guidance, based on lessons learned from working with the data, may be helpful for those just considering using trip-making data to inform decision-making.

1. Familiarize yourself with and consider a range of data sources.

It can be hard to approach big data – or any other data sources – and know how to use them without clear general knowledge of what is available. Cellular location data might be sufficient for getting regional flow patterns or origin-destination matrices, while route flows and trip characteristics might require GPS data. More specific requests, such as transit trip recognition, might require a provider willing to work with outside data sources, such as GTFS. Having this knowledge, you can avoid spending resources on data that don't truly meet your needs. The descriptions in this report offer some guidance and StreetLight Data provides two useful resources:

- [Tools for Collecting Travel Behavior Data](#), which describes and compares the different sources of data.
- [Big Data for Transportation](#), which describes applications of the data using real world examples.

While trip-making data and other big data can greatly improve our understanding of the world, they can't answer every question or replace other data sources entirely. For example, trip-making data can perform many functions of travel demand models without the need for travel surveys, land use data, traffic counts, and painstaking calibration. However, they can't necessarily produce forecasts and they don't include detailed traveler attributes for individual trips like surveys do (those attributes are generalized from census data). For some data needs – like spot speeds or event day trip counts – traditional field studies could be cheaper, more immediate, or more reliable than trip-making data.

Ultimately, big data add to the menu of options, replacing some items and complementing others.

2. Ask a specific question to address a planning need.

One major advantage of the trip-making data described here – its quantity – can also be a serious impediment. It may be tempting to begin by asking for as much information as possible, but this can make interpretation more difficult and drive up costs. Instead, have a clear question in mind. For example, at one point this report asks: Where do all the cars parked at the Meadowview lot during on a typical weekday in 2015 come from? This question points to the travel mode, the spatial and temporal resolution, the type of analysis, and ultimately the right source of data to use.

This focus will also make interpreting the data much simpler. When it comes to data requests, think of the simplest path to a useful answer. Data providers typically offer ways to dig deeper on an as-needed basis. For example, the StreetLight InSight® interface lets subscribers create simple analysis zones in an interactive map or upload more complex zones from GIS, specify a data source and trip characteristics of interest, and download data tables on-the-fly. Teralytics, after running its own analysis of the data, works with clients to decide on the relevant metrics to report.

3. Enlist the right people help interpret the data.

Data providers do most of the heavy lifting – analyzing the raw data, cleaning it, and offering it in a useful format – but using big data isn't as easy as pushing a button. Trip-making data are often provided in ways that require some basic analysis and interpretation. Even in the simplest applications, the best use of the data most likely requires basic GIS skills, spreadsheet-style data analysis, and knowledge of transportation planning principles to interpret and apply the findings. Many planning groups are proficient in these areas, but any gaps in skills or knowledge should be filled before purchasing large amounts of data.

4. Aggregate big data appropriately to leverage its full potential.

Trip-making data is typically more reliable and compelling when it's aggregated over larger areas or longer periods of time to achieve a larger sample size. Each additional detail, such as trip purpose, time of day, or demographic characteristics, demands a larger sample size. Working with the data may therefore require tradeoffs; smaller analysis zones may come at the cost of longer analysis periods or fewer trip details. For better results, always think about which features of the data can be aggregated while still providing meaningful results.

For example, if your initial question pertains to a small area, such as a neighborhood or large parking lot, the analysis will most likely require aggregating over a long period of time. A before and after study for a recent project, however, may require larger analysis zones to offset the short period of time from which data can be sampled.

More information on the *Connecting Sacramento* study and emerging uses of big data can be found at www.ssti.us.