TECHNICAL MEMORANDUM

State of Emerging Mobility Big Data Sources and its Applications

Task 1: Evaluate Mobility Datasets

SUMA
Support for Urban Mobility Analysis

A technical memorandum to

Support for Urban Mobility Analyses (SUMA)
FHWA Pooled Fund Study

Authors:
Gargi Singh
Vijayaraghavan Sivaraman
Ed Hard

August 2022

Submitted by the

Texas A&M Transportation Institute

Table of Contents

EXECUTIVE SUMMARY

The 2019 SUMA technical memorandum on Tools and Best Practices for Using Passive Origin-Destination (1) Data assessed three primary location estimation technologies: Global Positioning System (GPS), Location-Based Services (LBS), and Cellular. This memo serves as an update to the above mentioned by focusing on big data-based mobility products and their applications produced from such sources. The insights and metrics reported by the mobility products are affected by the diversity of data sources and the extent to which this data is processed and packaged. Therefore, as the market continues to evolve and data vendors offer amalgamated data products from multiple data source technologies, it is necessary to understand various data products in the market, their characteristics, and their potential applications.

This memo broadly categorizes some widely used data sources across one or more vendors, their products, features, and typical transportation applications. They are sub-divided by the extent to which the source data is processed and synthesized with one or more secondary data sources as:

- **Trip Trace (Raw) Data**: This data category is the most granular data available in the market. Depending on the data source technology, this type of data could either yield a location estimate (trip end) of the activity or an entire trajectory across locations of activities.

- **Transformed Data** are aggregated trip trace data and expanded to the population. This data is often sampled, cleaned, and aggregated to produce data products that could be used for further analysis. The data is aggregated either by time or space. E.g., Origin-Destination data and Point of Interest (POI) data.

- **Modeled Data** are modeled trip trace data that is enriched with other non-passive data sources such as demographic data, land use data, credit card data, etc. It is then processed to develop synthetic travel diaries to subsequently model a region's daily travel patterns analogous to an activity-based model and assigned to the regional transportation network.

Most of these products are delivered either through a cloud service as pre-packaged custom raw outputs or via web-based platforms. The latter, web-based platforms allow an analyst to perform a custom query online that can be visualized and downloaded for further offline analysis and reporting. The web platforms can be broadly categorized into three types –

- **Data provider platforms** offer one-time purchases and periodic summaries accessible via their platform as part of a subscription. They offer both trip trace data and transformed data through their platform. It also is a good alternative for users who might not have the time, technical skills, or necessary cloud infrastructure to handle raw trip trace data.

- **Third party platforms** serve as hosts to an array of big data sources as well as a user's custom datasets. These platforms allow an analyst to visualize and summarize the big data hosted on such platforms. Most of these platforms offer a data downloader feature that allows bulk data downloads of raw data that can subsequently be analyzed in-house.

- **Service Provider Platforms** primarily offer access to transformed data produced using their proprietary methods. The data output provided is either modeled or includes additional inferences such as mode or demographic attributes using data layering or machine learning techniques. Some of these platforms allow users to upload aggregation zones or choose road segments from the interactive map on the platform to conduct select link or select zone analysis.

As is apparent, the transportation data market offers myriad options for consuming passive data. No one size fits all, and each data product offers a unique set of strengths and challenges. The choice of data product depends on several factors: user control/access and flexibility to analyze, resources required, transportation applications, etc. The trip trace/raw data is the most granular data available in the market, making it the most flexible data to work with in developing in-house applications, sample selection/filtering, custom analysis, and data expansion. However, due to its massive size, it also comes with the need to have in-house big data expertise and additional data management and processing costs. Alternatively, transformed data, due to being aggregated, does not come with the added cost of data management and processing; but it is less flexible in that it does not allow for sample selection, assessing bias, developing expansion factors, and has restricted use. Lastly, modeled data is ready to use off the shelf and requires minimum to no processing. However, it is highly constrained because it is a modeled representation of a region. The consumer might only be able to use it for validation or assess base year or other scenarios. Therefore, users need to understand what each data product brings to the table to make an informed choice for their intended application and use cases.

INTRODUCTION

Over the past decade, passively collected big data for transportation analysis has become pervasive and continues to grow exponentially. Predominately, it is sourced from various sources such as smartphone applications, fleet navigation systems, transit smart card infrastructure, or connected car applications. This data comes with higher velocity, variety, and volume resulting in an almost real-time representation of population mobility patterns. It has led the transportation data marketplace to flourish, with vendors delivering an array of transportation data products developed from these big data sources. As consumers, it is crucial to understand the types of data and products available in the market and how they can serve a transportation agency's analytical needs. However, in this growing data marketplace, agencies often are faced with the dilemma of choosing from a multitude of data vendors, making it an onerous and resource-intensive task.

In 2019, a technical memorandum was delivered as part of SUMA (1) that discussed sources, methods, and technologies producing big mobility data and its derived products. It primarily focused on cellular, LBS, and GPS collection technologies and illustrated what the resulting data represents and how travel is estimated from these data sources. This memo focuses on the types of data products available in the consumer market, how they are distinct from each other, and their potential applications for transportation planning and traffic operation analysis. Specifically, the objective is to serve as a reference to identify the type of data/data product needed for an application. This technical memorandum pursues this by classifying big data products for transportation mobility analysis into three main categories. The categories are divided in terms of their granularity, source, and the extent to which it has been synthesized as:

- Trip Trace (Raw) Data
- Transformed Data
- Modeled Data

The above listed data and(or) their derived products are available through vendors either for bulk download through their cloud services or accessible via its platform; the latter is more likely as a data product after some transformation and aggregation. Also, each data product and delivery method have distinct advantages and challenges depending on the user and type of application.

This memo provides an overview of characteristics for each of the above data product types to assist consumers in making an informed decision when purchasing and applying transportation data products. It provides the details on the following elements of the data available in the abovementioned three categories:

- What the data represents,
- Data technology associated with the data product,
- Strength and challenges with utilizing the data product,
- Data vendors and schemas, and
- Best practice examples from several DOTs and MPOs from around the country.

DATA TECHNOLOGY – WHAT'S NEW?

The 2019 SUMA technical memorandum on Tools and Best Practices for Using Passive Origin-Destination Data assessed three primary location estimation technologies: Global Positioning System (GPS), Location-Based Services (LBS), and Cellular. Since the last memo, the market has evolved with vendors bringing to the market products developed using data from multiple big data sources. Due to this consolidation, it is necessary to understand various data products offered in the market, their characteristics, and their potential applications. Further, this market is in a constant state of flux with emerging big data sources. For instance, many cellular data providers who relied solely on cellular tower-based location estimation methods have gradually replaced cellular data sources with more precise ones such as LBS technologies (smartphone applications) and lately are transitioning into GPS data (e.g., connected car data). This memo serves as an update and, more specifically, focuses on big data-based mobility products and their applications for the transportation market.

As it offers essential communication services to the national subscriber base, data sourced from telecom carriers has always had a relatively larger representative sample. However, its estimated location of physical activity has been less precise due to its dependence on cell phone tower triangulation. This makes it less amenable to understanding travel at finer resolutions or geographies (e.g., pedestrian movements, multi-modal travel, etc.). LBS data that currently dominates this market is found to be more precise as it can be estimated either using telecom towers, GPS, or Wi-Fi services. This allows for better location precision but is dependent on the smartphone application design and the availability of the location estimation service across space and time. More importantly, unlike telecom, the LBS user base is primarily unknown, and their usage behavior may not necessarily be representative of a typical household travel behavior. Thus, such sample shows a trade-off between location accuracy and daily travel behavior representativeness.

Lately, vendors have also been transitioning their data source to that originating from connected car data. This data source estimates locations using in-vehicle infrastructure such as navigation GPS, and thus tends to be more persistent and have at most precision. However, this source, similar to LBS, is often confined to representing a specific segment of the population or fleet of vehicles, such as belonging to a particular brand or performing specific services. Furthermore, it primarily represents auto travel, thereby excluding the potential to capture non-auto travel modes.

As to the future, the telecom sector is undergoing significant upgrades with the emergence of the fifth generation (5G) mobile broadband. Due to its lower latency and greater bandwidth, we could potentially witness a dramatic increase in very granular and real-time data. This data would be sourced from vehicles and connected devices, such as smartphones, traffic signals, roadway signs, etc. (see Table 1 for more details). In addition, its interactions with vehicles, a phenomenon known as vehicle-to-everything(V2X), could aid in the transportation planning and operations sector (2), which could drastically advance the transportation industry.

Table 1: V2X subcategories and their prospective applications

V2X SUBCATEGORIES	DESCRIPTION	SERVICES EXAMPLES
VEHICLE-TO-VEHICLE (V2V)	Direct communication between vehicles.	1) **Intersection collision warning** - potential crash risk at an intersection 2) **Local hazard warning -** sharing of information from one vehicle to other(s) of abnormal activity on a road that could be a potential risk
VEHICLE-TO-INFRASTRUCTURE (V2I)	Communication between a vehicle and road infrastructure.	1) **Road work ahead warning –** Information on roadwork and potential lane closures due to it 2) **Traffic jam ahead warning** 3) **Green light optimal speed advisory (GLOSA)** to allow vehicles to pass intersections without stopping
VEHICLE-TO-PEDESTRIANS (V2P)	Communications between vehicles and pedestrian devices	**1) Speeding Car Alert** - Alert devices when a car is approaching at an unsafe speed. **2) Automatic Stop** - Enable cars to stop at traffic crossings automatically.
VEHICLE-TO-DEVICE (V2D)	Communication between vehicles and non-V2V enabled vehicles and cyclists	
VEHICLE-TO-NETWORK (V2N) **OR** **VEHICLE-TO-CLOUD (V2C)**	Communications with the broadband cellular mobile network for data exchange.	**Shared Mobility** - Drivers' preferences, such as the seat and (or) mirror position or music preferences, are saved in the cloud and adjusted when carsharing.
VEHICLE-TO-GRID (V2G) **OR** **VEHICLE-TO-HOME (V2H)** **OR** **VEHICLE-TO-BUILDING (V2B)**	Communication between power grid and vehicle battery.	**Dynamic Load Management (DLM)**[1] – Enable Electric vehicles (EVs) to charge anywhere, irrespective of location or surroundings. - Ability to push back power from the battery

Thus, while telecom technology is not as widely used in transportation at present, it is likely to return as a significant source for transportation analysis as the fifth generation (5G) mobile broadband technology becomes ubiquitous. Specifically, the introduction of 5G microcells (3) could result in more precise location estimates, yielding a much larger representative sample of travel with better location precision

[1] Vehicle-to-grid (v2g) is a technology that has the power to transform the energy system - https://www.virta.global/vehicle-to-grid-v2g

than what had been available from this source. However, as the five Vs of big data, namely, Volume, Velocity, Veracity, Variety, and Value, increase with 5G, the privacy concerns grow too:

1. The data would need to be stored safely to avoid data breaches.
2. Organizations need to establish procedures to mask or aggregate the data as the data gets more granular to ensure privacy.
3. The privacy laws might impact affect the sample size of data.

5G technology and privacy concerns will continue to shape the digital data landscape and, consequently, its utility for transportation planning and traffic operations analysis.

PASSIVE MOBILITY DATA PRODUCTS

As the previous section shows, transportation data products from diverse big data sources in the market have progressively evolved and will continue to do so. It is not only influenced by the diversity of data sources, but also by how this data is processed and packaged as a product. These products are broadly delivered through cloud computing resources either as raw data with minimal aggregation or as pre-packaged custom outputs after several transformations, often through web-based platforms. This allows an analyst to perform custom queries/aggregations and then download and visualize them offline without needing big data analytical tools. The transformed data products are the most widely accessed and available in the transportation marketplace. There have been advances where the modeled outputs are made available through cloud platforms as data products. Each has its tradeoffs, mainly between investing time and building resources in-house to produce custom analytics versus using off-the-shelf analysis results from vendors.

Sample Selection	**Yes**	**No**	**No**
Custom Analytics	Yes	No	No
TDM Input	Yes	Yes (Restricted Vendor Format | Sample)	No (Already Modeled)
Validation	Yes	Yes	Yes
Staffing Skills	High	Medium	Low
Inhouse Infrastructure	High	Medium	Low
Applications	Several	Restricted to Vendor Data	Low: Modeled Results
Subscription	One Time Purchase	One Time Purchase | Subscription	Possible Subscription | Evolving
Agency Data Architecture	Needed – Big Data Management	Not Needed	Not Needed

Figure 1: Mobility Big Data Sources and Delivery Platforms

This section broadly categorizes some of the widely used data sources, their products, their features, and typical transportation applications across categories based on the state of the data as either being raw, transformed, or modeled. Figure 1 illustrates the nature of data products, associated need for in-house skills and computational infrastructure, purchase frequency, etc. In terms of their representation, these data sources provide mobility insights on the passenger (non-commercial) and (or) commercial travel for a region of interest defined by potential customers. Following Figure 1, this section further summarizes mobility data products available in the market as trip trace (raw) data, transformed data, and modeled data with references to select vendor schemas. The vendors referenced in this document are not necessarily the only ones in the market but are prominent in this space and are used to discuss potential applications.

Trip Trace (Raw) Data

Trip trace[2] data is the most disaggregated form of passively collected crowdsourced data available to purchase for transportation/mobility analysis. Trip trace data depending on its location estimation technology and application (example: in-vehicle navigation, smartphone application), could either yield a location estimate (trip end) of the activity or an entire trajectory across locations of activities. Location technologies ranging from GPS, Cellular networks to Wi-Fi could significantly influence the accuracy of an activity's estimated location. The location for trip trace data originating from certain applications (example: connected cars, fleet navigation systems, etc.) is primarily estimated using GPS. In contrast, trip trace data originating from LBS applications (see 4 for more details) could be estimated from a mix of location estimation technologies ranging from GPS, cellular networks, or Wi-Fi (5). The latter source of mobility data originates from smartphone applications that use a combination of GPS, cellular network, and Wi-Fi and is often referred to as location-based service (LBS) data. This means that some samples could represent trajectories (example: travel navigation apps using GPS), and others could represent activity at a location (example: advertising apps using Wi-Fi or Bluetooth within a store or a mall). Thus, it is essential to understand the sub-samples to differentiate the samples representing travel trajectories compared to those representing activity at a location.

In contrast, those derived from embedded GPS devices such as commercial fleets providing services or connected cars primarily capture travel trajectory attributes such as location and timestamp logs. They also tend to provide an array of attributes besides location and time stamps. For instance, embedded GPS devices in commercial vehicle fleets provide vehicle attributes such as weight class and likely function (see INRIX schema for more details)**.** This also allows one to sample segments based on broader weight classes and estimate their potential underlying use or function, as shown in Table 2. These attributes could be valuable for transportation planning, such as understanding the representative sample of the population of such vehicles in a region and undertaking external and local commercial vehicle impacts. In contrast, the data originating from connected vehicles, using Internet of Things (IoT) sensors, generate additional vehicle attributes such as operating status and usage of vehicle components (see Wejo schema for more details). The IoT sensors from connected cars generate data on

[2] A trip trace is a set of location logs with timestamps passively gathered from device using one or more source of location estimation technologies (example: GPS, Cellular Network, and or Wi-Fi signals).

the usage and working conditions of components of a connected vehicle, such as seat belt usage, brakes, wipers, etc. These attributes are valuable to traffic operations and safety analysis. It is to be noted that though some vendors primarily sell non-commercial vehicles at this time, others offer for all connected vehicles (e.g., commercial fleet, rental cars, etc.). These data originate from select vehicle brands and models (primarily more recent ones with connected car capabilities). This information would be important to consider if there is interest to utilize such data to understand environmental impacts of electrification, and account for bias or representativeness in the sample, etc.

Table 2: INRIX Weight classification and their corresponding provider profiles

INRIX Weight Class	**Definition**	**Provider Type**	**Provider Driving Profile**	**Corresponding FHWA Vehicle Classification**
Weight Class 1	< 14000 lbs	- consumer - fleet	- Consumer Vehicle - Field Service / Local Delivery Fleets - Taxi / Shuttle / Town Car Service Fleets	Class 1: Motorcycle Class 2: Passenger Cars Class 3: Four tire, single unit
Weight Class 2	14000 – 26000 lbs	- fleet	- Field Service / Local Delivery Fleets - For Hire/Private Trucking Fleet	Class 4: Buses Class 5: Two axle, six tire, single unit Class 6: Three axle, single unit
Weight Class 3	> 26000 lbs	- fleet	- Field Service / Local Delivery Fleets - For Hire/Private Trucking Fleet	Class 7: Four or more axle, single unit Class 8: Four or less axle, single trailer Class 9: 5-Axle tractor semitrailer Class 10: Six or more axle single trailer Class 11: Five or less axle, multi trailer Class 12: Six axle, multi-trailer Class 13: Seven or more axle, multi-trailer

Overall, trip trace data obtained from non-cellular devices such as embedded GPS or connected cars produce the most detailed spatio-temporal log of mobility patterns as those devices are constantly pinging. In contrast, LBS data derived from smartphone devices can be sparse depending upon the devices' application configuration, usage, and the location technology accessed during their usage. However, with the advent of 5G technology, the cellular network could also significantly gain location precision.

Irrespective of the location estimation method and probe type, trip trace data is likely to involve high volume and velocity. If one must work with such data, one must have the computational infrastructure and technical capabilities to work with it. Increasingly, this is becoming feasible with growing cloud storage and computing services and could likely be accomplished through in-house training based on available resources. It also allows the flexibility to undertake multiple in-house analysis tasks with a clear understanding of the value and challenges of working with such data.

Transformed Data

Transformed mobility data is trip trace data aggregated and sometimes enriched using other data sources (e.g., U.S Census and AADT). This data is often sampled, cleaned, and aggregated using proprietary methods by vendors to produce data products that could be used for further analysis. As part of the transformation, the data is also sampled to exclude or filter out trip trace data based on vendor-defined criteria and logic in their algorithms. For instance, the filter process might exclude location logs that may not qualify as an activity location (trip end) based on the criteria designed by the data providers. These criteria are part of the vendor's proprietary algorithm that, in this example, may include dwell time, activity cluster, and its contextualization using secondary sources (e.g., land use). The resulting filtered data is then aggregated and packaged as is or expanded to produce mobility products representing travel for a region of interest. These products range from trip matrices (representing travel between zones within a region of interest) to analytics reporting visits to activity centers such as a point of interest – parks, malls to parking events. In some instances, the data vendors also provide derivative products such as a subset of the trip matrices for those seen passing through a particular section of the transportation network (roadway) or seen in a specific region – widely referred to as select link or select zone analysis.

The advantage of transformed data, due to it being aggregated, is that it does not have the same volume and velocity as trip trace data. This makes it convenient to process, analyze and visualize the transformed data offline using commonly available tools such as Microsoft Excel®. In essence, transformed data, in most cases, is no more big data and does not require high-performance computing resources such as a cloud service. However, the data also has its tradeoff as the transformed data loses some basic information that raw data might possess, which is essential to understanding the representative sample and its biases. Also, once trips are aggregated to produce trip matrices and aggregated to a zone, it is no longer possible to link a trip to its point of interest, thereby making it difficult to infer the likely trip purpose and other characteristics.

Further, the above aggregation also involves expansion, ranging from expanding the sample trips to representative samples' estimated home locations' resident population or roadway vehicle counts. The former is often a straightforward expansion that relies on the inverse sampling rate (or weight) of sample devices' estimated home location (census geography), such as that employed by AirSage and SafeGraph (see SafeGraph schema for more details). In contrast, the expansion to vehicle counts is a bit more involved, requiring information on local vehicular traffic volume, and is undertaken using an iterative proportional fitting procedure (6)**.** This procedure is like the Origin-Destination Matrix Estimation (ODME) procedure widely employed in travel demand analysis (see 7 for more details)**.** Thus, one might want to understand the underlying approaches and the data sources used in transformed data products before using them as input or as a source for validation.

The above transformation to the raw data means that one must rely on transformed data to use it as is, because vendors' proprietary methods may not provide the transparency required to understand the biases in the data and appropriately adjust and compensate for it, which is possible with traditional sources such as travel surveys. Thus, when choosing data to purchase, it would help to work closely with the vendors to understand their attributes and level of aggregation. This could provide additional

insights into the data product and help the user make an informed choice. Also, it could have significant implications regarding how to further use and transform the data, such as how the transformed data could be used as input for regional travel analysis. This section discusses various types of transformed data, broadly across two categories – commercial and non-commercial mobility data.

Commercial Vehicle Data - GPS

ELD Data

Electronic Logging Devices (ELD) were made mandatory for all interstate drivers of commercial vehicles as of December 2017 so that commercial drivers met the requirements for working and resting hours (8). ELDs are devices installed in trucks to track driver rest and service times and ensure compliance with federal Hours of Service (HOS) rules. It is synchronized with the engine of commercial vehicles so that they can monitor the engine operations automatically and capture data such as timestamp, location, engine hours, and vehicle miles traveled at certain intervals. In addition to trip information, it also records identification information of the driver, authenticated user, vehicle, and motor carrier. FMCSA does not require ELDs to capture vehicle performance attributes; however, some ELDs collect safety attributes such as speed and hard braking (9).

The data originating from ELD are anonymized and aggregated for privacy reasons and are delivered either as customizable zonal origin-destination matrices or geospatial vehicle activity clusters representing parking or harsh breaking events. By design, ELD data does not collect positional data continuously and has low accuracy (10); but it provides more contextual information than those available from most of the GPS probe data and connected vehicle data.

Non-Commercial Data - LBS

Point of Interest (POI) Data

Besides trip product (example: trip matrices), LBS data is also extensively used in estimating visits to Point of Interest (POI). This is generally extracted from LBS data collected by smartphone applications and aggregated spatially using catchment areas. This aggregated data product provides information on visits to a POI, which could be temporary activity areas, such as an outdoor public event, or permanent locations such as a national park or retail site.

POI data product is distinct from traditional trip data as the former focuses on visits to a POI rather than visitors' subsequent or previous trip ends. As observed in Figure 2, in addition to information on POI, it might also provide visitor attributes, such as their home location, stay duration etc. However, this information is aggregated either by space or time, and often is reported over a week or month (see SafeGraph or FourSquare schema for more details).

Figure 2: A summary of some attributes provided by SafeGraph

Home and Work Summary	Places Summary	Travel Pattern Summary	Polygon Data
•Visitor home census block group •Visitor daytime census block group	•NAICS code •Street address	•Median distance from home •Dwell time •Visits by day and hour •Unique visitors count	•Building Footprint Data

Modeled Data

The travel demand model development process typically managed by regional transportation agencies is developed by transportation consultants using travel survey data. More recently, such models are being developed by constructing synthetic travel dairies by fusing passive big data with publicly (e.g., U.S. Census) and privately (e.g, credit card transactions data) sourced data, as shown in Figure 3. These synthetic travel dairies may not be as detailed as traditional travel surveys but are found to report similar travel characteristics for a region as that obtained using traditional surveys. Furthermore, it yields a much larger and more diverse sample than traditional travel surveys.

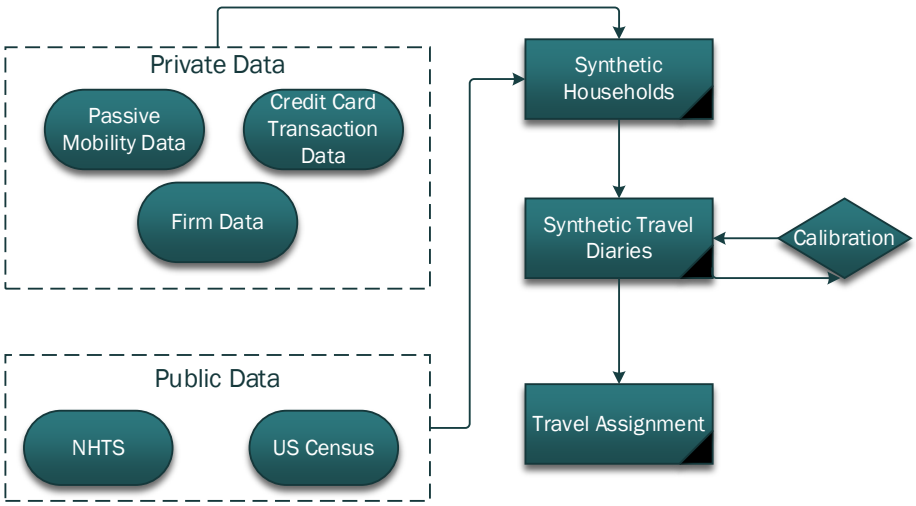


Figure 3: Modeled data (data driven model) approach

The abovementioned development of synthetic travel diaries has led to leap frogging the traditional regional travel demand model development process with big data-driven diaries. Specifically, these synthetic travel diaries are seamlessly integrated into the regional travel demand model architecture. For example, Transport Foundry (11) first conceived a big data-driven regional model as a discrete event simulation (DES) using synthetic travel diaries developed by fusing multiple big data sources, and subsequently utilizing these diaries as input to assign daily travel onto regional highway network using MATsim. They demonstrated this application for Asheville, N.C., as a case study (12) and validated it

against field traffic counts, adopting the widely used calibration/validation procedure applied in traditional four-step and activity-based models. Similarly, Replica (see Schema for more details) has developed a scalable data-driven multi-modal regional travel demand model architecture to produce base year model estimates for any metropolitan area across the U.S., utilizing passive big data to represent both passenger and commercial travel.

Several other firms, such as Bentley Systems (Streetlytics), and Cambridge Systematics (LOCUS), have either developed or are developing such big data-driven modeling platforms. The outputs from such synthetic data-driven models delivered by these consulting firms allow transportation analysts to selectively gain insights into the extent of travel by commercial and non-commercial vehicles along any given roadway link or region. Importantly, given the velocity of big data, these insights are available for the most recent time for any area across the U.S. For example, Atlanta Regional Council (ARC) used Streetlytics data (13) along with their in-house models to understand the origin-destination patterns of travelers passing through I-85 for alternate routing and the potential impact of route closures resulting from its collapse in 2017. A growing number of such applications of data-driven regional travel models are emerging across the U.S.

These developments could lead to significant turnaround times for regional agencies to evaluate and assess travel impacts under different scenarios, both for long-term infrastructure development and short-term scenarios. This could be attained through minimal investment in big data infrastructure or in-house or recruitment by remotely accessing such services on third-party cloud platforms. However, with this convenience comes the complete reliance on such platform and their methodologies with little control into the sampling process or analytical methods employed, due to the proprietary nature of these products.

Figure 4 illustrates the three data product types discussed above, their respective data source technologies, and some of the respective data vendors. Overall, if one had to choose across the three forms of big data available in the marketplace, i.e., raw, transformed, and modeled. Pursuing raw data would provide the most flexibility to sample the data, develop agency-specific data products, and undertake ad hoc analysis as such needs arise. However, this would require the agency to invest actively in developing a big data architecture, train in-house staff, and (or) acquire consulting support, widely available across cloud service providers. In contrast, if an agency is constrained in resources but want to pursue in house model development, they could choose to go with transformed data that allows them with some flexibility in terms of development of model inputs and would be able to control certain aspects of model development. Alternatively, an agency could use modeled data if the agency is comfortable with the entire data management and model development being entirely handled by third-party applications. This would also mean that they would likely not have much control over their sample or modeling procedure, which may or may not conform to the regional model development requirements. There could be some potential arrangements with consulting firms in line with the current model development practice between governing agencies and consulting firms. With either of these options, the big data-driven transportation analysis will likely add significant value to the current practice.

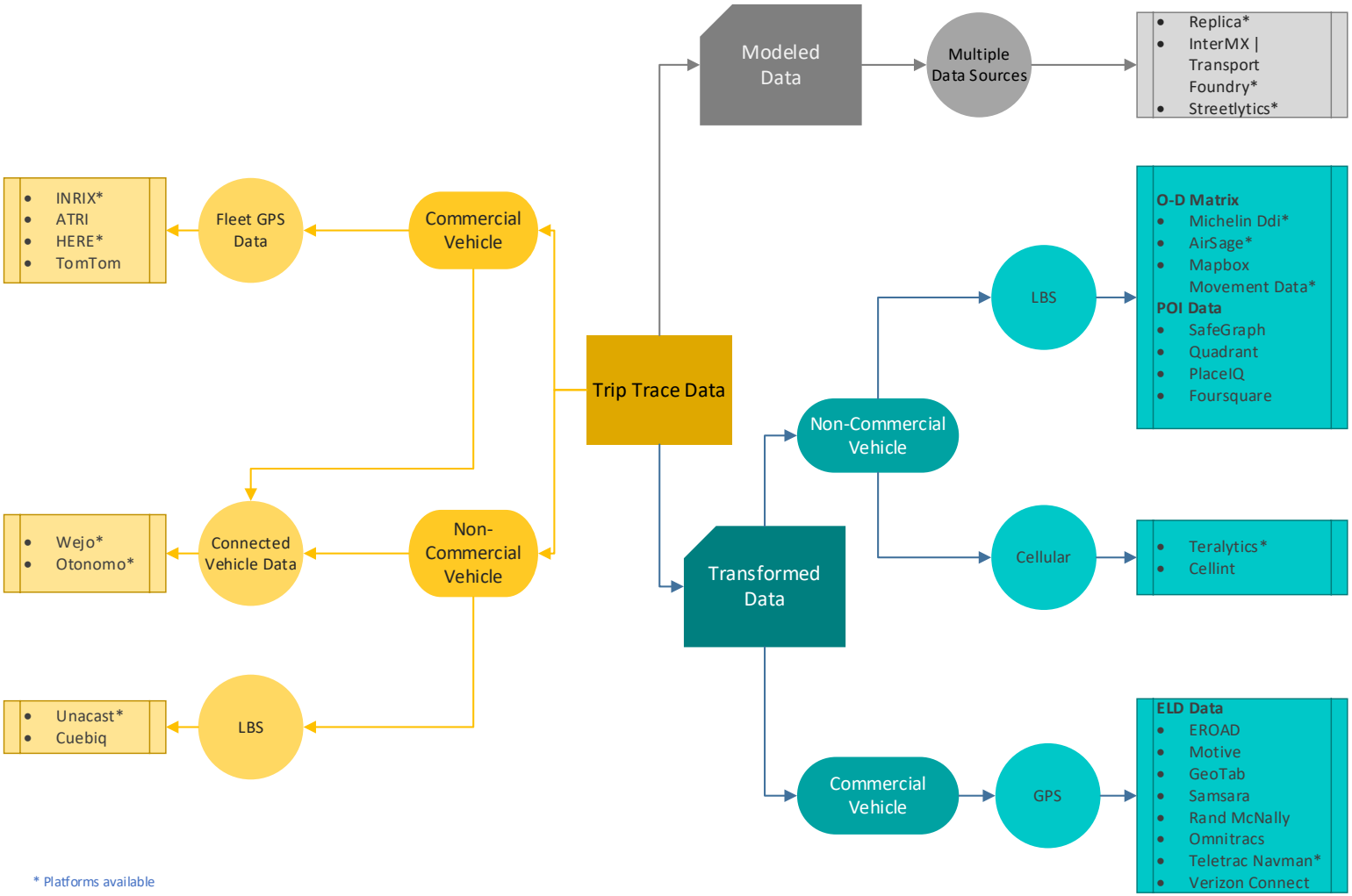


Figure 4: Data product types and their respective data sources and vendors[3]

[3] Schemas of some of the above data vendors can be accessed via https://passive-data-schemas.s3.amazonaws.com/index.html

SERVICES OR PLATFORM-BASED DATA

The above-discussed types of data products, namely, trip trace data, transformed data, and modeled data are either delivered through cloud services for bulk downloads or hosted through a web-based platform of services with additional analytical and visualization tools. The service or platforms can be broadly divided into three types: Data Provider Platforms, Third Party Platforms, and Service Provider Platforms, as shown in Figure 5.

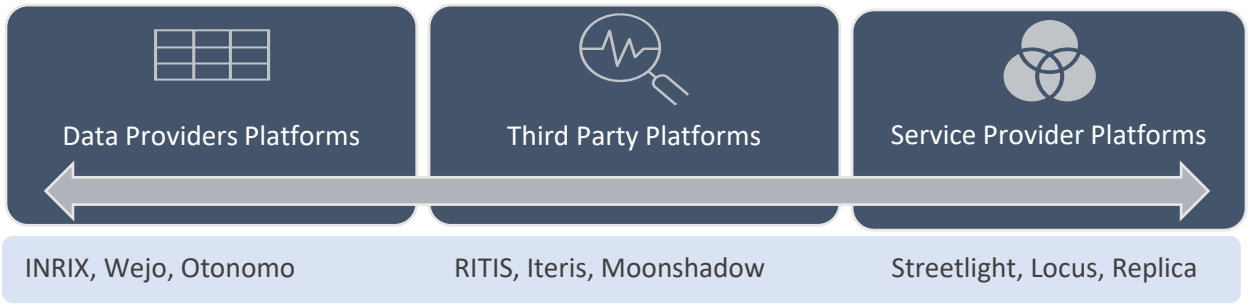


Figure 5: Types of Data Platforms and Services

Data Providers Platforms: While some data providers exclusively focus on platform development, some offer both, one-time data purchases and data summaries (transformed), access on their platform as part of a subscription. For instance, INRIX, as shown in Figure 6, offers trip trace data and transformed data through its platform. The platform serves as a good alternative for users who might not have the time, technical skills, or necessary cloud infrastructure to handle trip trace or raw data.

Granular Probe Data	Transformed Data
•Trip Data •Trajectory Data •Waypoints Data	•Trip Trends - Trip counts, lenth, duration, etc. •Trip Analytics - OD data •Volume - Traffic Count Data •Roadway Analytics •Corridor Analytics

Figure 6: INRIX - Raw products versus platform

Third party platforms such as *Moonshadow* and *Iteris ClearGuide,* serve as hosts to an array of big data sources as well as a user's custom datasets. These platforms allow an analyst to visualize and summarize the big data hosted on such platforms, predominantly used for traffic operation analysis. Most of these platforms, as listed in Table 3, also offer a data downloader feature where bulk data downloads of raw data could be also analyzed inhouse.

Some platforms allow the users to upload their data in addition to the choice of passive data available to undertake analysis on their platform. Also, some platforms, such as Iteris and RITIS, provide real-time

traffic data and historical data for comparison. These platforms primarily are geared toward supporting the transportation market's traffic analysis and operational sector.

The third category, **Service Provider Platforms**, also offers transformed data. The data is either modeled or includes additional inferences such as mode or demographic attributes using data layering or machine learning techniques. Some of these platforms, like StreetLight Data, allow users to upload aggregation zones or choose road segments from the interactive map on the platform. Once the study area selection is made, the users are offered a multitude of analysis options to choose from that are either pre-processed and available instantly or processed at the backend and delivered when ready. The platform allows an agency and (or) their consulting staff to download a subset of transformed raw data per their study requirements (e.g., select link or zone analysis in a region). In addition, at times, the platforms use multiple raw trip trace data sources to represent a complete picture of travel patterns. Like this, the service providers such as Locus purchase the most disaggregated passive data (raw trip trace data) and aggregate and package these data tailor-made to the users' needs.

Most platforms offer visualization capabilities where the aggregated or raw data can be spatially or temporally visualized. However, there might be restrictions on its use, such as it being confined to a specific project or due to data transformation, one might not be able to control or account for the potential biases in the data. Table 3 lists some of the data platforms available in the market (14).

Table 3: Platforms and their attributes

	Data Source	Data Retention	Data Downloader	Ease of Use		Historical Data
				Planning	Operations	
StreetLight	INRIX and Cuebiq*	☑	☒	4.3/5.0	3.7/5.0	☑
Moonshadow	INRIX, WEJO, MICHELIN, OTONOMO or Your Own	☒	☒	3.0/5.0	3.0/5.0	☑
Iteris ClearGuide	HERE, INRIX, WEJO	☑	☑	3.5/5.0	4.5/5.0	☑
RITIS	INRIX, HERE	☑	☑	4.3/5.0	4.5/5.0	☑
INRIX Roadway Analytics	INRIX	☑	☑	3.7/5.0	4.3/5.0	☑

*(*StreetLight, https://www.streetlightdata.com/wp-content/uploads/StreetLight-Data_Methodology-and-Data-Sources_181008.pdf)*

Recently, there has been a surge of data vendors that house multiple data sources on their websites and can be accessed without reaching out to individual data vendors. For instance, AWS marketplace (see website for more details) sells and offers free samples of a multitude of data vendors, such as SafeGraph, INRIX, AirSage, etc. Similarly, CARTO and Explorium are service providers that also house multiple public and private datasets.

CONCLUSIONS

The transportation data market offers a myriad of options for consuming passive data. No one size fits all, and each data product offers a unique set of strengths and challenges. The choice of data is a product of various factors, such as user control, resources required, applications, etc.

Trip Trace/Raw data are the most granular of data available in the market, which offers the flexibility to develop multiple in-house transportation applications and undertake multiple ad hoc analyses as needs arise out of one data purchase. Further, it allows the ability to assess and sub-sample the data for custom analysis, data expansion, detection, correction of data bias, etc. However, due to its granularity, hence large data size, it can be challenging to work with depending on the type of data source (e.g., connected vehicle data) or the size of the study area. Transportation agencies or consultants require specialized staff such as data analysts, data scientists, and data engineers to manage (procure, store and secure) and process the big data securely on the cloud, which are added costs.

Transformed Data, unlike Trip trace data, are aggregated and often expanded to the population. Due to its aggregation, these data are easier to handle and require less computational resources and minimal big data analytics skills. Users also save time and cost on analyzing and storing the data, as this data product requires minimal processing. However, their aggregated nature offers less flexibility and control over analysis, bias control, sample selection, expansion, and applications.

Modeled Data, in a transportation planning context, is an output from a regional travel demand model produced using synthetic travel diaries developed from big data sources rather than traditional travel surveys. The big mobility data forms the primary input for such travel diaries enriched with other data, such as credit card transactions, land use data, transit data, economic data, demographic data, etc. to develop daily travel and activity profile. This data product is ready to use and requires minimum to no processing and hence, does not come with added cost or time of data processing and management. However, it is highly constrained because it is a modeled representation of a region, and one might only be able to use it for validation or assess base year or other scenarios. In comparison, one could potentially consider transformed data (example: trip tables) for validation purposes and input to regional models.

The above-discussed types of data products are either delivered through cloud services for bulk downloads or hosted through a web-based platform of services with additional analytical and visualization tools. While Trip trace data are primarily delivered as bulk download, platforms are generally used to consume either transformed data or modeled data. However, platforms are also used to download trip trace data for a targeted area/analysis, which makes it easy to handle computationally due to their smaller geographic size.

Platforms are either provided by trip trace data vendors (e.g., Wejo) or third-party platforms (e.g., Iteris) and service providers (e.g., Streetlight Data) that consume trip trace/raw data. While the former platform only offers products using one data source, the latter often provides multiple data sources to allow the user to use data layering techniques to conduct analysis.

The mobility data landscape continues to change due to technological advancement and privacy concerns. The big data sources emerged as source for transportation analysis a few decades ago, originating from telecom carriers to recently being dominated by LBS-based smartphone applications. This is also in transition, with connected car data becoming the latest alternate source for understanding population mobility patterns. Soon, one might see the revival of telecom data with the implementation of 5G technology. This could potentially result in a dramatic increase in highly granular and real-time data, which might bring privacy concerns and regulations. As privacy laws and evolving technology continue to shape the big data landscape, we will likely see more amalgamation of data source technologies across data products in the market. With this perspective, it would be valuable to invest in developing in-house infrastructure for big data management as well as train and recruit staff with an appropriate mix of transportation and big data analytics skills.

APPENDIX A: Applications

Table 4: Suitability of Passive O-D Data Products by Study Type and Use

	APPLICATIONS	DATA PRODUCTS						
		Trip Trace Data			Transformed Data			Modeled Data
		Fleet GPS Data	Connected Vehicle Data	Non-Auto Sources **	OD Matrix	POI Data	ELD Data	
>>> PLANNING	Site Trip Generation	✓*	✓*	✓*	✓*	✓	✓*	✓
	Origin-Destination	✓	✓	✓	✓	X	✓	✓
	External Travel - Commercial	✓	✓	X	✓	X	✓	✓
	External Travel – Non-Commercial	✗	✓	✓	✓	X	X	✓
	Time of the Day (TOD)	✓	✓	✓	✓	✓	✓	✓
	Freight Studies	✓	✓	X	✓	✗	✓	X
	Route Analysis	✓	✓	✓	✗	✗	✗	✓
	Trip Length Distribution	✓	✓	✓	✗	✗	✗	✓
	Special Events Studies	✓	✓	✓	✓	✓	✓	✓
	Bike and Ped Studies	✗	✗	✓	✓**	✗	✗	✓
	Transit Studies	✗	✗	✓	✓**	✗	✗	✓
	Parking Need Studies/Planning	✓	✓	✗	✗	✗	✓	X
	Safety Studies	✓	✓	✗	✗	✗	✓	X
	Select Link Analysis	✓	✓	✓	✗	✗	✗	✓
	Select Zone Analysis	✓	✓	✓	✓	✗	✓	✓
<<< OPREATIONS	Traffic Impact Analysis	✓	✓	✓	✗	✗	✗	X
	Roadway impact and user fee studies	✓	✓	✓	✗	✗	✗	X
	Intersection and Corridor Analysis	✓	✓	✓	✗	✗	✗	X
	Travel Time and Speed Studies	✓	✓	✓	✗	✗	✗	X
	Performance Measures	✓	✓	✓	✗	✗	✗	X
	Toll road studies	✓	✓	✗	✗	✗	✗	✓
	Incident management	✓	✓	✓	✗	✗	✗	✓

✓* In-house/custom analysis.

** GPS data from transit agencies, Strava, and other non-auto mobility data sources

✓** If the mode is imputed using trip trace data

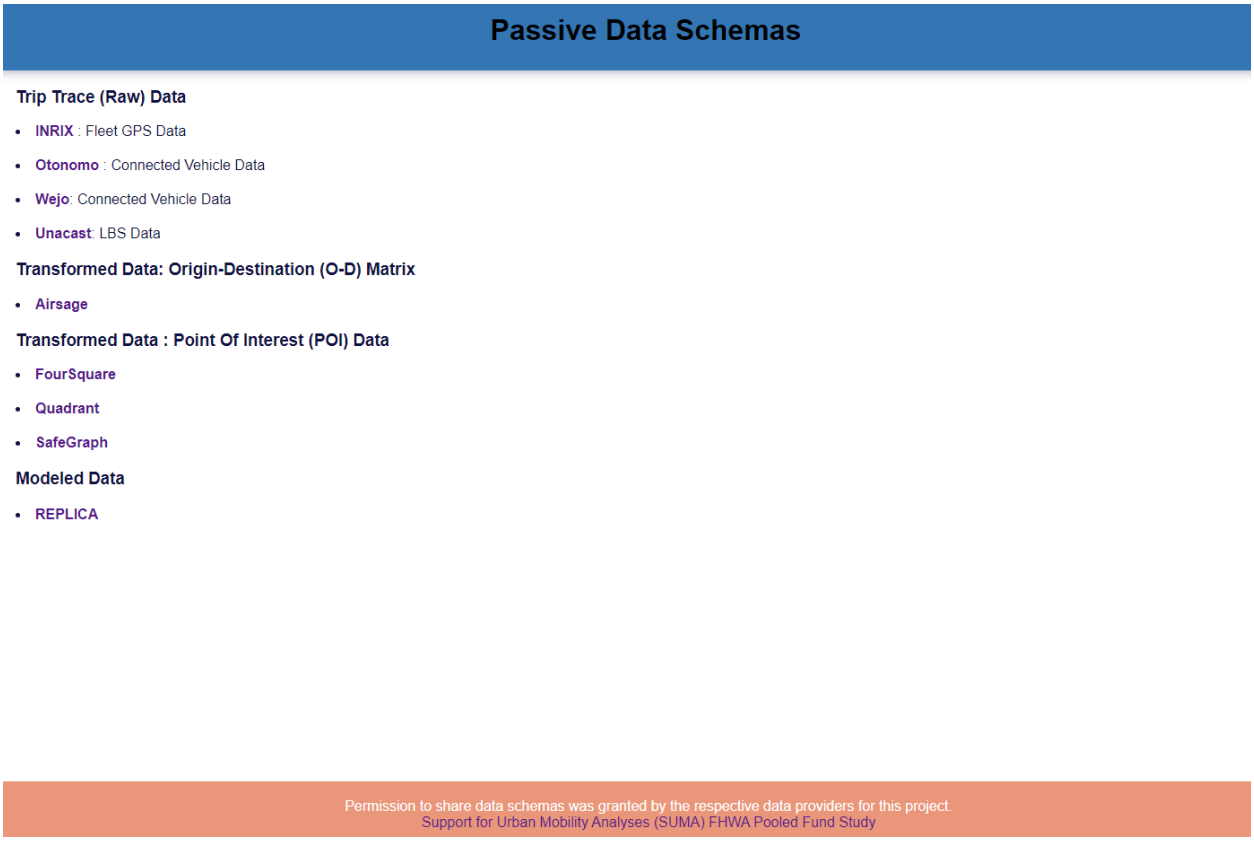
APPENDIX B: Case Studies

DATA	SUMMARY	DATA VENDOR | PLATFORM
Trip Trace Data	**Origin-Destination Study by Clark County, WA**: LINK TO STUDY **Application**: Origin Destination An O-D study performed for Southwest Washington Regional Transportation Council (RTC) used INRIX data and Moonshadow's DB4IoT tool to conduct an in-depth analysis of freeways corridors in Clark County. The Urban Freeway Corridor Operations (UFCO) Study analyzed results from a one-year sample data (January 2018-January 2019) to understand four freeways in the study area. The report also discusses evaluation criteria developed to choose dataset for the study.	INRIX | Moonshadow
	Evaluation of the Impact of Presence Lighting and Digital Speed Limit Trailers on Interstate Speeds in Indiana Work Zones: LINK TO STUDY **Application**: Safety Analysis, Time of Day analysis A joint study between Purdue University and Indiana DOT focused on improving safety around nighttime construction zones. Researchers used connected vehicle data and accident reports and found a correlation between the accidents and hard-braking events. Using the connected car data, the researchers identified areas to deploy advanced warning vehicles. Once the lighting and speed limit trailers were deployed, the results were compared with a work zone without the queue trucks. The results were encouraging – there was a significant reduction in vehicle speeds and hard-breaking events in the work zone with mitigation measures.	Wejo
	2017 Corpus Christi External Study: LINK TO STUDY / RESULTS **Application**: External Travel (commercial + Non-commercial) External study conducted using GPS and LBS data to estimate external-to-external (E-E), external-to-internal (E-I), and internal-to-external (I-E) travel for passenger vehicles and trucks.	INRIX, Airsage
	CDOT Strava Metro Data Analysis Summary: LINK TO STUDY **Application**: Bike and Ped Studies	Strava Metro

	Colorado DOT purchased 24 months of Strava data to gain insights about bicyclist activity patterns in the state. This CDOT study objectives were to develop best practices for database management, since uncompressed individual GPS points is a big data to handle. They correlated the permanent continuous bicycle counter data with Strava data to estimate bicycle activity across the state. Lastly, CDOT identified bicycle corridors in the state and categorized them into low, medium and high use based on Strava bike trip data.	
	Gameday Transportation Evaluations: LINK TO STUDY **Application**: Special Events Study, Time of Day analysis A comparison of pregame and postgame congestion analysis for college football games at City of College Station conducted by TTI.	INRIX
Transformed Data	**TETC Mileage-Based User Fee Study (MBUF):** LINK TO STUDY **Application**: Roadway impact and user fee studies This report summarizes the results of the MBUF truck pilot conducted by The Eastern Transportation Coalition (TETC). TETC collected and analyzed the truck pilot data that was collected using EROAD ELD devices. The study summary shows the estimated costs of fuel, federal fuel tax, state fuel tax and hypothetical MBUF to show a comparison between the estimated costs under the current fuel tax system versus a potential MBUF approach.	EROAD
	Corridor User Analysis for Oregon Department of Transportation (ODOT): LINK TO STUDY **Application**: Toll Road studies A toll studies for ODOT was conducted using the StreetLight (STL) Data. The corridors were analyzed to understand the travel characteristics of the current users. The results from the studies were used by ODOT for the development, screening, and analysis of alternatives for the I-5 and I-205 toll projects.	StreetLight Data
	Data-driven traffic signals in Las Vegas: LINK TO STUDY **Application**: Intersection and Corridor Analysis A traffic signal timing plan was created using connected-vehicle data and other data sources. The data was fed into a traffic management platform and create predictive model to inform traffic signal operators. The results from the model are used to create traffic signal timing plan to measure performance related to safety and efficiency.	GeoTab

Modeled Data	**Sacramento Area Council of Governments (SACOG) Implemented SB 743 with VMT Data:** LINK TO STUDY **Application**: Roadway impact and user fees SACOG worked with the data provider to generate VMT for the entire region since SACOG's Travel demand Model does not include the trips occurring outside of its jurisdiction. The data from Replica was successfully validated by SACOG against on-the-ground observed data gathered from Caltrans' traffic data, population data and other local agencies. After validation, the tool was used to assess impact of new projects on vehicle miles traveled (VMT).	Replica
	Study Name: LINK TO STUDY **Application**: Traffic Impact Analysis Atlanta Regional Council (ARC) used Streetlytics data along with their in-house models to understand the origin-destination patterns of travelers passing through I-85 for alternate routing and the potential impact of route closures resulting from its collapse in 2017.	Streetlytics

APPENDIX C: Schemas



The above website houses schemas of some of the data vendors, it can be accessed via http://passive-data-schemas.s3-website-us-east-1.amazonaws.com/

References

1. Martin, M., Chigoy, B. and Hard. E. (2019). Tools and Best Practices for Using Passive Origin-Destination Data. A technical memorandum to Mobility Measurement in Urban Transportation (MMUT) Pooled Fund Study. Texas A&M Transportation Institute.
2. Monserrat, J.F., Diehl, A., Bellas-Lamas, C. and Sultan S. (2020). Envision 5G Enabled Transport. International Bank for Reconstruction and Development/ The World Bank.
3. Essentra (2021). A guide to 5G small cells and macrocells. https://www.essentracomponents.com/en-gb/news/guides/guide-to-5g-small-cells-and-macrocells
4 .Haosheng H, Georg G., Jukka M. K., Martin R. and Van de Weghe N. (2018). Location based services: ongoing evolution and research agenda, Journal of Location Based Services, 12:2, 63-93, DOI: 10.1080/17489725.2018.1508763
5. Peterson L. and Groot R. (2009). Location-Based Advertising: The Key to Unlocking the Most Value in the Mobile Advertising and Location-Based Services Markets. www.petersonmobility.com
6 . Streetlight (2019). Streetlight Insight – Our Methodology and Data Sources.
7. CDM Smith, Horowitz, A. , Crease, T. , Pendyala, R. and Chen, M. (2014). NCHRP Report 765: Analytical Travel Forecasting Approaches for Project-Level Planning and Design. Transportation Research Board of the National Academies.
8 . Federal Motor Carrier Safety Administration (2018). General Information about the ELD Rule. https://www.fmcsa.dot.gov/hours-service/elds/general-information-about-eld-rule#:~:text=The%20ELD%20rule%3A,carriers%20are%20required%20to%20keep
9 . Prozzi, J., Christensen, S., and Farzaneh, R. (2021). Exploring the Use of Electronic Logging Device (ELD) and Telematics Data for Informing Freight Policy, Planning, and Operations. Texas Depart of Transportation.
10. Vander Lan, Z., Zahedian, S., Aliari S. (2021). The Eastern Transportation Coalition Vehicle Probe Project: HERE, INRIX and TomTom Data Validation
11. Kressner, J.D. (2017). Synthetic Household Travel Data Using Consumer and Mobile Phone Data. IDEA Program Final Report. NCHRP-184.
12. Kressner, J.D., Macfarlane, G.S., Huntsinger, L. and Donnelly, R. (2016). Using passive data to build an agile tour-based model: A case study in Asheville, in Proc. 6th Innovations in Travel Modeling Conference, 2016.
13. "I-85 Bridge Collapse Dataset." (2020). Federal Highway Administration. Accessed January 1, 2020.: www.atlantaregional.org/I85BridgeCollapseDataset.
14 . Singh, S., Gerke, M., Jha. K., Sivaraman, V. and Hard. E. (2022). Summary of Feedback Gathered on Strengths and Challenges of Mobility Data Platforms. A technical memorandum to Mobility Measurement in Urban Transportation (MMUT) Pooled Fund Study. Texas A&M Transportation Institute.