May 2021

INDEPENDENT EVALUATION OF BIG DATA FOR REGIONAL TRAVEL AND MOBILITY ANALYSES

National Capital Region Transportation Planning Board

Metropolitan Washington Council of Governments

Kimley»Horn

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AADT	Average Annual Daily Traffic
ABM	Activity-Based Model
ACS	American Community Survey
AI	Artificial Intelligence
APC	Automated Passenger Counter
API	Application Program Interface
ARC	Atlanta Regional Commission
ATRI	American Transportation Research Institute
ATSPMs	Automated Traffic Signal Performance Measures
AVL	Automatic Vehicle Location
BEA	Bureau of Economic Analysis
BLS	Bureau of Labor Statistics
BMC	Baltimore Metropolitan Council
BRTB	Baltimore Regional Transportation Board
CATT	Center for Advanced Transportation Technology
CAV	Connected and Automated Vehicle
CCP	Connected Citizens Program
CCTV	Closed-Circuit Television
CCWP	Commuter Connections Work Program
CMAP	Chicago Metropolitan Agency for Planning
CMP	Congestion Management Process
COG	Metropolitan Washington Council of Governments
CSV	Comma Separated Values
CTPP	Census Transportation Planning Products
DDOT	District of Columbia Department of Transportation
DFHV	Department of For-Hire Vehicles
DES	Department of Economic Security
DOT	Department of Transportation

DRPT	Virginia Department of Rail and Public Transportation
ETA	Estimated Time of Arrival
ETC	Eastern Transportation Coalition
FAST	Fixing America's Surface Transportation Act
FHWA	Federal Highway Administration
FTA	Federal Transit Administration
FY	Fiscal Year
GIS	Geographic Information System
GPS	Global Positioning System
GUI	Graphical User Interface
HOT	High Occupancy Toll
HOV	High Occupancy Vehicle
HPMS	Highway Performance Monitoring System
HTS	Household Travel Survey
IT	Information Technology
ITS	Intelligent Transportation Systems
LBS	Location-Based Services
LOS	Level of Service
LRT	Light Rail Transit
MaaS	Mobility as a Service
MAG	Maricopa Association of Governments
MDOT	Maryland Department of Transportation
MDOT-SHA	Maryland Department of Transportation – State Highway Administration
MDS	Mobility Data Specification
MPO	Metropolitan Planning Organization
MSA	Metropolitan Statistical Area
MTA	Maryland Transit Administration
MUMA	Moovit Urban Mobility Analytics
MWAA	Metropolitan Washington Airports Authority
MWCOG	Metropolitan Washington Council of Governments

NHS	National Highway System
NHTS	National Household Travel Survey
NPMRDS	National Performance Management Research Data Set
NVTA	Northern Virginia Transportation Authority
O-D	Origin-Destination
OCTO	Office of the Chief Technology Officer
OEM	Original Equipment Manufacturers
OIPI	Office of Intermodal Planning and Investment
PAG	Pima Association of Governments
POI	Points of Interest
PUMS	Public Use Microdata Sample
QA/QC	Quality Assurance/Quality Control
QCEW	Quarterly Census of Employment and Wages
RFI	Request for Information
RITIS	Regional Integrated Transportation Information System
RM3P	Regional Multimodal Mobility Program
RTDC	Regional Transportation Data Clearinghouse
RTS	Regional Travel Survey
SAFETEA- LU	Safe Accountable Flexible Efficient Transportation Equity Act - A Legacy for the Users
SCAG	Southern California Association of Governments
SEMCOG	Southeast Michigan Council of governments
SWG	Study Working Group
TA	Transportation Alternatives
TAP	Transportation Alternatives Program
TAZ	Transportation Analysis Zone
TDF	Travel Demand Forecasting
TDM	Travel Demand Management
TLC	Transportation-Land Use Connection
TNC	Transportation Network Company
TPB	National Capital Region Transportation Planning Board

UCFE	Unemployment Compensation for Federal Employees
UPWP	Unified Planning Work Program
VDOT	Virginia Department of Transportation
VMT	Vehicle Miles Traveled
VPP	Vehicle Probe Project
VRE	Virginia Railway Express
WMATA	Washington Metropolitan Area Transit Authority
LEHD	Longitudinal Employer Household Dynamics

STUDY WORKING GROUP

A study working group (SWG) was created to collaborate in the development of this research and report. The SWG consisted of staff from key functions of the Metropolitan Washington Council of Governments (COG)/National Capital Region Transportation Planning Board (TPB). Development of this report was based on the state-of-the-practice review and Big Data evaluation. The SWG was responsible for discussing and vetting issues throughout the study process.

COG/TPB STAFF

- Tim Canan – Planning Data and Research Program (Project Manager)
- Nicole McCall – Planning Data and Research Program
- Charlene Howard – Planning Data and Research Program
- Yu Gao – Planning Data and Research Program
- Ray Ngo – Travel Forecasting and Emissions Analysis Program
- Stacy Cook – Plan Development and Coordination
- Jaleel Reed – COG Department of Community Planning and Services
- James Li – Systems Performance Planning Program
- Andrew Burke – Systems Performance Planning Program

CONSULTANT TEAM

- Britton Hammit – Kimley-Horn of DC
- Anthony Gallo – Kimley-Horn
- Alan Toppen – Kimley-Horn
- Jiaxin Tong – Kimley-Horn
- Gui Vendemiatti – Kimley-Horn
- Aichong Sun – AECOM
- David Roden – AECOM
- Senanu Ashiabor – Intermodal Logistics Consulting Inc.
- Nick Hinze – Intermodal Logistics Consulting Inc.
- Melissa Sacknoff, Hannah Brusca, Jin Choi – Kimley-Horn

NOTE TO THE READER

This report was prepared in a format such that it can be used as a reference document as Big Data procurement strategies are developed within the TPB. Please find a summary of each chapter and how to use each section below.

Executive Summary | Condensed overview of the report summarizing details related to the project background, purpose, and research methodology; the state-of-practice review; findings from the Big Data product evaluation; the menu of applicable Big Data products; and the implementation plan.

- Use the Executive **Summary** to provide an overview of the project and point you towards the chapter that best suits your needs.

Chapter 1: Introduction and Purpose | Project background, purpose, research methodology, and summary of the TPB's identified research areas for the evaluation.

- Use **Chapter 1** to better understand the motivation for the project and the specified research areas of interest in which the report is based upon.

Chapter 2: State-of-Practice Review | Findings from transportation agency surveys and interviews originally presented to TPB in the Task 2 State-of-Practice Review Memorandum.

- Use **Chapter 2** to learn more about what peer-agencies are doing with Big Data.

Chapter 3: Big Data Product Evaluation | Findings from the independent evaluation of Big Data products. Details of the evaluation methodology are provided, and the evaluation results are presented in multiple ways to meet the diverse needs of TPB and agency partners.

- Use **Chapter 3** as a reference for the Big Data product evaluation findings. This chapter was designed with overlapping content to meet the needs of readers with varying objectives.
- Reference **Section 3.1** for details of the evaluation methodology, evaluation criteria, and Big Data products surveyed within the context of this report.
- Reference **Section 3.2** for details of applicable Big Data products organized by product. This section provides detailed product-specific information including an overview of the product, validation documentation and case studies, cost and pricing structure, data licensing and sharing data storage, processing, and analysis considerations.
- Reference **Section 3.3** for details of applicable Big Data products organized by the specified research areas. This section consolidates the applicable Big Data products for each specified research area and should be used to understand what options exist for a specific research need. Product details can then be referenced in **Section 3.2**.
- Reference **Section 3.4** for a summary of the evaluation findings and a comprehensive table of opportunities for, and limitations of, Big Data for each specified research area.

Chapter 4 Recommendations | The ultimate recommendation and procurement plan resulting from the independent evaluation process, and a discussion of promising Big Data solutions for future consideration and research.

- Use **Chapter 4** to learn more about the variety of procurement options available for investing in Big Data products and reference the immediate-term and long-term recommendations that resulted from this independent evaluation of Big Data.

EXECUTIVE SUMMARY

EXECUTIVE SUMMARY

INTRODUCTION

> **Big Data** is characterized as an **information asset** with such high volume, velocity, variety, and veracity that specific technology and analytical methods are required for its transformation into value.

An independent evaluation of Big Data and its use and limitations in regional transportation planning applications was completed for the National Capital Region Transportation Planning Board (TPB). The consultant team worked closely with the TPB study working group (SWG) to clearly identify the TPB's data needs within each of the agency's research areas prior to beginning the evaluation. The evaluation explored the strengths, limitations, applicability, and acquisition costs of common Big Data products used by transportation agencies. The evaluation included a detailed state-of-practice review consisting of a literature review, peer agency survey and interviews, and Big Data vendor interviews. Each Big Data product was evaluated against a scoring rubric developed by the consulting team and the SWG. The scoring rubric used seven evaluation criteria to evaluate whether each product would be an effective investment for the TPB based on the defined research areas and data needs. Once the evaluation was completed, additional feedback was gathered from the SWG to further refine the consultant team's understanding of the most critical needs for Big Data at the TPB. Next, meetings were conducted with local peer agencies (i.e., the Virginia Department of Transportation [VDOT], District Department of Transportation [DDOT], Virginia Department of Rail and Public Transportation [DRPT], Washington Metropolitan Area Transot Authority [WMATA], Maryland Department of Transportation – State Highway Administration [MDOT-SHA]) to learn about each agency's experience using Big Data, their greatest needs for Big Data, and to discuss potential collaboration in procuring new Big Data products. Using the insight gleaned from each step of this process, the consultant team recommended a pilot procurement plan for the TPB to trial one or two Big Data products that offer promise for multiple data needs. Lastly, an overview of the evaluation and partner agency feedback was presented to TPB leadership in February 2021.

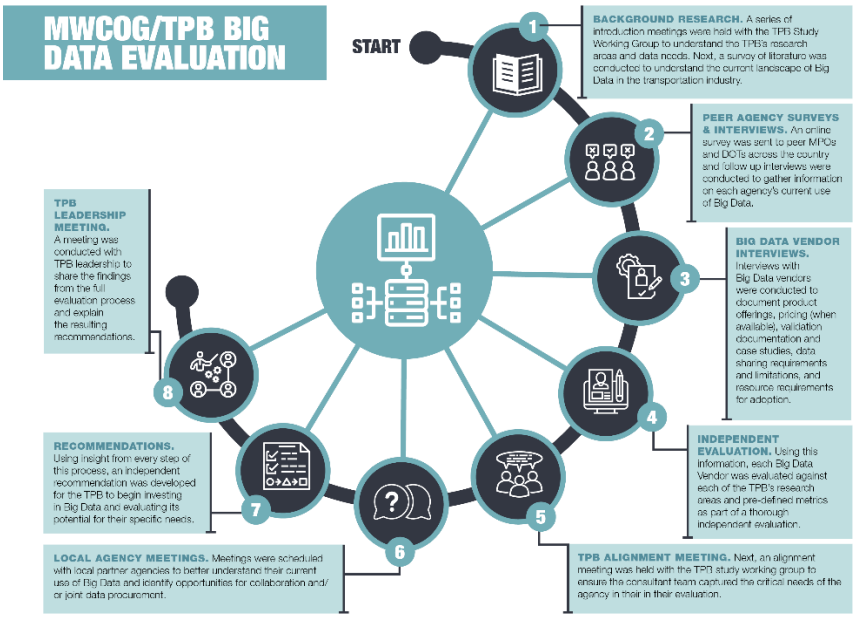


Figure ES–1 | Independent Evaluation Process

PURPOSE AND BACKGROUND

The purpose of this report is to present the TPB with an independent evaluation of Big Data for regional transportation planning applications. It identifies opportunities for, and limitations of, Big Data; evaluates product offerings; and ultimately recommends an implementation strategy. Big Data products with transportation planning applications include but are not limited to the following:

- Passively collected data from mobile applications (including global positing system [GPS] traces and location-based services)
- Large socioeconomic or "location-based" datasets to better understand population and employment characteristics
- Data from public infrastructure (including transit on-board vehicle sensors, traffic sensors, cameras used to monitor traffic flow), and smart card data
- Emerging Big Data sources (including connected and automated vehicle [CAV] data), transportation network company (TNC) data, and micromobility data

The TPB has previously applied Big Data products in a few of its core functions, such as travel demand forecasting, regional congestion monitoring and performance measurement, and travel/traffic data collection. Seeking opportunities to broaden the use of these products, this evaluation identifies specific Big Data products that have the potential to further enhance the technical analyses performed within the TPB program as well as to inform critical decision-making for regional transportation planning, performance measures, and programming.

The evaluation provides a roadmap for the TPB to implement and procure Big Data products and to advance their current practice using Big Data solutions. The evaluation also assesses the feasibility of applying Big Data products to a series of research areas identified by a TPB SWG

A TPB SWG identified eight research areas for potential application of Big Data. Each research area is summarized in **Table ES-1**.

Table ES-1 | Research Areas for Evaluation

Research Area		Description
1	Travel Demand Forecasting	• Origin-Destination (O-D) information by mode with trip purpose/destination/origin type • Region-wide speeds and volumes • Region-wide signal delay estimation/queue length
2	Travel Demand Management	• Estimating network demand using historical data • Survey of employer telework policies • Forecasts of future telework policies
3	System Performance/Congestion Management	• Monitoring network-wide congestion and performance • Detailed network performance data for strategy development • Event traffic data
4	Transit and Active Travel	• Impact of mode choice on network level of service (LOS) (scenario analysis) • Monitoring real-time transit performance • Monitoring active travel demand and delay

Research Area		Description
5	Transportation Network Companies	• Improved understanding of usage and demand for TNCs: temporally, geographically, O-D patterns • Socioeconomic and demographics of TNC riders • Relationship between TNCs and transit (e.g., first/last mile problem)
6	Traffic Counts	• Obtaining permanent/temporary traffic count data • Variability analysis of traffic counts (e.g., day-by-day) • Methods to validate traffic counts collected in traditional methods • Vehicle classification data
7	Connected and Automated Vehicles	• Impact of CAVs to vehicle occupancy • Impact of CAVs on multimodal travel demand • Impact of CAVs on infrastructure (e.g., parking) and revenue streams • Available data from CAVs
8	Other Research Areas	• Regional freight and commercial vehicle travel forecasting • Household survey data validation and supplementation • O-D patterns based on population demographics • Improved project selection

The research team conducted the following tasks in developing this report:

- State-of-practice literature review of Big Data sources and their application areas.
- Online survey sent to peer agencies—including peer metropolitan planning organizations (MPOs) and state departments of transportation (DOTs)—regarding their use of Big Data.
- Follow-up phone interviews with a subset of peer agencies to gather more information.
- Interviews with Big Data product vendors to better understand product offerings, pricing (when available), validation documentation and case studies, data sharing requirements and limitations, and resource requirements for adoption.
- Independent evaluation of Big Data products based on their applicability to the TPB's identified research areas.
- Development of a menu of Big Data solutions with a list of implementation options and recommendations for consideration by the TPB.
- Meetings with local partner agencies to better understand their experience with and interest in Big Data products.
- Development of recommendations for the TPB to begin procuring promising Big Data products.
- Meeting with TPB leadership to share key report findings and present the recommendations.

STATE-OF-PRACTICE REVIEW FINDINGS

The initial step of this effort consisted of a state-of-the-practice research on how peer agencies nationwide are using or planning to use big data in transportation. Peer agencies were selected based upon their geographic proximity to the metropolitan Washington region or comparable metropolitan regional size and functions. Twenty-two peer agencies responded to an online survey and eight were selected for follow-up interviews based on their responses to the survey. This review found that the use and implementation of Big Data products is still relatively new to most transportation agencies; despite this, some agencies are more familiar with or using a larger number of Big Data products than others. The survey and interview revealed the following perspectives about Big Data products:

Agencies maintain a healthy level of skepticism on the quality of the Big Data products.

- Agencies maintain a high-level of interest in Big Data due to the potential applications, as Big Data supplements traditional data collection in terms of its relative high volume, low latency, and wide spatial coverage.
- Agencies also maintain a healthy level of skepticism on the quality of the Big Data products, which stems from perceived vendor black box methodologies and a lack of clear and transparent validation. These factors have prevented the agencies from adding Big Data products as a permanent fix to their toolbox, particularly in areas outside system performance/congestion management.

The majority of peer agencies queried as part of this state-of-practice review are leveraging traditional data sources (e.g., Census data and travel surveys) as well as emerging Big Data products to address their agency needs. Big Data products from mobile devices are the most frequently used, with common vendors including INRIX and StreetLight. The prominent application areas for these Big Data products were travel demand forecasting and system congestion monitoring.

Implementing Big Data requires new information technology resources, hiring strategies, and staff training.

In most cases, peer agencies did not suggest that Big Data products were replacing existing data streams; rather, these emerging data sources were being used to supplement and validate existing methods and programs. The challenges agencies are facing in the adoption of Big Data products range from ensuring sufficient data validity and quality to overcoming budget and timeline constraints to technical requirements for storing and processing Big Data. Despite these challenges, several agencies found the insights offered by Big Data valuable to their organizational goals and plan to continue using multiple sources of Big Data for model validation and other application areas.

In conversations with local partner agencies in the metropolitan Washington area, similar themes emerged.

- Many of the agencies talked about the barrier of the “Black Box”. They described that it is not only difficult to not know the underlying algorithms and data sources, however, it is also challenging to keep up with the regularly evolving evaluation methodologies and data sources that make it difficult, if not impossible, to compare data year-over-year.
- Implementing Big Data involves both information technology resources, new strategic hiring strategies, and staff training programs. The information technology needs include data processing, data storage, and cloud computing as well as implementing data dashboards and visualization. The agencies also discussed the importance of integrating new team members with data science backgrounds as well as providing additional training opportunities for staff with more traditional experience for their agencies.

BIG DATA PRODUCT EVALUATION FINDINGS

The research team conducted an independent evaluation of Big Data products by appraising their use and limitations for the eight identified research areas. A wide selection of Big Data products was evaluated for potential value in each research area and ultimately given a yes or no recommendation based on findings from the agency surveys and interview, vendor interviews, literature review, and evaluations from technical experts identified by the research team. Each product was evaluated based on seven evaluation criteria, which were developed from the research objectives. **Figure ES–2** shows the evaluation methodology conducted for each Big Data product. This evaluation methodology was developed by the research team in close partnership with the SWG to ensure that the results would satisfy the TPB's objectives for conducting this thorough evaluation.

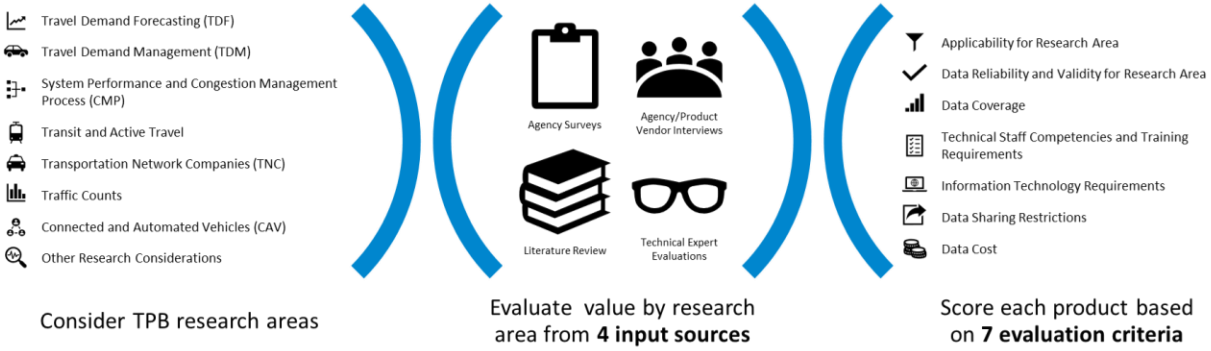


Figure ES–2 | Evaluation Methodology

Each Big Data product was considered separately for each research area; therefore, a product could be recommended for one research area and not another. For each research area, the research team developed a scoring matrix to rate each data product against each evaluation criterion based on the needs of the specific research area.

Based on the findings related to the various products, the research team organized the products into five different groups:

- **Mobile Device-Based Travel Patterns Data** — vendor products inferring information on travel patterns using aggregated, anonymized data from mobile devices. From this pool of products, four products were specifically identified that are potentially applicable to a large number of the TPB's research needs. Other mobile device-based products also were reviewed that were deemed to have a more limited applicability; those products are described in this report in less detail.
- **Socioeconomic or Location-Based Data** — products that provide information about specific locations or various socioeconomic data elements for a region. Some of these products may not necessarily reside in the domain of Big Data by definition but do have applicability to the TPB for understanding the underlying demographics and economic generators of travel in the region. These data sources also may be applicable to one or a few specific TPB research areas.
- **Data from Public Infrastructure** — data derived from physical infrastructure deployed by an agency in the field, such as traffic count sensors, traffic signals, and transit intelligent transportation systems (ITS) devices. In these cases, data feeds vary from agency to agency and the type of sensors in which they have invested.
- **End-User Platforms for Data Analytics** — products that feature dashboards geared toward an end-user of big data; some of these platforms ingest data produced by external sources and claim to be data-agnostic. Thus, these products are used in combination with data from other sources to produce analytics, summary metrics, and visualizations. The selection of end-user platforms for data analytics

is dependent on the data product being analyzed. Numerous Big Data product vendors provide unique platforms for analytics and visualization.

- **Emerging Big Data Sources** — products that the research team considers to be in the emerging stage or any commercial products for transportation planning purposes that are very novel or are being developed for integration with end-user products. These represent areas where the research team anticipates significant progress during the next several years but does not currently recommend that the TPB invest in a solution. Rather, the research team recommends that the TPB take a proactive approach in monitoring the development and progress of these various emerging data sources and ensuring regional coordination in their applications.

Finally, for each of the eight identified research areas, the research team identified the most applicable Big Data products. Details of this comprehensive evaluation process can be found in **Chapter 3** and **Appendix B**.

MAJOR TAKEAWAYS

The research team identified several high-level trends during the Big Data products evaluation:

- **There is no one size fits all solution to all the TPB's research area needs**. Each of the Big Data products evaluated had unique advantages and disadvantages, such as only offering data for certain modes or certain trip purposes. In particular, the research team found that mobile device-based Big Data sources face challenges in filtering trips by various sub-modes, with perhaps the exception of truck trips. The products evaluated differ in data resolution and level of detail as well as data validity and transparency.
- **The landscape of Big Data products is constantly evolving, even throughout the course of this study**. The products evaluated generally fall somewhere along a spectrum as shown in **Figure ES-3**. Typically, products from novel or emerging data streams provide data in a very raw format requiring significant data storage and processing abilities to glean information. As various products mature and gain more widespread use, dashboards and visualization tools to more easily interpret data are made available as an accompaniment to the products. These end-user analytic platforms can be built within a vendor's product offering or may be provided by third parties that ingest data from other sources (such as the University of Maryland Regional Integrated Transportation Information System [RITIS]). Vendor products are continuously evolving across this spectrum.

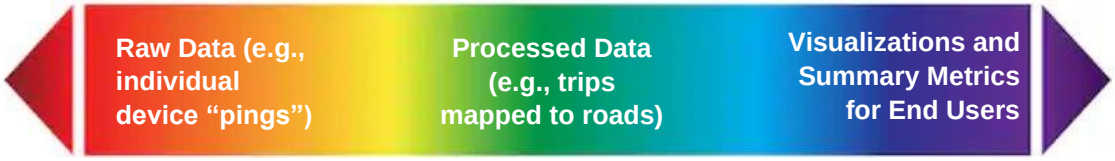


Figure ES-3 | Big Data Product Spectrum

- **Opportunities for collaboration and data sharing exist to aid in procurement and downstream processing of data.** For data sharing, opportunities exist to expand a dataset's coverage from the current state agency agreements with vendors. The TPB plays an important role as a central repository of planning data for the region, including its Regional Transportation Data Clearinghouse (RTDC)[1]. The RTDC contains information varying from traffic counts to transit to aviation data. The TPB should continue to evaluate and identify opportunities to expand the scope of the RTDC. The TPB also certainly faces challenges given the multijurisdictional and multistate makeup of the region. One primary challenge in operating the RTDC is the dependency of the TPB on partner-agencies for obtaining data.

[1] rtdc-mwcog.opendata.arcgis.com

Additional challenges associated with differing data formats and the like are often more easily overcome than obtaining the data in the first place.

- **Agency-deployed field infrastructure is still needed for ground-truth data**. While this study sought to examine the potential for "Big Data to validate field data", the findings showed that, at this stage, field data is overwhelmingly used as ground-truth data to validate Big Data. For this reason, continued investment in upkeep and maintenance of field infrastructure such as traffic count stations, traffic signal infrastructure, and transit ITS is warranted to allow for reliable ground-truth data.
- **Certain datasets likely exist but are not yet available for public agency use**. These datasets include tolling and pricing data from privately-owned and operated managed lanes facilities in Northern Virginia, data on TNC trips—outside of those reported to the District of Columbia, and similar data on micromobility trips (e.g., scooters).
- **The TPB can take a proactive approach in monitoring the development and progress of various emerging data sources and ensuring regional coordination in their applications**. Emerging data sources that the TPB should focus on include taxi/TNC data, micromobility data, and CAV data. Further, the more that the jurisdictions can coordinate to encourage common data standards and reporting, the more useful this data will be to transportation professionals and decision-makers in the region.

OPPORTUNITIES AND LIMITATIONS

After completing the independent evaluation of Big Data products geared towards the identified TPB research areas, the opportunities, and limitations for advancing each research area were summarized. This summary is provided in **Table ES-2**.

Table ES-2 | Opportunities and Limitations by Research Area

Research Area		Description of Big Data Opportunities and Limitations
1	Travel Demand Forecasting	• Several platforms exist for providing vehicular origin-destination (O-D) data at various levels of granularity: for example, some products provide this data at the transportation analysis zone (TAZ) or census tract level, while other platforms offer O-D data for specific roadway segments (e.g., INRIX, StreetLight, Replica). The most common agency applications for O-D data for travel demand modeling is for validating model survey data, especially external travel nodes. • O-D information for full set of transportation modes has become available from some (e.g. Replica) but not all data vendors • Trip purpose data is inferred based on device movement characteristics; as such, ascertaining a location other than home or work to a device for other trip purposes remains a challenge. However, Replica seems to have some success in this regard using a complex data integration and synthetic population modeling approach. • The accuracy and availability of O-D patterns by route and road segment is subject to the resolution of data needs; for example, there is currently no cost-effective means to collect accurate O-D data for facilities that are closely-spaced (e.g., Express Lanes and general purpose lanes that are only separated by lane markings or barriers). O-D data for lower-volume facilities may suffer from low sample rates. • While daily and subdaily traffic count data is offered by several Big Data vendors; such information is less readily available or valid for lower-volume roadway facilities at this time. • Region-wide signal delay estimation/queue length is not available for incorporating into a travel demand model at this time but may evolve in the near future. This information can be potentially obtained from Automated Traffic Signal Performance Measures (ATSPMs), but the deployment of

Research Area		Description of Big Data Opportunities and Limitations
		ATSPM infrastructure and systems would be on an individual agency level and either has not started or is still in its early stages in the region. Data from CAVs with high temporal resolution (less than 3 seconds) offers the potential for obtaining these measures but is unproven at this time.
2	Travel Demand Management (TDM)	• Evaluating the influence of telework policies on travel behavior relies on present and historical travel patterns data; it is not real-time as it takes time for travel behavior to change and the data needs to be archived throughout a period of time. • The travel behavior information, depending on the resolution desired—mode of travel, home/work locations—may have limitations and caveats as outlined above under Travel Demand Forecasting. For example, home/work location data by TAZ is much more reliable than route or mode data. • Additional contextual data and customized analysis may be needed to differentiate visitors versus home-based workers.
3	System Performance/Congestion Management	• Big Data is widely used for system performance/congestion management for vehicular travel on freeways and higher-volume arterials. The validity and accuracy of data for non-National Highway System (NHS) roads is subject to further evaluation, particularly for rural areas, because of sample rates of mobile device-based data and the number of road sensors deployed on these facilities. • Multimodal system performance data using mobile device-based travel pattern data is currently unattainable or very novel due to difficulties to chain the trips by different modes as well as to differentiate between certain modes (e.g., bus as mentioned above in Travel Demand Forecasting). • Monitoring network-wide congestion and performance can be achieved using a combination of mobile device-based travel pattern data and data from public infrastructure to produce different congestion and performance metrics; several end-user platforms are widely used for this purpose. This requires agencies to continue to invest and maintain this public infrastructure including traffic count sensors, traffic signals, and ITS devices, including those that are part of transit systems.
4	Transit and Active Travel	• Monitoring real-time transit performance is feasible using automatic vehicle location (AVL) data, and some end-user platforms for data analytics provide both hardware and software to report real-time performance; however, integrating data across transit agencies can be challenging, as they may use different end-user platforms and may not share standardized data formats for easy exchange of information. • Monitoring active travel demand and delay (e.g., for bicycles and pedestrians) using Big Data for these modes is fairly novel, and validation is limited at this time, in part due to a lack of validation data available from public infrastructure.
5	Transportation Network Companies (TNCs)	• There are no proven and large-scale methods to understand usage and demand for TNCs given the proprietary nature of these datasets; some Big Data vendors offer custom analyses attempting to infer TNC trips based on device movement characteristics. • TNCs are required by D.C. law to report detailed trip data to the D.C. government, from whom TPB could potentially request access for agreed-upon internal analysis purposes. The details for how to access the data are contained in the agreement between TPB/ WMATA and the D.C. government. This dataset is limited to TNC trips that either start or end in D.C.

Research Area		Description of Big Data Opportunities and Limitations
		• Socioeconomic and demographics of TNC riders are only available from regional or National Household Travel Surveys (NHTS). • The relationship between TNCs and transit (e.g., first/last mile problem) could be inferred using the datasets mentioned above but would not be straight forward at this time. • Emerging end-user platforms for data analytics, such as SharedStreets, allow for summarizing TNC and micromobility data; DDOT is currently using this platform.
6	Traffic Counts	• Mobile device-based Big Data counts are still being validated by counts collected using traditional methods, not the other way around. Continuous count data from road sensors plays a critical role in providing control points for mobile device-based Big Data solutions and is anticipated to continue to do so in the future. These control points are needed to validate Big Data solutions as they are continuously being refined to estimated ubiquitous counts. • Variability analysis of traffic counts (e.g., day-by-day) is obtainable using several Big Data products, subject to the caveat of limited sample sizes on facilities for individual days, particularly for lower functional class roads and facilities in rural or less dense areas. • Vehicle classification data is heavily dependent upon data from road sensors. Big Data products do not provide sufficient information about the vehicle classification that is adequate with road sensors (e.g., continuous count stations). • Big data solutions currently cannot collect vehicle occupancy data. The road sensor technology for collecting vehicle occupancy data is available but the accuracy is not there yet, especially for tolling industry deployment. • Mobile device-based Big Data products could be used to support Highway Performance Monitoring Systems (HPMS) for certain data types, but information such as roadway conditions and heavy vehicle percentages are difficult or unattainable via Big Data.
7	Connected and Automated Vehicles (CAV)	• Implementable CAV data that could be used for widespread understanding of travel patterns and traffic operations is not yet available. The research team feels that the most promising near-term data sets are likely coming from third-party solutions integrating high-resolution data from vehicle original equipment manufacturers (OEMs), which report data in very short time intervals (3 seconds or less) and provide additional data relating to vehicle trajectory (e.g., braking/acceleration).
8	Other Research Areas	• Regional freight and commercial vehicle travel O-D data is offered by Big Data vendors for medium and heavy trucks. • Household survey data validation and supplementation can rely on mobile device-based travel pattern data. For example, the Federal Highway Administration (FHWA) is embarking on the Next Generation NHTS using mobile device-based Big Data products. • Big Data vendors that provide O-D patterns based on population demographics are inferring this data from underlying census information.

RECOMMENDATIONS

This evaluation resulted in a thorough menu of potential Big Data product procurement options that were provided to the SWG for feedback (**Appendix D**). Further refinement of these options led the research team to the following recommended procurement plan and long-term considerations.

IMMEDIATE PROCUREMENT PLAN

For the initial procurement, it is recommended that a **pilot program approach** be taken to allow the TPB to trial one or two products prior to the agency making a large investment. Three pilot procurement options were developed that meet the greatest number of TPB data needs.

- **Option 1** | Informal Trial of StreetLight
- **Option 2** | Formal Trial of Replica
- **Option 3** | Informal Trial of StreetLight & Formal Trial of Replica

OPTION 1: STREETLIGHT

As described in the report, StreetLight was rated well during the independent evaluation for offering numerous products that met the data needs identified by the TPB. Additionally, partner agencies in the Metropolitan Washington region have experience using StreetLight and have made substantial investments in the product.

Informal Trial: As an established company in the industry, StreetLight is not likely to offer a free data trial. Therefore, it is recommended that the TPB explore options for conducting an informal trial of StreetLight product offerings to determine whether a larger long-term investment of the data is warranted.

Evaluation Strategies: In order to evaluate the applicability of StreetLight products for the TPB's research areas and needs, it is recommended to select one or more specific projects with clear applications and objectives to focus on during the trial period. The scope of these projects should be used to determine how the data is acquired, which data is most appropriate for supporting the purpose and need of the project(s), and how much data is needed to achieve a valuable result.

Data Acquisition: The team identified two primary options for procuring a sample of StreetLight Data to support this initial pilot procurement.

- First, the TPB can request free access to the VDOT StreetLight subscription (subscription allows free access to partner MPOs). The VDOT subscription will not provide data for trips in Maryland and Washington, D.C., that do not touch Virginia, but it includes a 15-mile buffer around the commonwealth boarder. Therefore, TPB staff could set up analyses within with VDOT subscription in this constrained geographic area to gain familiarity with the latest StreetLight product offerings and further understand the strengths and limitations of the tool. The TPB could also use this data to perform validation analyses on the data provided to determine whether a larger investment in the product would provide sufficient value to the TPB.
- Second, the TPB can define a series of zones within the Metropolitan Washington region and purchase a finite amount of data for those zones directly from StreetLight. For example, 50 zones can be defined within the region and data could be requested for each season of the year (i.e., spring, summer, fall, winter). This data set should be available for approximately $25,000 based on preliminary conversations with StreetLight sales, assuming the TPB will want to obtain data for various non-vehicular modes (e.g. pedestrians, bike, bus, rail).

OPTION 2: REPLICA

The second option is completing a formal trial of a new Big Data product, Replica, which also was identified during the evaluation as being a promising product for addressing the TPB's data needs. Replica is a new online platform that models "synthetic population" of a given area at a very detailed level for a wide range of modes. Replica was formed as part of Alphabet, Google's parent company, in 2018, and became an independent company in 2019. The underlying dataset includes LBS and GPS data, consumer resident data, land use and real estate data, credit transaction data, and ground truth data (such as traffic counts and transit boarding). This online platform includes three main products:

- **Places |** a series of detailed, activity-based travel models for specific regions at specific points-in-time, with data down to the network-link level. Replica currently offers a Baltimore-Washington "Places" model, which includes the entirety of Maryland, Virginia, and the District of Columbia.
- **Trends |** a higher-level dashboard comparing mobility patterns over time, with metrics including: total trip counts, mode split, the percentage of people commuting to work, the percentage of people sheltering-in-place, and consumer spend, as well as COVID-19 data. Trends data is available for the entire US and can be obtained at the census tract, city, county, MSA, or state level.
- **Scenarios (not yet available) |** "what-if" analysis for changing network links/land use.

Additional detail on Replica product offerings can be found in the final report.

Formal Trial: Due to their new entry into the Big Data market, Replica is offering agencies trial packages to increase exposure to their products. The TPB could take advantage of these formal trials offered by Replica to learn more about their product offerings and perform validation analyses on the data provided to determine whether a larger investment in the product would provide sufficient value to the TPB.

Evaluation Strategies: The recommended evaluation strategy for Replica is similar to that suggested for StreetLight; however, because an actual trial version of the Replica platform is likely available, the TPB may have less ability to define the exact data acquired and will likely need to perform the evaluation at a more rapid pace. For these reasons, it is recommended to open conversations with Replica to define the scope of the trial and determine which products will be included and for what duration. Next, prior to the trial beginning, the TPB should define a handful of specific projects or research questions that will be evaluated as part of the trial. This way, when the data becomes available, the TPB will be able to make efficient use of this pilot evaluation.

Data Acquisition: The TPB can likely reach out directly to Replica to initiate a formal trial of the platform. The project team is aware of other agencies, including DDOT, being introduced to or conducting trials of Replica, but it is unclear at this time if those trials are still ongoing. All data must be obtained through the online platform but can be downloaded for further analysis if needed. The project team has explored the synthetic population data from the Places model and notes that the individual person-trip tables are very large files.

OPTION 3: STREETLIGHT & REPLICA

The third option combines the first two options, offering the TPB the opportunity to create an informal trial of StreetLight, while also conducting a formal trial of Replica. This option would allow the TPB to conduct similar analyses on both data sets to compare their effectiveness for addressing data needs and identify which platform may offer greater long-term benefits for the agency.

Evaluation Strategies: The recommended evaluation strategy for the combined acquisition of StreetLight and Replica is similar to those recommended for both independently. To achieve the greatest comparison of the two data products, it is recommended to select a few projects or research questions that can be evaluated with both vendors. Since the trial with Replica may be a formal trial, there is likely less flexibility

in which data can be provided; therefore, it may be advantageous for the TPB to identify the scope of the Replica offering, and then seek to procure a similar data sample from StreetLight.

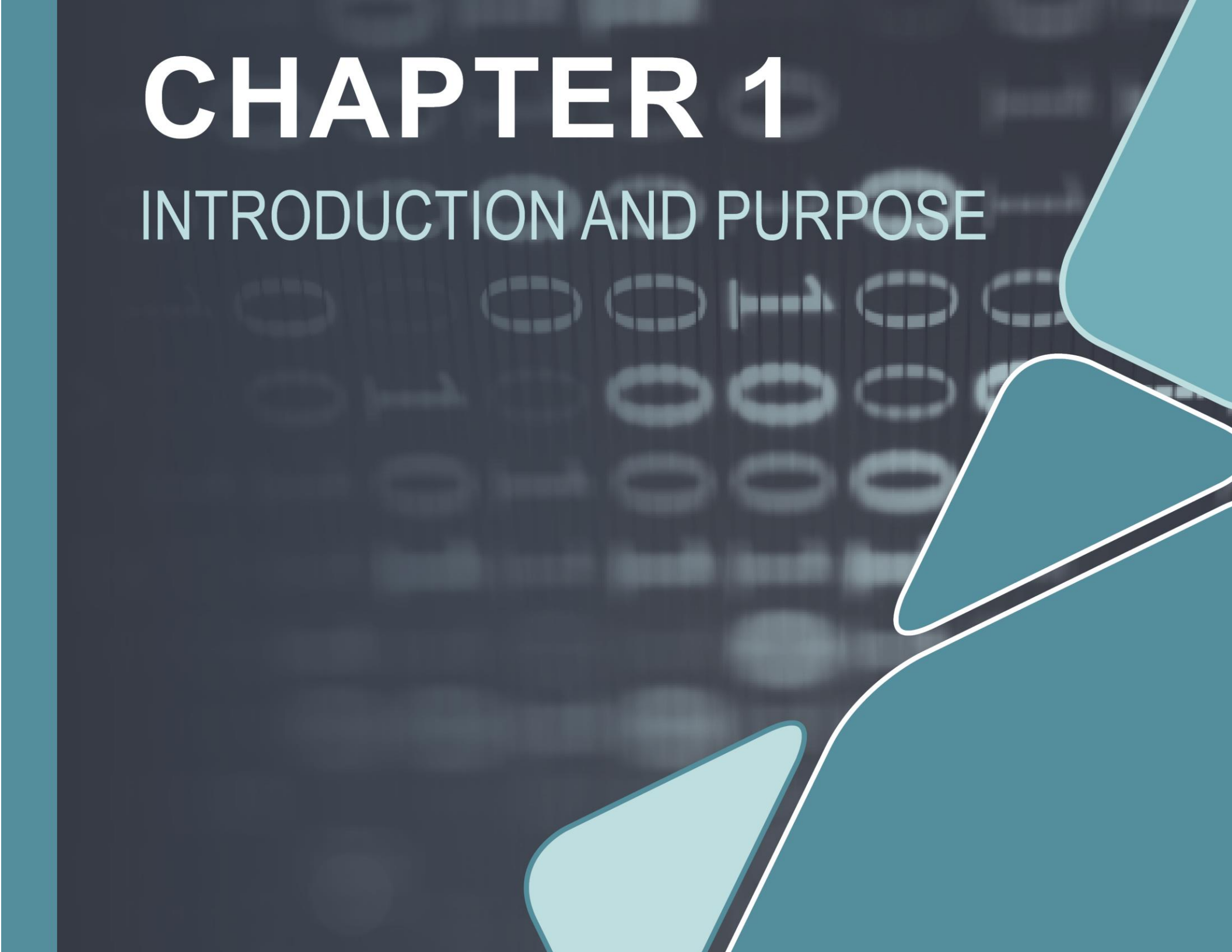
LONG-TERM RECOMMENDATIONS

The research team also identified a number of long-term recommendations for the TPB to consider as the landscape of Big Data continues to evolve.

- **Identifying Promising Research Leading to Practice**. Promising research is currently being sponsored through FHWA and other entities, such as universities and research institutes, to reduce the unknowns in using Big Data and introduce transparency to the analytic process. These projects are resulting in open-source code and documentation that sheds light on details of Big Data processing procedures and introduces new opportunities for processing raw data in the future. While these projects are still at a highly granular level and may not be ready for the TPB's application at the time of this report, they hold promise and could be evaluated further as research evolves into practical application. The TPB should partner with universities in the region to continue carrying out research on Big Data.
- **Staying Proactive**. As the landscape of Big Data products continues to evolve at a rapid pace, the TPB should take a proactive approach to monitoring the development of new products, innovative technology in data acquisition, and new product validation studies. This could be done by conducting an annual or biannual survey of peer agencies and reporting updates of this Big Data product evaluation, specifically geared towards identifying how the industry is embracing new Big Data products (e.g., new product vendors within existing data categories as well as data from completely new data streams, such as CAVs, micromobility, TNC, and gig travel). This continuous update also would enable the TPB to stay up to date on the latest experiences with common Big Data products and recent validation information from research and practice.
- **Regional Advocate for Big Data and the Importance of Maintaining Infrastructure**. Finally, the TPB can be a regional leader advocating for the continued investment in transportation infrastructure that supports the use of Big Data. Currently, Big Data validation relies on the baseline data provided by transportation infrastructure; therefore, the continued maintenance of this infrastructure is imperative. The TPB can actively support continued investment in regional traffic count stations, sensors for monitoring bicycle and pedestrian activities, multimodal traffic signal infrastructure and communications, transit fare collection and counter technology, advanced transit systems (real-time performance, scheduling, crowding), and further ITS infrastructure. The TPB can also advocate for data standards for various Big Data sources (for example, the Mobility Data Standard for micromobility data). Finally, the TPB can continue to be an advocate for the behind-the-scenes "infrastructure" for storing and archiving data, such as bringing Big Data products or outputs into the Regional Transportation Data Clearinghouse or another regional solution facilitating open data access.

CHAPTER 1

INTRODUCTION AND PURPOSE



1 INTRODUCTION AND PURPOSE

The Metropolitan Washington Council of Governments (COG) is the regional organization of the Washington area's 24 major local governments and their governing officials, plus area members of the Maryland and Virginia legislatures and the U.S. Senate and House of Representatives.

COG provides a focus for action on issues of regional concern such as comprehensive transportation planning, air and water quality management, environmental monitoring, tracking economic development and population growth and their effects on the region, coordinating public safety programs, and promoting childcare and housing for the region. COG is supported by financial contributions from its participating local governments, federal and state government grants and contracts, and through grants and contracts from foundations and the private sector.

COG is the administrative agent for the National Capital Region Transportation Planning Board (TPB) that is responsible for coordinating transportation planning at the regional level in Northern Virginia, Suburban Maryland, and the District of Columbia. The TPB is the federally designated Metropolitan Planning Organization (MPO) for the region and plays an important role as the regional forum for transportation planning. The TPB prepares plans and programs that the federal government must approve for federal-aid transportation funds to flow to the metropolitan Washington region.

Members of the TPB include representatives from the transportation agencies of the states of Maryland, Virginia, and the District of Columbia, local governments; the Washington Metropolitan Area Transit Authority (WMATA); the Maryland and Virginia General Assemblies; non-voting members from the Metropolitan Washington Airports Authority (MWAA); and federal agencies.

The TPB was created in 1965 by local and state governments in the Washington region in response to a requirement in the 1962 highway legislation mandating the establishment of official MPOs. The TPB became associated with COG in 1966, serving as COG's transportation policy committee.

1.1 PURPOSE

The purpose of this report is to present the TPB with an independent evaluation of Big Data and its use and limitations in regional travel and mobility analyses and modeling and ultimately recommend an implementation strategy. The scope of this project was developed through the TPB, in collaboration with the District of Columbia Department of Transportation (DDOT), the Maryland Department of Transportation (MDOT), the Virginia Department of Transportation (VDOT), the Virginia Department of Rail and Public Transportation (DRPT), and WMATA.

For the purposes of this study, Big Data is characterized as an information asset with such high volume, velocity, variety, and veracity that specific technology and analytical methods are required for its transformation into value[2]. In addition, traditional transportation data sources that are large in volume are considered Big Data for the purposes of this evaluation.

- "Volume" refers to the flagship characteristic that makes Big Data big: its sheer volume. Big Data sources contain a lot of data.
- "Velocity" refers to the frequency in which new data are being collected.
- "Variety" refers to the type of data being collected across the spectrum from structured and unstructured data. Structured data refers to data that can fit in a traditional database, whereas unstructured data

[2] De Mauro, Greco, Grimaldi, (2016) "A formal definition of Big Data based on its essential features", Library Review, Vol. 65 Issue: 3, pp.122-135, https://doi.org/10.1108/LR-06-2015-0061

refers to data that cannot easily be structured in a database (e.g., social media feeds, traffic camera footage).

- "Veracity" refers to the trustworthiness and reliability of the incoming data stream.

With processing and analytics tools, Big Data can illustrate patterns and trends in human behavior and activity. Big Data products with transportation planning applications include but are not limited to passively collected data from mobile applications, including global positioning system (GPS) traces and location-based services; data from public infrastructure, including on-board vehicle sensors, traffic sensors and cameras used to monitor traffic flow, and smart card data; emerging Big Data sources, including connected and automated vehicles (CAVs) data, Transportation Network Company (TNC) data, and micromobility data, among others.

The TPB has previously applied Big Data products in a few of its core functions, such as travel demand modeling, regional congestion monitoring and performance measurement, and travel/traffic data collection. Recent applications of Big Data solutions include the following:

- TPB purchased AirSage data in June of 2014 and used the data to enhance the external trip model validation for the TPB Version 2.3 Travel Model[3] (the regional travel demand model).
- TPB staff investigated the applicability of Google Application Programming Interfaces (API) travel time data to their modeling and planning functions[4]. TPB staff obtained Google travel time for vehicle and transit mode for a sample of areas within the region, verified them using travel demand model travel time predictions, and investigated regional traffic and transit travel patterns.
- INRIX data are being used for regional congestion monitoring and performance measures reporting.
- National Performance Management Research Data Set (NPMRDS) data are being used as specified in the federal Performance-Based Planning and Programming (PBPP) congestion and freight reporting and target-setting activities.
- A mobile device application was used to collect travel behavior data for a 7-day panel survey conducted in conjunction with the 2017-2018 Regional Travel Survey (RTS).
- TPB staff regularly summarize traffic sensor data as part of the support they provide to DDOT's Highway Performance Monitoring System (HPMS) program.

This evaluation will provide the TPB with necessary information and understanding of the strengths, limitations, applicability, and acquisition costs of various potential Big Data sources before committing to a significant investment in Big Data. It identifies specific Big Data products that have the potential to further enhance the technical analyses performed within the TPB program and informs critical decision making for regional transportation planning, performance measures, and programming. The evaluation will provide a roadmap for TPB to implement and procure Big Data products in addition to their current practice in Big Data solutions. It determines the feasibility of applying Big Data products in the following areas:

- Analyze regional travel and mobility to help validate data TPB collects through its conventional survey and traffic monitoring programs and to provide additional information and understanding of aspects of mobility that are not easily captured through conventional programs.
- Estimate, calibrate, and validate its regional travel demand model.
- Fulfill federal mandates for a regional congestion management process (CMP).
- Understand the characteristics and impacts of emergent travel options and technologies on travel throughout the region, including, but not limited to, the rapid rise in the use of TNCs, such as Uber and

[3] TPB internal technical memorandum: Year 2014 AirSage Trip Files and Data Processing.
[4] Feng Xie and Dusan Vuksan. *Potential Applications of Google-Based Travel Time Data in Transportation Planning Analyses.* Submitted to Transportation Research Record (TRR) for presentation and publication. August 2019

Lyft, traffic information and route navigation apps such as Google Maps and Waze, and real-time arrival and trip-planning for transit, ridesharing, bikesharing, and other forms of shared mobility.

- If the findings of this study result in a recommendation to obtain Big Data to conduct regional travel and mobility analyses, the TPB intends to apply Big Data analytics in answering key research questions on regional travel and mobility as well as travel demand forecasting.

1.1.1 TPB PROGRAMMATIC NEEDS

The TPB serves numerous functions and responsibilities including maintaining responsibility for the federally required metropolitan transportation planning process, serving as a forum for regional coordination, and providing technical resources for decision-making in the metropolitan Washington region. The TPB's primary responsibilities are listed below and shown in **Figure 1-1**.

- Long-Range Transportation Plan
- Transportation Improvement Program
- Performance Measures and Targets
- CMP
- Travel Forecasting
- Mobile Emissions Planning
- Socioeconomic Forecasts
- Regional Travel Trends
- Regional Plans and Planning Studies
- Continuous Airport Systems Planning

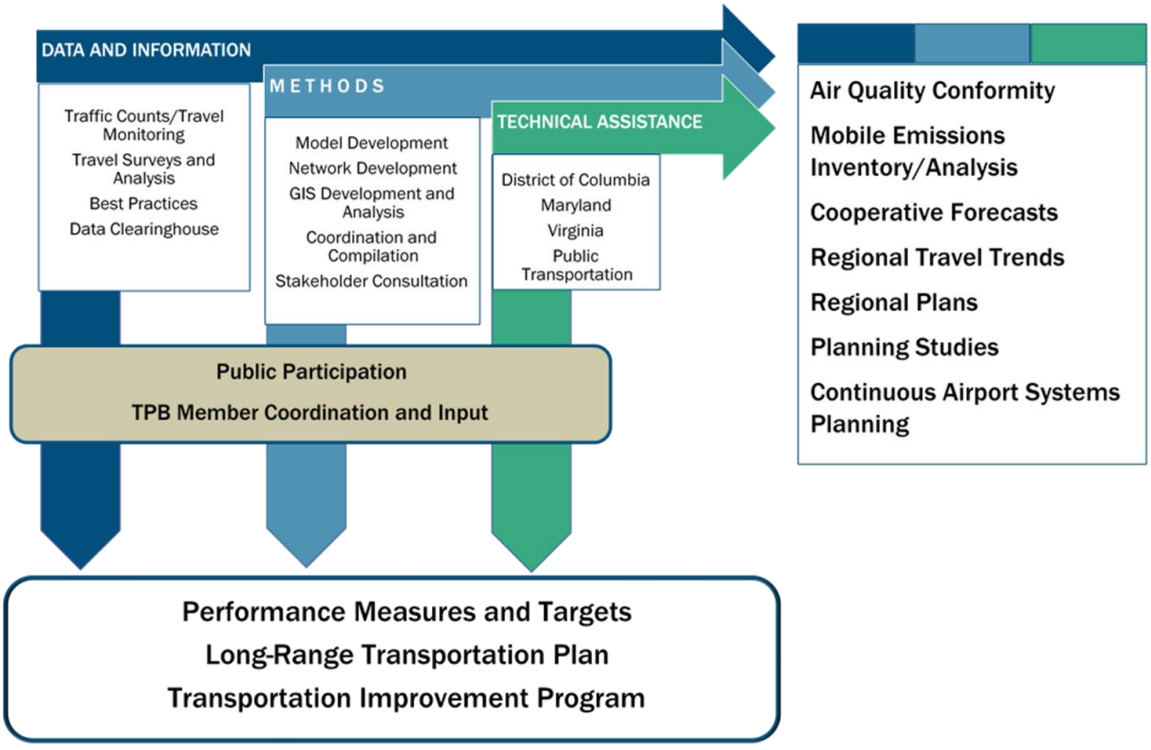


Figure 1-1 | TPB Core Programs and Data and Methods to Support These Programs[5]

[5] FY 2020 Unified Planning Work Program (UPWP)

The TPB's program needs are characterized by the Unified Planning Work Program (UPWP), which provides a platform to coordinate regional transportation planning activities. Transportation elements in the UPWP relevant to this study include the long-range transportation plan, the transportation improvement program, travel forecasting, the travel monitoring and data programs, system performance and congestion management programs, and the mobility and enhancement. The fiscal year (FY) 2020 UPWP describes all the TPB's federally assisted state, regional, and local transportation planning activities scheduled between July 1, 2019 and June 30, 2020. The FY 2020 regional planning priorities include:

- MAP-21 implementation: transition to performance-based planning and programming
- Regional coordination of transportation planning beyond traditional boundaries
- Ladders of opportunity: access to essential services

1.1.2 TPB RESEARCH AREAS

The TPB faces increasing needs for understanding local and regional travel behavior and mobility in support of their core functions to achieve the priorities set out in the UPWP and to address key regional multimodal mobility challenges. Integrating Big Data products into the TPB's programs has the potential to fill in gaps and enhance the efficiency, reliability, and validity of numerous efforts. These data products may introduce new data sources to supplement or replace existing sources of data and introduce previously unavailable types and magnitudes of data.

For this effort, the TPB project manager convened a study working group (SWG) composed of professional and technical staff within COG and TPB. The purpose of the SWG was to guide the focus and define the priorities for this independent evaluation of Big Data to ensure the resulting product would be beneficial for the TPB. The SWG identified eight core research areas for consideration in the Big Data evaluation. These research areas were: (1) travel demand forecasting, (2) travel demand management (TDM), (3) system performance and congestion management, (4) transit and active travel, (5) TNCs, (6) traffic counts, (7) CAVs, and (8) other Research Areas. To provide context for each of these research areas, the following sections summarize the TPB's existing programs, efforts, and needs within each research area.

TRAVEL DEMAND FORECASTING (TDF)

The TPB maintains the regional TDF model. This model provides a mathematical representation of the supply and demand for travel within the metropolitan Washington region. There are two major inputs into the model: (1) the transportation network representing all major roads and transit connections in the region and (2) land activity data for each transportation analysis zone (TAZ). The TPB travel forecasting model covers 6,800 square miles throughout three states (i.e., Virginia, Maryland, and West Virginia) and 22 jurisdictions/counties. The network is comprised of approximately 45,000 links and more than 600 transit routes. A single model run typically takes 30 hours of processing time on a high-end workstation. The cost of maintaining and applying this model constitutes almost half of the region's transportation planning budget, including data support (preparing and procuring data for model development, calibration, validation, and update).[6]

The current regional travel demand model is Version 2.3.78 released in April 2020. The model distinguishes travel patterns based on four time periods: (1) AM peak period, (2) midday, (3) PM peak period, and (4) nighttime/early morning hours. Multimodal trip generation within this model is conducted for five primary trip purposes: (1) home-based work, (2) home-based shop, (3) home-based other, (4) non-home-based work, and (5) non-home-based other. Three other trip purposes also are incorporated to represent commercial and freight trips: (1) non-freight commercial vehicles, (2) medium trucks, and (3) heavy trucks. Trips are then distributed through the model using the standard gravity model formulation to take into consideration roadway and transit travel times. A mode choice model is used to differentiate between 15 travel modes

[6] www.mwcog.org

including: (a) three auto modes (i.e., single occupancy vehicle, dual occupancy vehicle, three or more occupancy vehicle), (b) four transit modes (i.e., commuter rail, all bus, all Metrorail, combined bus and rail), and (c) three modes of access to transit (i.e., park and ride, kiss and ride, and walk). Traffic is assigned throughout the model using static user-equilibrium assignment.[7]

The latest model was calibrated and validated using the following data sets[8]:

- Census Data
    - 2000 Census Transportation Planning Products (CTPP)
    - American Community Survey (ACS)
- 2007/2008 TPB Household Travel Survey
- Transit on-board surveys
    - 2008 Metrorail survey
    - 2008 Regional bus survey
    - 2007-2008 On-Board survey of Maryland Transit Administration (MTA) riders, including users of the MARC train service
    - 2005 Virginia Railway Express (VRE) passenger survey
- 2005 COG TPB Commercial Vehicle Survey
- State and local government traffic counts
- 1994 COG TPB Auto External Survey
- 1996 COG TPB Truck Internal Survey
- 1996 COG TPB Truck External Survey

TPB is currently embarking on an effort to develop their Next Generation TDF Model, which will be developed as a simplified activity-based model (ABM). This model will have higher needs in data input compared to the previous model. Therefore, the TPB may consider Big Data products that could be used to supplement the data that is currently used to develop, calibrate, and validate their regional travel demand model.

TRAVEL DEMAND MANAGEMENT

The TPB TDM program area is focused on strategies that contribute to more effective and safer use of the region's transportation systems. As part of this program, a number of strategies have been developed[9]:

- **Alternative Commute Programs** | TPB actively encourages numerous alternative commute programs with an overarching goal of taking more cars off the road. These alternative commute programs include carpooling, vanpooling, telecommuting, transit, guaranteed ride home, bicycling to work, and outreach.
- **High Occupancy Vehicle (HOV) Facilities and Value Pricing** | TPB analyzes and documents the number of commuters using HOV facilities. In addition, TPB has studied a variety of pricing structures for managing congestion and environmental impacts as well as the public acceptability of pricing. These pricing studies have contributed to the MD 200 Intercounty Connector (ICC) and I-495/I-95 Northern Virginia Express Lanes projects.

[7] www.mwcog.org
[8] www.mwcog.org
[9] www.mwcog.org

- **Public Transportation Improvements** | In recognition of the importance of public transportation for reducing regional congestion, TPB compiles transit data to identify needs and locations for public transportation improvements.
- **Multimodal Facility/Pedestrian/Bicycle Improvements** | TPB is working to maintain, update, and implement new active transportation facilities to enhance multimodal travel within the region. The "Bicycle and Pedestrian Plan for the National Capital Region" includes approximately 410 bicycle and pedestrian facility improvement projects. Similarly, bike-sharing and car-sharing services are being explored as further means of reducing personal vehicle ownership, which in turn may further reduce congestion.
- **Growth Management** | At a regional level, TPB integrates lane use and transportation planning to promote transit and multimodal commuting options. Strategies related to growth management are developed from regional activity centers, Transportation-Land Use Connection (TLC) Program, and TPB's 2006 Regional Mobility and Accessibility Study. From these sources, evidence suggests that locating jobs and housing closer together can provide alternative commuting options that may reduce congestion.

In summary, the TPB is seeking an evaluation of Big Data products to identify which, if any, would be applicable to further enhancing these programs.

SYSTEM PERFORMANCE AND CONGESTION MANAGEMENT

For decades, the TPB has been exploring methods to efficiently monitor transportation system performance throughout the region. In 2005, the Safe Accountable Flexible Efficient Transportation Equity Act - A Legacy for Users (SAFETEA-LU) federal legislation mandated a CMP. This requirement was continued and expanded with the Fixing America's Surface Transportation (FAST) Act. Today, the CMP is fully incorporated into the region's long-range transportation plan, Visualize 2045.[10]

From the early 1990s through 2011, the TPB pursued a mobility-monitoring program using aerial survey technologies. Every 3 years, approximately 300 centerline miles of limited-access highway facilities in the metropolitan Washington region were surveyed. This effort provided vehicle counts and traffic quality ratings based on observed congestion. Since 2010, the TPB has been using INRIX data provided from the I-95 Corridor Coalition Vehicle Probe Project (VPP) to monitor system performance and congestion.[11] Today, TPB staff uses probe data analytics tools hosted by the Regional Integrated Transportation Information Systems (RITIS) at the University of Maryland to monitor system performance and congestion.

The TPB has identified the following congestion management strategies to expand capacity, reduce congestion, and mitigate the impact of congestion:

- **Demand Management Strategies** | This focus area aims to influence traveler behavior for the purpose of redistributing or reducing travel demand. Demand management strategies employed by TPB include: commuter connections program, promotion of local TDM, support for public transportation improvements, advocacy for pedestrian and bicycle facility enhancements, car sharing, and land use strategies.
- **Operational Management Strategies** | This focus area targets improvements to the transportation system that enables the network to continue functioning effectively. Operational management strategies employed by TPB include HOV facilities, variably-priced lane facilities, incident management, regional transportation operations coordination, and intelligent transportation systems (ITS).
- **Integrated/Multimodal Strategies** | This focus area targets the overlap between demand management and operational management to fully integrate these strategies. Integrated/multimodal

[10] www.mwcog.org
[11] www.mwcog.org

strategies include integrated corridor management, advanced traveler information systems, and the I-270 Innovative Congestion Management project.

- **Additional System Capacity** | Adding capacity is another congestion management strategy, which can eliminate bottlenecks, improve safety, and target specific traffic operational improvements. In comparison to the other strategies deployed, there have been relatively fewer capacity increase projects in recent years and emphasis has been placed on demand and operational management strategies.

In addition to leveraging the strengths of Big Data to understand the state of the regional transportation network during typical conditions, the TPB is interested in understanding the transportation impacts of major scheduled and unscheduled events. For example, during the 2018-2019 partial federal government shutdown, the board investigated the travel and congestion impacts within the metropolitan Washington region. To complete this analysis, staff used travel speeds collected by vehicle probe data from the University of Maryland's Probe Data Analytics (PDA) Suite. In the memorandum documenting this investigation, it was indicated that the breadth of this analysis was limited and the analysis would be enhanced with additional data that depict factors, such as traffic volume, weather, and additional commuting modes.[12]

TRANSIT AND ACTIVE TRAVEL

Transit and active travel are high priorities for the TPB. The TPB analyzes regional transit ridership trends and forecasts from the WMATA Metrorail system, Virginia and Maryland commuter railroad networks, and local/commuter bus networks throughout the region. The TPB brings together the operating agencies to coordinate long-term planning of major transit improvements and organize programs to promote greater use of transit. In addition, the TPB regularly brings transit operators together to share best practices for improving short-term operations and identifying long-term funding needs.[13]

The TPB has a dedicated bicycle and pedestrians subcommittee and this committee has developed a prioritized list of unfunded or partially funded bicycle and pedestrian projects throughout the metropolitan Washington region. The purpose of this list is to raise awareness of the projects and increase the likelihood these projects will be funded.[14] The TPB also provides a forum for local planners and decision-makers to share best practices for increasing active travel within the region. One of the major initiatives the TPB Planning Committee is working on is a National Capital Region Trails Network that will be made up of 61 miles of connected trails for biking and walking.[15]

Additionally, the TPB prioritizes the safety of vulnerable road users with numerous campaigns, program areas, and policies, including:

- **Street Smart Safety Campaign** | Vulnerable road uses (e.g., pedestrians and bicyclists) comprise more than a quarter of the region's traffic fatalities. The TPB developed this campaign with the goal of reducing the number of pedestrian and bicyclist injuries and deaths in the region.
- **Transportation Alternatives Program** | This program provides federal funds for small-scale projects, such as those encompassing pedestrian and bicycle facilities, trails, safe routes to school (SRTS), community improvements, and environmental mitigation. The TPB is responsible for selecting projects throughout the region to receive funding from this program.[16]
- **Complete Streets Policy** | The Complete Streets Policy was developed by TPB in 2012 with input from the Citizens Advisory Committee. As part of this policy, TPB defined a "complete street" as one that

[12] www.mwcog.org
[13] www.mwcog.org
[14] www.mwcog.org
[15] www.mwcog.org
[16] www.mwcog.org

safely and adequately accommodates motorized and non-motorized users, including pedestrians, bicyclists, motorists, freight vehicles, emergency vehicles, and transit riders of all ages and abilities, in a manner appropriate to the function and context of the facility.[17]

TPB works with the state DOTs in D.C., Maryland, and Virginia through two programs, the Technical Assistance Program (through its TLC Program)[18] and the Transportation Alternatives (TA) Set-Aside Program to fund bike and pedestrian projects. The TA Set-Aside Program provides federal funds for small-scale projects such as pedestrian and bicycle facilities, SRTS projects, streetscaping, and community improvements, and environmental mitigation)[19]. Additionally, it is part of the Federal Surface Transportation Block Grant Program, which was previously known as the Transportation Alternatives Program (TAP), and that name is still commonly used in some states.

TRANSPORTATION NETWORK COMPANIES

TNCs, or ride-hailing companies, such as Uber and Lyft, are transforming the way people travel in the region. The TPB is interested in understanding how this change in travel behavior is impacting network congestion and regional travel patterns. Regional partners, such as WMATA, also are interested in understanding how ride-hailing is affecting demand for transit. In their 2019 report compiling initial findings from the one-in-a-decade Regional Travel Survey, the TPB reported that TNCs are commonly used in the regional core, especially in the District of Columbia, where nearly one in five D.C. residents and one in 10 Arlington County residents use ride-hailing services at least once per week.[20]

Additionally, in 2019, the TPB conducted a study of the use of ride-hailing applications to access local airports as part of their Regional Air Passenger Survey. Data collected from more than 23,000 departing passengers at the region's three largest airports found decreased use of private vehicles, taxis, or Metrorail to access the airports. In total, 51 percent of travelers used private vehicles to access the airport, 24 percent of travelers used TNCs and, the remaining 25 percent of travelers used taxis, transit, walking, or bicycling.[21] The greatest effects were seen at the Ronald Reagan Washington National Airport (DCA), which is the closest airport to the District of Columbia. Between 2015 and 2017, taxi use decreased from 25 to 18 percent and TNC use increased from 14 to 21 percent.[22]

The TPB is interested in understanding how travel patterns (e.g., mode choice, trip destination, trip generation) are changing with the continued growth of TNCs in the region, but data is currently limited. However, the District of Columbia, per Code § 50–301.29a (general requirements for private vehicles-for-hire), is requiring ride-hailing companies to comply with a number of regulations to operate in the District, including a requirement to submit information including anonymized data regarding trips[23].

TRAFFIC COUNTS

The TPB relies on understanding travel demand and travel patterns throughout the region to build models, develop policies, and support programs. Traffic counts are an important subset of many different research areas within the TPB; therefore, the TPB would like to better understand how existing sources of traffic count data can be supplemented and validated. Additionally, traffic count data collection can be an intensive process and is often conducted for short durations; therefore, additional interest lies in how Big Data sources may be able to provide continuous traffic counts that could be used to compare travel changes

[17] www.mwcog.org
[18] www.mwcog.org
[19] www.mwcog.org
[20] www.mwcog.org
[21] www.mwcog.org
[22] www.mwcog.org
[23] code.dccouncil.us

hourly, daily, weekly, monthly, seasonally, and yearly. Aggregated traffic count data for the region is compiled by TPB and provided to the Regional Transportation Data Clearinghouse (RTDC)[24].

CONNECTED AND AUTOMATED VEHICLES

One of the TPB's Vision Goals is that "the Washington Metropolitan region will use the best available technology to maximize system effectiveness". One strategy listed to reach this goal is "maintaining international leadership in taking advantage of new technologies for transportation, such as automated highway systems and personal rapid transit".[25] While CAV technologies are not yet widespread, significant effort is being undertaken to develop and evaluate these technologies in controlled testing environments. In efforts to stay at the forefront of the industry, the TPB has already begun asking questions of how new CAV technologies may change travel patterns, travel demand, mobility services, and land use.

OTHER RESEARCH AREAS

The TPB identified a few other specific areas in which a Big Data solution may be feasible. These topics include: (a) investigation of freight and commercial vehicle travel throughout the region, (b) data that could be used to validate and supplement survey findings (e.g., National Household Travel Survey [NHTS] and Regional Travel Survey), (c) data that can be used to support analyses related to community-specific travel patterns and mode choice, and (d) data that could be used to inform transportation-related project selection at the TPB.

The TPB recently finished its two-part RTS[26]. The first part is a recruitment questionnaire, which asks about household characteristics in addition to new questions on alternative travel options that reduce the need for solo driving, such as ride-hailing and bikeshare. The second part is a travel diary in which participants record every trip made by members of their household on their assigned weekday such as school, work, errand, and recreation trips. TPB staff have begun analyzing the results from the first part of the survey and are currently in the process of performing quality control and analysis of the travel diary. Data from the travel diary will be released later in 2020.

Ongoing impacts of the COVID-19 pandemic on the region's transportation network is another topic of interest that is currently being investigated using Big Data. As the situation continues to evolve, the impact of the pandemic on travel demand, mode choice, and system performance, both for the short- and long-term, will be important for all agencies to understand.

SUMMARY OF DATA NEEDS BY RESEARCH AREA

A summary of the aforementioned research areas and the focus areas for identifying Big Data solutions were provided by the TPB SWG and are summarized in **Table 1-1**. The specific research needs are described in further detail in **Section 3.3**.

Table 1-1 | Research Areas for Evaluation

Research Area		Description
1	Travel Demand Forecasting	• Origin-Destination (O-D) information by mode with trip purpose/destination/origin type • Region-wide speeds and volumes • Region-wide signal delay estimation/queue length

[24] rtdc-mwcog.opendata.arcgis.com
[25] www.mwcog.org
[26] www.mwcog.org

Research Area		Description
2	Travel Demand Management	• Estimating network demand using historical data • Survey of employer telework policies • Forecasts of future telework policies
3	System Performance/Congestion Management	• Monitoring network-wide congestion and performance • Detailed network performance data for strategy development • Event traffic data
4	Transit and Active Travel	• Impact of mode choice on network level of service (LOS) (scenario analysis) • Monitoring real-time transit performance • Monitoring active travel demand and delay
5	Transportation Network Companies	• Improved understanding of usage and demand for TNCs: temporally, geographically, O-D patterns • Socioeconomic and demographics of TNC riders • Relationship between TNCs and transit (e.g., first/last mile problem)
6	Traffic Counts	• Obtaining permanent/temporary traffic count data • Variability analysis of traffic counts (e.g., day-by-day) • Methods to validate traffic counts collected in traditional methods • Vehicle classification data
7	Connected and Automated Vehicles	• Impact of CAVs to vehicle occupancy • Impact of CAVs on multimodal travel demand • Impact of CAVs on infrastructure (e.g., parking) and revenue streams • Available data from CAVs
8	Other Research Areas	• Regional freight and commercial vehicle travel forecasting • Household survey data validation and supplementation • O-D patterns based on population demographics • Improved project selection

1.2 RESEARCH METHODOLOGY

An independent evaluation of Big Data and its use and limitations in regional transportation planning applications was completed for the TPB. The consultant team worked closely with the TPB SWG to clearly identify the TPB's data needs within each of the agency's research areas prior to beginning the evaluation. The evaluation explored the strengths, limitations, applicability, and acquisition costs of common Big Data products used by transportation agencies. The evaluation included a detailed state-of-practice review consisting of a literature review, peer agency survey and interviews, and Big Data vendor interviews. Each Big Data product was evaluated against a scoring rubric developed by the consulting team and the SWG. The scoring rubric used seven evaluation criteria to evaluate whether each product would be an effective investment for the TPB based on the defined research areas and data needs. Once the evaluation was completed, additional feedback was gathered from the SWG to further refine the consultant team's understanding of the most critical needs for Big Data at the TPB. Next, meetings were conducted with local peer agencies (i.e., VDOT, DDOT, DRPT, WMATA, Maryland Department of Transportation – State Highway Administration [MDOT-SHA]) to learn about each agency's experience using Big Data, their greatest needs for Big Data, and discuss potential collaboration in procuring new Big Data products. Using the insight gleaned from each step of this process, the consultant team recommended a pilot procurement plan for the TPB to trial one or two Big Data products that offer promise for multiple data needs. Lastly, an overview of the evaluation and partner agency feedback was presented to TPB leadership in February 2021. The complete evaluation methodology is illustrated in **Figure 1-2**.

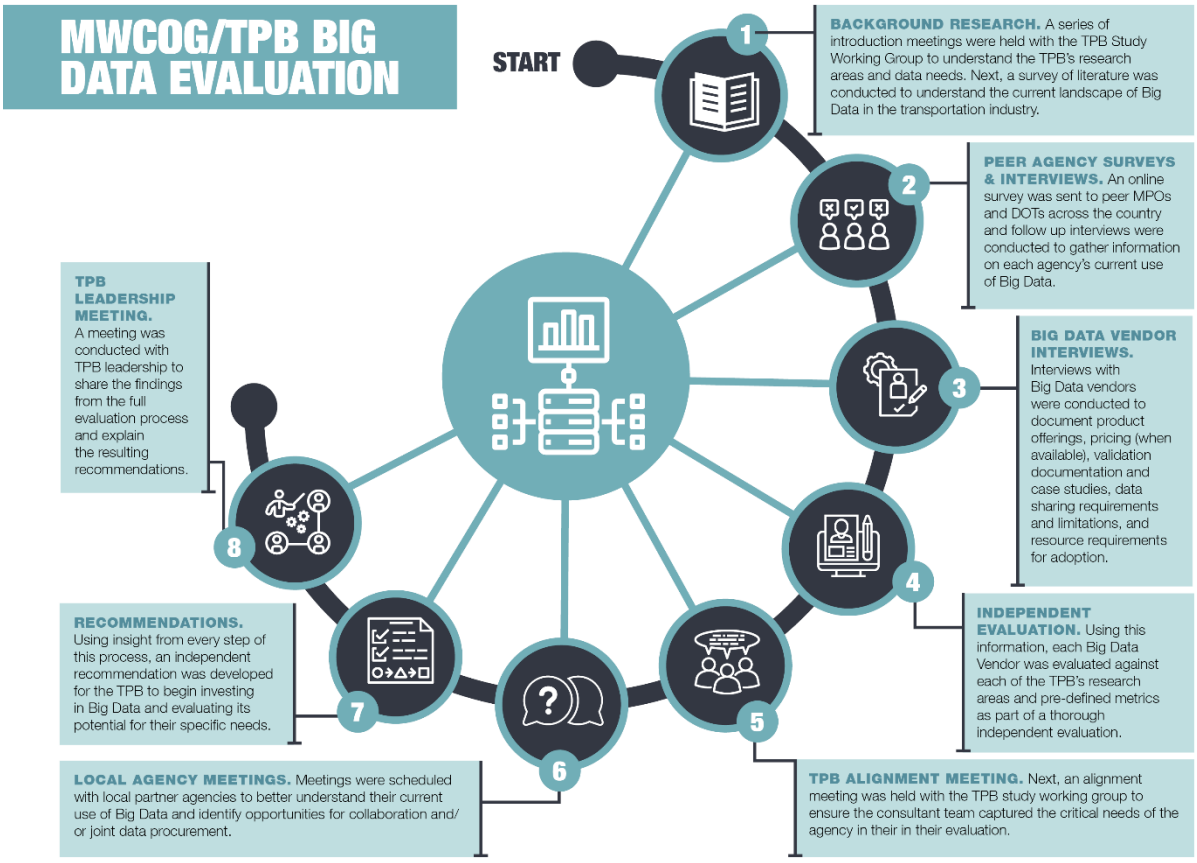


Figure 1-2 Independent Evaluation Process

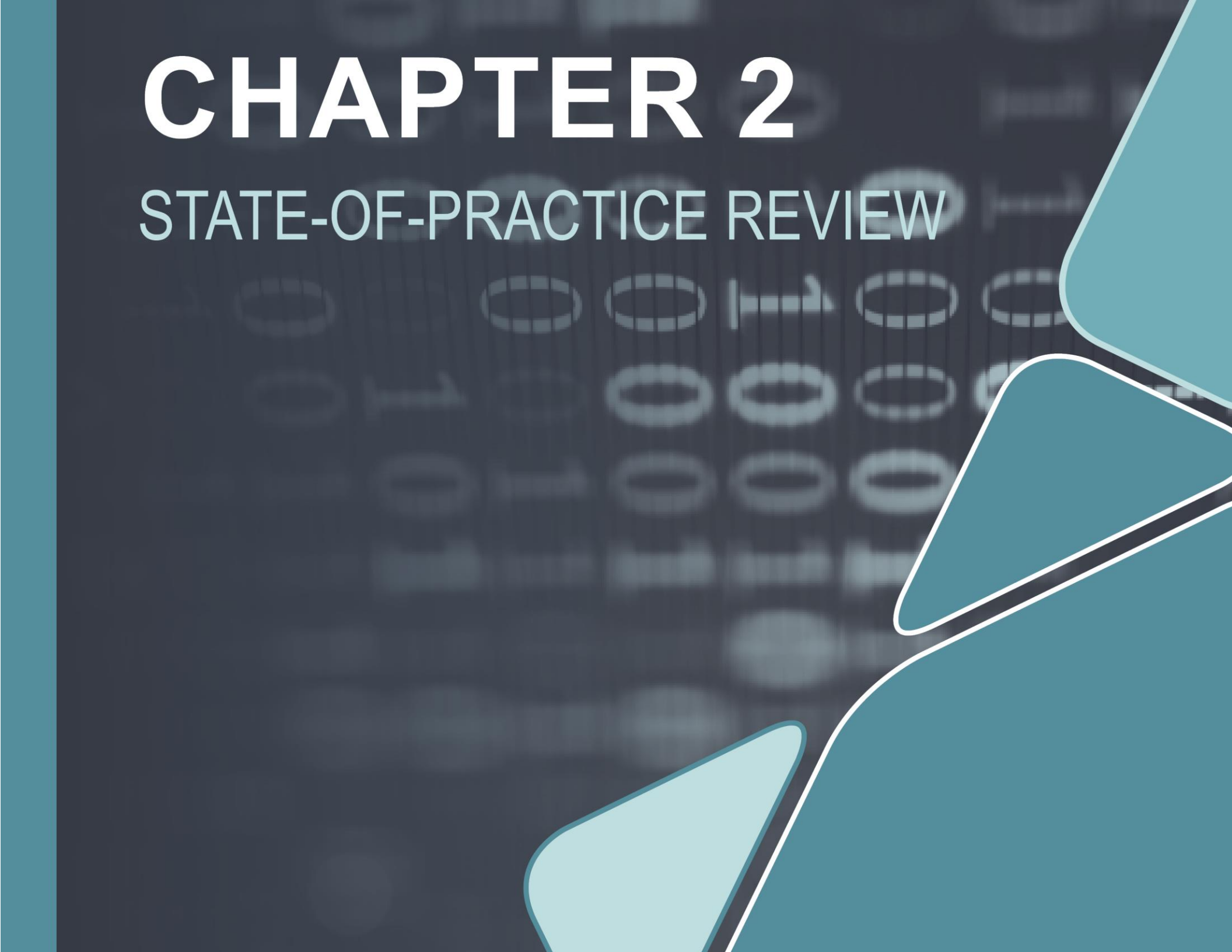
1.3 REPORT OUTLINE

Each chapter of this report is summarized below.

- **Chapter 1 Introduction and Purpose** | Project background, purpose, and research methodology and provides a summary of the TPB's identified research areas for the evaluation.
- **Chapter 2 State-of-Practice Review** | Findings from transportation agency surveys and interviews originally presented to TPB in the Task 2 State-of-Practice Review Memorandum.
- **Chapter 3 Big Data Product Evaluation** | Findings from the independent evaluation of Big Data products. Details of the evaluation methodology are provided, and the evaluation results are presented in multiple ways to meet the diverse needs of the TPB and agency partners.
- **Chapter 4 Recommendations** | The ultimate recommendation and procurement plan resulting from the independent evaluation process, and a discussion of promising Big Data solutions for future consideration and research.
- **Appendix A |** Comprehensive summary of peer agency survey and interviews.
- **Appendix B |** Comprehensive evaluation results for all evaluated vendors in reference to all research areas.
- **Appendix C |** Summary of the local partner agency meetings.
- **Appendix D |** A comprehensive menu of Big Data solutions and recommended options.

CHAPTER 2

STATE-OF-PRACTICE REVIEW



2 STATE-OF-PRACTICE REVIEW

2.1 REVIEW METHODOLOGY

The state-of-practice review involved two primary components: (1) nationwide agency online survey with follow-up phone interviews, and (2) local agency meetings. The survey was designed using the Qualtrics online survey tool with questions developed to elicit responses from peer agencies that would indicate their experience using Big Data in application areas related to the TPB's established needs. The survey was sent to a total of 35 peer agencies, and responses were received from 21 agencies; these agencies are listed in **Table 2-1**. Feedback from agencies electing to remain anonymous are included in the findings; however, identifier information has been removed from this report. The original survey questions and responses from each of these agencies are provided in **Appendix A**.

Of the 21 agencies that responded to the online survey, nine peer agencies were identified for follow-up phone interviews based on their responses, geographic location in relation to the metropolitan Washington region, and comparable metropolitan regional size. The intention of these phone interviews was to gather more information and clarify responses to survey questions. Each interview was scheduled for one hour and agencies were encouraged to share as many thoughts as they wished during the interview. Questions were structured to gather insight on each peer agency's experience using Big Data for application areas related to the TPB needs. The interview was conducted in a hybrid fashion to address all major questions identified from the survey feedback and enable fluid conversation based on the interviewed agency's direct experience with Big Data. Of the nine interviews requested, eight peer agencies were interviewed, and their responses were used to further inform this state-of-practice review. The peer agencies interviewed as part of the state-of-practice review are indicated in **Table 2-1**.

Table 2-1 | Online Survey Respondents

No.	Agency Name	Metro Area	State(s)
1	Atlanta Regional Commission (ARC)*	Atlanta	GA
2	Chicago Metropolitan Agency for Planning (CMAP)*	Chicago	IL
3	Delaware Valley Regional Planning Commission (DVRPC)	Philadelphia	PA, DE
4	East-West Gateway Council of Governments (EWGCOG)	Saint Louis	MO
5	Fredericksburg Area Metropolitan Planning Organization (FAMPO)	Fredericksburg	VA
6	Houston-Galveston Area Council (HGAC)	Houston	TX
7	Maricopa Association of Governments (MAG)*	Phoenix	AZ
8	Metro (Portland, Oregon)*	Portland	OR
9	Miami-Dade Transportation Planning Organization [Florida Metropolitan Planning Organization Advisory Council (MPOAC)]*	Miami, Tampa, Orlando	FL
10	Mid-Ohio Regional Planning Commission (MORPC)	Columbus	OH
11	New York Metropolitan Transportation Council (NYMTC)*	New York	NY
12	North Central Texas Council of Governments (NCTCOG)	Dallas/Fort Worth	TX
13	North Jersey Transportation Planning Authority (NJTPA)	Newark	NJ
14	Northeast Ohio Areawide Coordinating Agency (NOACA)	Cleveland	OH
15	Pima Association of Governments (PAG)	Tucson	AZ
16	Regional Transportation Commission of Southern Nevada (RTCSNV)	Las Vegas	NV
17	Southeast Michigan Council of governments (SEMCOG)*	Detroit	MI
18	Southern California Association of Governments (SCAG)*	Los Angeles	CA
19	Virginia Department of Transportation (VDOT)	Virginia statewide	VA

No.	Agency Name	Metro Area	State(s)
20	Wasatch Front Regional Council (WFRC)	Salt Lake City	UT
21	Anonymous**	Anonymous	-

** Peer agencies who participated in a follow-up phone interview.*
*** Respondents were given option of not identifying themselves although they could provide a response.*

After gathering input from agencies nationwide, meetings were set up with the TPB's local partner agencies to better understand each agency's biggest need(s) related to Big Data and learn more about their current Big Data investments, experience, and lessons learned. These meetings also served to query interest and potential for pooled funding of Big Data procurements or sharing of data resources across the region. The agencies in which the TPB held meetings are provided in **Table 2-2**. A comprehensive summary of these meetings is provided in **Appendix B**.

Table 2-2 | Local Partner Agency Meeting Participants

No.	Agency Name	State(s)
1	District of Columbia DOT (DDOT)	DC
2	Virginia DOT (VDOT)	VA
3	Maryland DOT State Highway Administration (MDOT-SHA)	MD
4	Washington Metropolitan Area Transit Authority (WMATA)	DC
5	Virginia Department of Rail and Public Transportation (DRPT)	VA

2.2 SURVEY AND INTERVIEW FINDINGS

The objective of the state-of-practice review was to gain insight into how peer agencies are currently using (or not using) Big Data in transportation planning. Specifically, *which Big Data products are being used, what are they being used for, how are they acquired and processed, and what challenges are agencies facing with their adoption?* The following sections summarize the findings related to each of these questions from both the surveys and interviews.

2.2.1 WHICH BIG DATA PRODUCTS ARE AGENCIES USING?

Using the survey data, a matrix was developed to pair each surveyed agency with the Big Data products they are using; this information is provided in **Table 2-3**. Nearly all the peer agencies who responded to the survey indicated they use some form of Big Data.

The agencies who use a large number of Big Data products (i.e., defined as those using 10 or more products) include Atlanta Regional Commission (ARC), Chicago Metropolitan Agency for Planning (CMAP), and Miami-Dade Transportation Planning Organization (Miami-Dade TPO). From the responses, a correlation was identified between agency size and the number of adopted Big Data products. In general, larger agencies were more likely to use multiple Big Data products than smaller agencies. This finding is likely related to agency needs and available resources.

The top three Big Data products being used by the surveyed peer agencies are listed below.

- National Performance Management Research Data Set (NPMRDS)/INRIX
    - NPMRDS provides INRIX speed and travel time data for all National Highway System (NHS) roads.
    - NPMRDS data is used widely as it can be procured by agencies for free. However, most agencies noted that the network coverage of this data set is not extensive enough to support many applications.
- Disaggregate Census Data
- StreetLight Data

Additional Big Data products commonly used by peer agencies include government records, InfoUSA (Data Axle USA) and Dun & Bradstreet Business Listings, HERE, and continuous traffic counts. It is noted that HERE was previously the source data for NPMRDS, that is mostly commonly used by metropolitan planning organizations (MPOs).

2.2.2 HOW ARE AGENCIES USING BIG DATA PRODUCTS?

In addition to understanding which Big Data products have been procured by peer agencies, the survey inquired how agencies were using Big Data once procured. A summary of the number of agencies using various procured Big Data products for different application areas are provided in **Table 2-4**. These areas matched TPB's original research needs that were carried out for the state-of-practice review phase of the study and reorganized during the evaluation phase of the study.

Table 2-3 | Big Data Products Adopted by Surveyed Agencies

Surveyed Peer Agencies	Disaggregate Census data product	Government record	InfoUSA (Data Axle USA)/ Dun & Bradstreet business listing	Google Places	Co-Star data	LIDAR and other airborne data	Job postings data	NPMRDS /INRIX*	HERE	StreetLight Data	American Transportation Research Institute truck GPS data	Google Travel Time	Transit Automatic Vehicle Location	Farecard/ Toll Data	Social media	Continuous traffic data	Other data sources
Atlanta Regional Commission	X	X	X		X	X		X	X	X	X	X	X	X		X	X
Chicago Metropolitan Agency for Planning	X	X	X		X	X		X	X		X					X	X
Delaware Valley Regional Planning Commission								X									
East West Council of Governments (Missouri)		X	X			X		X	X			X					
Fredericksburg Area Metropolitan Planning Organization (Virginia)								X	X	X					X		
Houston-Galveston Area Council	X	X	X					X		X	X			X		X	X
Maricopa Association of Governments (Florida)	X	X						X	X	X	X	X				X	X
Metro (Portland, Oregon)	X					X		X	X	X		X				X	
Miami-Dade Transportation Planning Organization	X		X	X				X		X	X	X					
Mid-Ohio Regional Planning Commission	X	X	X	X				X	X	X		X	X			X	
New York Metropolitan Transportation Council	X							X		X							
North Central Texas Council of Governments	X	X		X				X	X	X	X	X			X		
North Jersey Transportation Planning Authority	X	X	X		X			X									
Northeast Ohio Areawide Coordinating Agency	X			X				X		X							
Pima Association of Governments (Arizona)	X	X	X	X		X		X		X							X
Regional Transportation Commission of Southern Nevada	X		X					X	X		X		X				
Southeast Michigan Council of Governments	X	X						X		X						X	X
Southern California Association of Governments	X	X	X					X		X	X			X		X	X
Virginia Department of Transportation	X		X				X	X	X	X						X	
Wasatch Front Regional Council (Utah)	X	X							X			X				X	
Anonymous Agency	X	X		X				X		X	X	X				X	
Number of Agencies Using Data	**18**	**13**	**11**	**6**	**3**	**5**	**1**	**20**	**11**	**15**	**9**	**9**	**3**	**3**	**2**	**11**	**7**
Percent of Total Surveyed	**86%**	**62%**	**52%**	**29%**	**14%**	**24%**	**5%**	**95%**	**52%**	**71%**	**43%**	**43%**	**14%**	**14%**	**10%**	**52%**	**33%**

Includes Big Data Products specified in the survey.
This list represents the agencies who responded to this question of the survey.
**The survey focused on NPMRDS that is currently based on INRIX.*

TRAVEL DEMAND MODELING

Travel demand modeling is a common need across most of the surveyed peer agencies. Big Data solutions offer multiple streams of data that are frequently used to supplement these efforts. The majority of the surveyed agencies use Big Data products for travel demand modeling efforts. These agencies suggested Big Data originating with mobile devices can fill in the gaps for travel demand model validation at subregional and corridor levels, as these Big Data products can be acquired at a relatively cost-effective price, compared with field data collection. A list of specific use cases identified for Big Data products in travel demand modeling is provided below.

- Origin-Destination (O-D) travel data, including underlying information such as trip purpose and traveler demographics, is used as an input to develop travel demand models.
    - StreetLight and AirSage appear to be the most commonly used Big Data products for regional and corridor level O-D pattern analysis. The Big Data products are often used by the peer agencies (e.g., SEMCOG, MAG, ARC and SCAG) to validate the O-D patterns from traditional data sources (e.g., household travel survey). However, such validation is mostly conducted at highly aggregate level (e.g., super district that consists of a cluster of TAZs) and usually only for minimum reasonableness checking since there is hardly any evidence that could determine which data source is more reliable when differences exist. Using Big Data for O-D validation is more likely to support consensus among different data sources.
- Regional socio-economic data is used as an input to travel demand models.
    - InfoUSA (Data Axle USA) and Dun & Bradstreet were identified as sources for external travel attributes and other attributes used in model calibration.
- Behavioral data is used in estimating community activity levels and activity center spatial data for developing activity-based models.
- Trip and service operations data is used to validate travel demand model.
    - Transit operational data (e.g., Farecard, automated passenger counter [APC] and automatic vehicle location [AVL] data) were used to better understand and validate the travel pattern by transit modes and the transit operations (e.g., transit travel time, on-time service and service reliability).
- Traffic data is used to estimate and validate speeds in travel demand models for both vehicular traffic and transit.
    - INRIX and Google travel time data were cited as being used to validate speed estimates for both vehicles and transit in travel demand modeling.
- Freight data is used to estimate and validate freight performance in travel demand models.
    - American Transportation Research Institute ATRI) freight data is cited as being used to support freight performance validation in model development.
- Several agencies suggested Big Data is used to better understand managed lane usage by time-of-day.

Table 2-4 | Number of Surveyed Agencies using Big Data Products for Specific Application Areas

Application Areas		Disaggregate Census data product	Government record	InfoUSA (Data Axle USA)/ Dun & Bradstreet business listing	Google Places	Co-Star data	LIDAR and other airborne data	Job postings data	NPMRDS /INRIX*	HERE	StreetLight Data	American Transportation Research Institute truck GPS data	Google Travel Time	Transit Automatic Vehicle Location	Farecard/ Toll Data	Social media	Continuous traffic data	Other data sources
Travel Demand Modeling	Regional and corridor level O-D patterns	9	4	2	2	1	1	0	7	3	4	4	3	0	1	0	2	3
	Managed lane time of day usage	2	1	1	1	0	0	0	2	1	1	1	2	0	2	0	3	0
	Trip purpose and destination type	8	3	2	2	0	0	1	4	1	3	1	2	0	1	0	1	2
	External travel attributes	5	4	5	3	0	0	0	4	2	4	5	2	0	0	0	3	2
	Other model calibration/validation	14	7	6	5	1	0	1	17	8	7	2	6	3	2	0	8	5
Transportation Network Companies (TNC)	TNC O-D characteristics, (e.g., time/location)	2	1	1	2	0	0	0	3	1	1	0	1	0	0	0	0	1
	TNC rider demographics	1	1	0	1	0	0	0	2	1	1	0	1	0	0	0	0	0
	Assess the impact of TNC on other modes and traffic congestion	2	0	1	1	0	0	0	2	1	0	0	1	0	0	0	0	0
Travel Demand Management (TDM)	TDM policy evaluation	4	3	0	2	0	0	0	3	1	3	0	1	0	0	0	1	4
	Telecommuting trends	4	2	0	0	0	0	0	1	1	2	0	1	0	0	0	0	1
Traffic Counts	Day-to-day traffic variation	7	0	1	1	0	0	0	7	4	0	1	3	0	1	0	7	2
	Vehicle classification	5	1	0	2	0	0	0	2	1	1	2	1	0	0	0	7	0
System Performance Monitoring	System performance monitoring/evaluation	6	2	0	2	0	0	0	16	6	2	1	4	1	1	0	5	2
	Impacts and trend of congestion management strategy	5	1	0	1	0	0	0	8	4	1	1	2	0	0	0	3	1
Other applications	Freight and commercial vehicle travel	5	1	1	2	0	0	0	5	2	1	9	2	0	0	0	2	0
	Inter-city bus travel	4	0	0	1	0	0	0	2	1	0	0	2	0	0	0	0	0
	Travel behavior of special population group	5	3	1	1	0	0	0	3	1	3	0	2	0	0	0	0	0
	Traffic dynamics and interactions with transit	3	0	0	1	0	0	0	2	1	0	0	2	0	0	0	0	0
	Project selection	1	0	0	1	0	0	0	4	3	0	0	1	0	0	0	3	1
	Data validation and integration of Household travel survey	6	1	1	1	0	0	0	3	2	1	1	2	1	1	0	1	1
	Other (economic development, etc.)	2	5	5	2	2	4	0	2	2	5	0	4	0	1	2	0	2

TRAVEL DEMAND MANAGEMENT

Every agency has its own flavor of travel demand management that is highly dependent on the makeup of the regional community. Managing demand can be related to strategic development growth, multimodal commute options, and policies encouraging new travel behaviors. Many of these strategies are being supplemented with data. Datasets and products being used to better understand and interpret travel behavior and travel patterns by the agencies surveyed include:

- Static socioeconomic and demographic data are used to better understand and interpret regional travel behavior and travel patterns.
    - Disaggregated census data
    - Government records (Quarterly Census of Employment and Wages [QCEW], Longitudinal Employer Household Dynamics survey [LEHD][27])
- O-D and trip generation information are important for managing demand.
    - AirSage and StreetLight are cited as common sources of O-D data.
    - High-resolution global positioning system (GPS) data is collected from a third-party application[28].

SYSTEM PERFORMANCE MONITORING AND TRAFFIC COUNTS

Monitoring system performance is crucial for all agencies. Numerous surveyed peer agencies revealed that Big Data solutions are currently being used to support these efforts. Big Data products derived from mobile-devices are commonly used in this application area because these sources are considered by many agencies as suitable for reporting network-wide performance trends over time and comparing the performance of facilities within a region.

- Traffic data is used to:
    - Monitor regional transportation network performance (TNP) and monitor congestion
    - Estimate average annual daily traffic
    - Estimate multimodal O-D travel patterns

The major product cited by agencies is NPMRDS (input data is from INRIX). Other major vendors that provide this kind of data are HERE, StreetLight, and ATRI (for truck data).

TRANSPORTATION NETWORK COMPANIES

TNCs, such as Uber, Lyft, Via, and others, are transforming travel in dense urban areas. While beneficial in reducing private vehicle ownership, the impact of TNCs on the transportation network is largely unknown. Big Data products are a primary mechanism for investigating the effect of TNCs on congestion, mobility, and transit.

- O-D data of TNC-related trips can be used to identify hotspots where TNC pick-ups and drop-offs are most prominent in the region.
    - Custom analyses are available from StreetLight to obtain TNC-trip specific O-D data.

In the survey responses, numerous agencies expressed interest in better understanding TNC travel patterns and network impacts; however, in most cases Big Data products were not yet established and

[28] Metropia App www.metropia.com

adopted for these purposes. This sentiment was carried through many of the interviews with peer agencies. A few discussed their desire to identify evaluation strategies that leverage the TNC data available through regional agreements with Uber, Lyft, and taxi companies, such as NYMTC.

OTHER APPLICATION AREAS

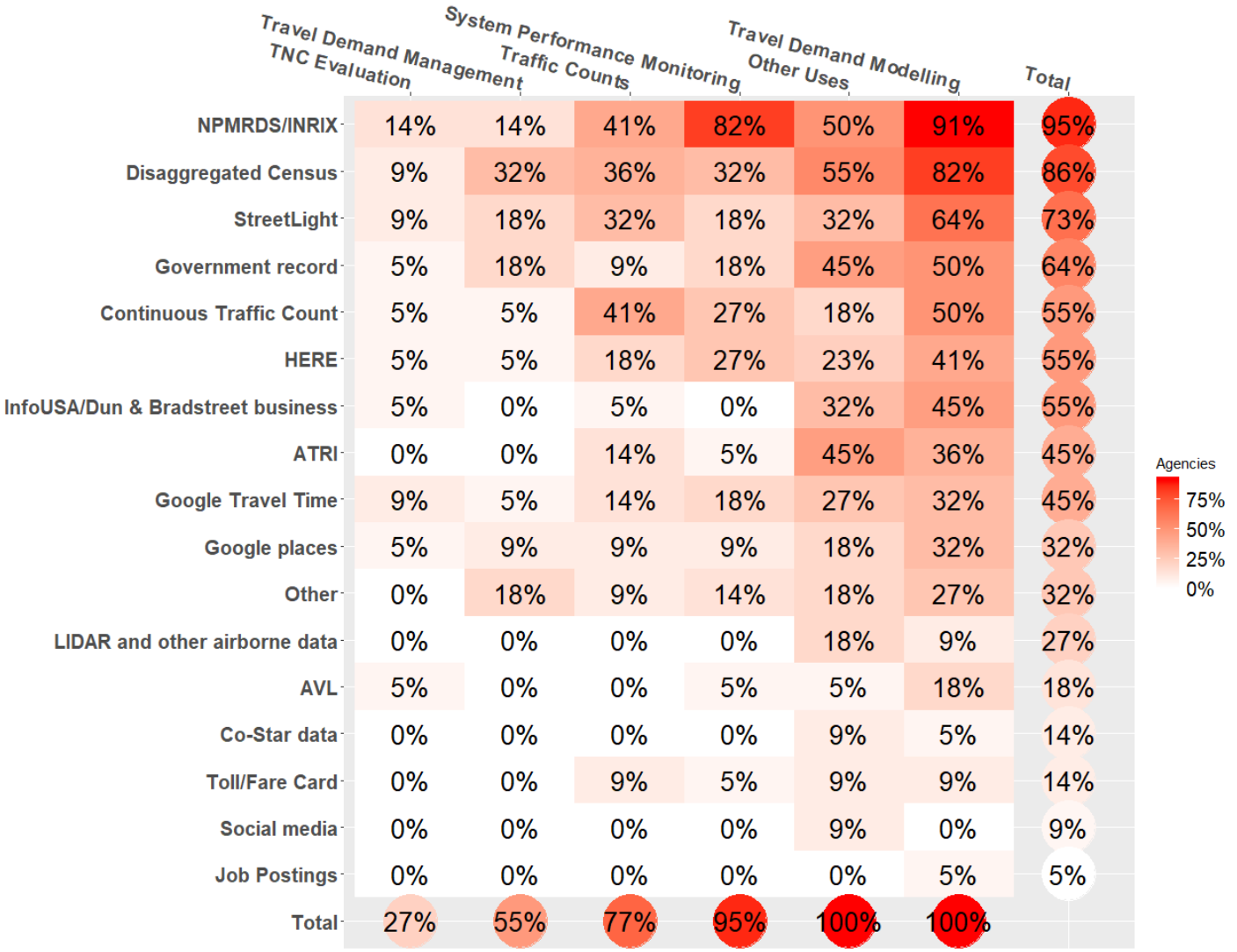
As shown in **Figure 2-1**, peer agencies are implementing Big Data solutions to support additional application areas, including transit, active and sustainable travel planning, economic development, socioeconomic modeling, lane use modeling, freight analysis, and speed validation.

- Transit farecard and toll data streams are used to infer demographic distribution and transit travel patterns.
- Social media data issued for public involvement.
- Use of socioeconomic and location-based data to identify the location of freight generators.
- Lidar and other forms of aerial-based data are being used for watershed modeling and construction, urban heat island analysis, land use modeling, and elevation analyses.
- Agency asset management databases (e.g., regional sidewalk inventory and current land use conditions) and other similar data sources are being used for additional analyses, including active and sustainable travel planning.

The major data product categories cited by agencies for “other applications” during the survey are Disaggregate Census data, NPMRDS, Government Records, and ATRI.

SUMMARY OF AGENCY BIG DATA USE

Figure 2-1 provides a matrix pairing Big Data products with common application areas. Within each cell of this figure, the percentage of surveyed agencies using the data product for the specific application area is indicated. For example, 91 percent of peer agencies are using NPMRDS/INRIX data for travel demand modeling. The application areas are shown in an ascending order of propensity for Big Data solutions.



** The survey focused on NPMRDS that is currently based on INRIX.*
*** Traffic counts application includes count and speed data.*
**** Agencies surveyed reported no applications of Big Data products in connected and automated vehicles (CAV).*

Figure 2-1 | Application Areas Benefiting from Big Data Products

2.2.3 HOW ARE AGENCIES ACQUIRING BIG DATA PRODUCTS?

The growing popularity of Big Data solutions in the transportation industry is calling for new data management and procurement strategies. In the survey responses, three primary acquisition strategies for Big Data products were identified:

- Collected in-house or acquired via partnerships with other agencies in the region (e.g., federal, state, and local governments)
- Purchased directly from a vendor with a one-time purchase or subscription supported solely by the peer agency interviewed
- Purchased directly from a vendor with a one-time purchase or subscription in sharing costs with partner agencies in the region

The method for acquiring Big Data products is largely dependent on the type of product and source of information. Some of the commonly used Big Data products by surveyed peer agencies are publicly available (e.g., disaggregate census data), whereas others require purchase directly from a data vendor. When purchasing data from a vendor, there is often the option to obtain a single set of historical data (e.g., all data from 2019) or a subscription to gather updated data in real-time (e.g., all data from 2020 as it becomes available). Purchasing fees and subscriptions also may include the analytic platform provided by the vendor for data query and manipulation. Further, multiple vendors offer opportunities for regional agencies (e.g., MPOs, state transportation agencies, and local transportation agencies) to share costs for access to region-wide data. Sometimes procurement costs are shared among agencies; other times, these costs are covered by a larger agency and shared with smaller agencies.

Table 2-5 denotes the acquisition strategies adopted by each peer agency. Each percentage represents the frequency that acquisition strategy was used to procure Big Data products for their agency. Product acquisition mechanisms under the "other" category primarily include web-downloads from state, county, and local agencies.

Table 2-5 | Big Data Product Acquisition Strategies by Surveyed Agencies

Surveyed Peer Agencies	Collected In-house	Purchase/ Subscription	Shared Cost	Other (specify)
Atlanta Regional Commission	25%	69%	6%	0%
Chicago Metropolitan Agency for Planning	0%	36%	0%	64%
Delaware Valley Regional Planning Commission	33%	33%	33%	0%
East West Council of Governments (Missouri)	86%	0%	0%	14%
Fredericksburg Area Metropolitan Planning Organization (Virginia)	25%	0%	25%	50%
Houston-Galveston Area Council	22%	78%	0%	0%
Maricopa Association of Governments (Florida)	50%	42%	8%	0%
Metro (Portland, Oregon)	29%	14%	57%	0%
Miami-Dade Transportation Planning Organization	33%	50%	17%	0%
Mid-Ohio Regional Planning Commission	0%	40%	0%	60%
New York Metropolitan Transportation Council	50%	0%	0%	50%
North Central Texas Council of Governments	40%	33%	20%	7%
North Jersey Transportation Planning Authority	50%	33%	0%	17%
Northeast Ohio Areawide Coordinating Agency	100%	0%	0%	0%

Surveyed Peer Agencies	Collected In-house	Purchase/ Subscription	Shared Cost	Other (specify)
Pima Association of Governments (Arizona)	30%	50%	10%	10%
Regional Transportation Commission of Southern Nevada	17%	50%	17%	17%
Southeast Michigan Council of Governments	0%	17%	0%	83%
Southern California Association of Governments	56%	33%	0%	11%
Virginia Department of Transportation	33%	33%	33%	0%
Wasatch Front Regional Council (Utah)	0%	0%	17%	83%
Anonymous Agency	30%	50%	10%	10%
Average	**34%**	**32%**	**12%**	**23%**

2.2.4 WHAT TOOLS ARE USED TO PROCESS BIG DATA?

Survey results showed that most agencies must post-process the obtained Big Data feeds to derive actionable information specific to their application areas. Three primary types of processing tools were identified by the surveyed agencies:

- Spreadsheet tools, such as Microsoft Excel
- Database tools, such as GIS, Microsoft Access, and SQL
- Custom computer scripts, using programming languages such as R and Python as well as application programming interfaces (APIs) for querying data directly from a vendor

The tool selected for data analysis and processing is highly dependent on the data type and the format of the received data. For example, data that is geospatially referenced is difficult to interpret in basic spreadsheet tools and would likely need to be processed first through data analytics software and custom analytical scripts or a tool such as GIS to ascertain meaning. Similarly, aggregated summary data sets can be easily analyzed in a spreadsheet tool, whereas raw disaggregate data often are too large for spreadsheet tools and require the use of more advanced data processing tools (e.g., custom computer scripts).

Table 2-6 provides a summary of the data processing tools used by agencies to process their acquired Big Data sources. Each percentage represents the frequency that processing tool was used to process their respective Big Data products.

Table 2-6 | Big Data Analysis Tools Used by Surveyed Agencies

Agency	Spreadsheets (e.g., MS Excel)	Database tools (GIS)	Custom Computer Scripts
Atlanta Regional Commission	16%	44%	41%
Chicago Metropolitan Agency for Planning	5%	50%	45%
Delaware Valley Regional Planning Commission	33%	33%	33%
East West Council of Governments (Missouri)	43%	29%	29%
Fredericksburg Area Metropolitan Planning Organization (Virginia)	33%	33%	33%
Houston-Galveston Area Council	40%	60%	0%
Maricopa Association of Governments (Florida)	31%	31%	38%

Agency	Spreadsheets (e.g., MS Excel)	Database tools (GIS)	Custom Computer Scripts
Metro (Portland, Oregon)	40%	30%	30%
Miami-Dade Transportation Planning Organization	50%	50%	0%
Mid-Ohio Regional Planning Commission	31%	31%	38%
New York Metropolitan Transportation Council	25%	50%	25%
North Central Texas Council of Governments	36%	64%	0%
North Jersey Transportation Planning Authority	44%	44%	11%
Northeast Ohio Areawide Coordinating Agency	60%	20%	20%
Pima Association of Governments (Arizona)	27%	36%	36%
Regional Transportation Commission of Southern Nevada	0%	83%	17%
Southeast Michigan Council of Governments	43%	43%	14%
Southern California Association of Governments	25%	0%	75%
Virginia Department of Transportation	36%	43%	21%
Wasatch Front Regional Council (Utah)	46%	31%	23%
Anonymous Agency	17%	33%	50%
Average	**32%**	**40%**	**28%**

2.2.5 WHAT CHALLENGES ARE AGENCIES FACING IN USING BIG DATA PRODUCTS?

Peer agencies were surveyed about the challenges they have faced using Big Data. Specifically, agencies were asked to rate the degree of difficulty in the following areas:

- Information technology infrastructure needed to acquire, store, and access data
- Project and task timeline constraints
- Project and task budget constraints
- Data processing and interpretation complexity
- Quality assurance and quality control requirements
- Data validity and available sample size
- Data availability and coverage

Peer agency responses to these questions are summarized in **Table 2-7**.

Table 2-7 | Reported Level of Difficulty Using Big Data Products

Data Set/Application	Infrastructure*	Task Timeline	Budget	Complexity	Other**
Disaggregate Census data product (Specify e.g., PUMS)	4	4	4	4	2
Government record (Specify e.g., appraisal business licensing QCEW)	3	5	3	4	2
InfoUSA (Data Axle USA)/Dun & Bradstreet business listing	3	6	4	5	2
Google Places	4	5	4	6	2
Co-Star data	2	4	4	3	1

Data Set/Application	Infrastructure*	Task Timeline	Budget	Complexity	Other**
LIDAR and other airborne data	4	5	4	5	1
National Performance Management Research Dataset (NPMRDS)/INRIX	4	5	3	5	1
HERE	4	5	4	5	0
StreetLight Data	3	5	3	4	2
American Transportation Research Institute (ATRI) truck GPS data	4	5	5	5	2
Google Travel Time	4	5	4	5	2
Transit automatic vehicle location	5	3	2	5	0
Farecard/Toll Data (specify)	3	3	2	3	1
Social media (specify, Facebook, LinkedIn, etc.)	2	2	2	2	1
Continuous traffic data (e.g., high-resolution signal automatic traffic recorders)	4	5	4	5	2
Other data sources	4	5	5	5	2
MODE	4	5	4	5	2

Scores represent averages calculated from survey respondents.
Scores represent a scale from 1 to 10, with 10 being the greatest difficulty.
** Infrastructure refers to the IT Infrastructure needed to acquire, store, and access data (severs, network, etc.)*
*** Other comments included quality assurance/quality control (QA/QC), validity of data (sample size), accuracy, and coverage (time and facilities)*

The most frequently cited challenges were timeline constraints and data complexity, followed closely by technological infrastructure requirements and budget constraints. Agencies were encouraged to elaborate on their responses, and in doing so, suggested that the degree of difficulty in each area is highly dependent on the type of Big Data and the application area itself. Further explanations of these challenges and others also were provided and are summarized below.

DATA VALIDITY, COVERAGE, AVAILABILITY, AND QUALITY ASSURANCE

- Many Big Data products appear promising for meeting agency needs; however, since many sources of Big Data are still relatively new, limited external validation is available to support their accuracy.
- Big Data products offer a lot of information that was not previously available; however, they are not comprehensive. For example, many Big Data vendors pull information from mobile devices, which provides only a sample of the full population. Data coverage and sample size is an important characteristic of each Big Data product and must be considered prior to use. This challenge is often amplified in regions with lower population density, such as rural areas.
- Challenges and concerns related to the quality of Big Data sources were raised. Emphasis on the importance of rigorous quality assurance was discussed, especially in the travel demand modeling application area.
- One agency indicated that in their experience, the Big Data products (e.g., StreetLight) often bias analyses, as the time coverage may not be sufficiently broad to reveal true yearly, seasonally, weekly, and daily travel pattern variation.
- Numerous agencies expressed concern with the validity and the overall quality of Big Data products. These concerns largely developed from the lack of transparency in how data samples were collected, processed, and inflated by data vendors (e.g., StreetLight and AirSage).
- Several agencies suggested that O-D data stemming from Big Data sources should only be used at high geographic aggregation levels (e.g., districts comprised of a cluster of transportation analysis zones [TAZs]) for the following reasons:

  - Data sampling is not random, and the sampling time frame is not transparent; therefore, multiple agencies question the sample inflation or expansion procedures.
  - Big Data vendor's definitions of trip purpose often do not directly align with those used in many regional travel demand models.
  - Big Data products typically lack vehicle classification information at a detailed level.

- A handful of agencies highlighted their efforts to validate big data products, such as StreetLight O-D data versus the Household Travel Survey (HTS). One agency suggested that while discrepancies between the data sources were observed, it was not possible to ascertain which was more accurate due to their unique limitations. Another agency compared average annual daily traffic (AADT) and vehicle classification data from StreetLight and AirSage and concluded that the data was not at an acceptable level of accuracy for their use yet. Most agencies that have conducted some form of comparison analysis concluded that Big Data products were not mature enough to replace data collected using traditional methods.
- The required minimum sample size and representation of the traveler population for some application areas (e.g., development of a travel demand model) is often not met with Big Data products.
- Big Data vendors maintain their edge on the market with proprietary means of processing raw data sources. However, the lack of transparency in data collection and processing methodologies limits the certainty agencies can have in the data source's reliability.

BUDGET CONSTRAINTS

- Budget constraints were rated at a varying degree of difficulty, also dependent on the Big Data product itself. Two products highlighted for their high costs were the Census and National Household Travel Survey (NHTS). The Census is completed once every 10 years, whereas the NHTS is typically conducted every 5 to 7 years. Although the Census and NHTS data is free to MPOs and state and local agencies, the cost associated with conducting these traditional travel surveys is high and often out of reach for smaller agencies if they ever want to conduct an independent survey at a regional or local level on a more frequent basis than what is administered at the national level. Even at the federal level, there are gaps in the NHTS results due to budget constraints.
- Budget constraints and overall concerns come into play with newer sources of Big Data obtained through product vendors. Challenges were suggested by the high cost of acquiring these data as well as difficulty interpreting opaque pricing models.
- Big Data pricing models also were cited as influencing the feasibility of adoption. For instance, high upfront costs were prohibitive for some agencies.
- One agency suggested its budget constraints were alleviated by making use of additional funding sources (e.g., Transportation Management Area fund).

DATA PROCESSING AND INTERPRETATION COMPLEXITY AND INFORMATION TECHNOLOGY INFRASTRUCTURE

- Data processing and interpretation can be time intensive due to the complexity and size of the incoming data stream.
- Specialized skills are needed to analyze some forms of raw data. These skills include the ability to develop and run specialized scripts to digest and process large datasets into an aggregated format that can be used to address the application area needs.
- While these specialized skills may be needed for many data sources, many agencies suggested that the technical challenges of processing this data was minimal because their existing staff is equipped with these skill sets, largely stemming from use of traditional data sources requiring spatial, statistical, and other analyses.

- Fusion of different types of data also was mentioned as a challenge. One specific example was related to challenging procedures required to georeference segment-based travel time and speeds from vendors such as INRIX and HERE for integration with spatially referenced data in geographic information system (GIS) or travel demand models.
- No significant analytical effort has been conducted to review TNC data, even though data from Uber, Lyft, and taxi companies are becoming more commonly available to local transportation agencies. A few agencies stated they were actively looking into strategies to leverage the data, whereas another agency discussed the challenge with having limited in-house expertise to process the TNC data available to them.
- The often differing definition of “trip purpose” between emerging Big Data vendors and traditional data streams also is a challenge for traditional transportation analysis and travel modeling.
- In discussing the promising advances in traffic count collection (e.g., traffic signal detection and surveillance data), one agency highlighted an ongoing IT challenge of data archiving, processing, and reporting.

2.3 LOCAL AGENCY MEETING FINDINGS

Overall, the meetings with local agencies raised many of the same key elements identified in the nationwide survey of agencies and follow-up interviews. A few of the common themes from these discussions are provided below. A summary of each local agency meeting can be found in **Appendix C**.

- **The Black Box** | Many of the agencies talked about the barrier of the "Black Box". They described that it is not only difficult to not know the underlying algorithms and data sources, however, it is also challenging to keep up with the regularly evolving evaluation methodologies and data sources that make it difficult, if not impossible, to compare data year-over-year.
- **New Realities |** Implementing Big Data involves both information technology resources, new strategic hiring strategies, and staff training programs. The information technology needs include data processing, data storage, and cloud computing as well as implementing data dashboards and visualization. The agencies also discussed the importance of integrating new team members with data science backgrounds as well as providing additional training opportunities for staff with more traditional experience for their agencies.
- **Primary Data Needs |** Throughout these meetings, the discussion converged on two primary data needs: Origin-Destination Data (by mode, trip purpose) and Traffic Performance Data (speeds and volumes). These two primary data needs meet the desire for multi-modal data, which allows agencies to better understand active travel patterns and provides insights into transit travel patterns as well.
- **Data Granularity |** Agencies indicated a smaller appetite for less granular data (e.g., census block groups) as these low-resolution data sources are not sufficient for answering common questions at the agency. Rather, they expressed interest in data resolutions smaller than traffic analysis zones (TAZ) and census-group levels that provide insight on trip purpose and traveler attributes.
- **Interest and Barriers to Shared Procurement |** The agencies shared an interest in pooling resources in the metropolitan Washington region. Potential barriers to shared procurement include the cost and funding due to the COVID-19 pandemic, fiscal year planning and early budget allocation, procurement methods (some agencies requiring competitive process), and scope of data (data type and region).

2.4 REVIEW SUMMARY

Big Data products are still relatively new to most transportation agencies. This state-of-practice review suggests that agencies maintain a high level of interest in Big Data due to the application potentials as well as a healthy level of skepticism on the quality of the data. The majority of peer agencies queried as part of this state-of-practice review are leveraging traditional data sources (e.g., Census data and travel surveys) as well as emerging Big Data products to address their agency needs. Big Data products from mobile devices are the most frequently used, with common vendors including INRIX and StreetLight. The prominent application areas for these Big Data products were travel demand modeling and system congestion monitoring.

In most cases, peer agencies did not suggest that Big Data products were replacing existing data streams; rather, these emerging data sources were being used to supplement and validate existing methods and programs. The challenges agencies are facing in the adoption of Big Data products range from ensuring sufficient data validity and quality and overcoming budget and timeline constraints to technical requirements for storing and processing Big Data. Despite these challenges, several agencies found the insights offered by Big Data valuable to their organizational goals and plan to continue using multiple sources of Big Data for model validation and other application areas.

CHAPTER 3

BIG DATA PRODUCT EVALUATION

3 BIG DATA PRODUCT EVALUATION

3.1 EVALUATION METHODOLOGY

The purpose of this report is to conduct an independent evaluation of Big Data products by appraising their use and limitations for application in eight of the transportation planning board's (TPB's) study working group (SWG) research areas. An illustration of the evaluation methodology is provided in **Figure 3-1**.

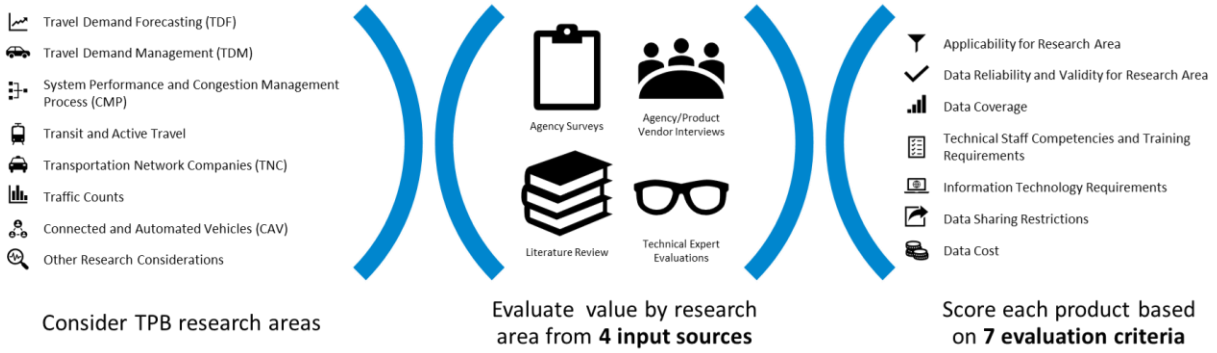


Figure 3-1 | Evaluation Methodology

For each of the TPB's research areas, selected Big Data products (summarized in **Table 3-1**) were evaluated for potential value in each research area and ultimately given a "yes" or "no" recommendation based on findings from each input source (i.e., agency surveys, agency interviews and meetings, literature review, and technical expert evaluations). These input sources were used to score each Big Data product evaluated based on seven evaluation criteria, which were developed from the research objectives. These were applicability for research area, data reliability and validity for research area, data coverage, resource requirements for technical competencies and information technology, data sharing restrictions, and data cost.

The seven evaluation criteria are described in detail in **Table 3-2**. Each Big Data product was considered separately for each research area; therefore, a product could be recommended for one research area and not another. Detailed explanations, guidance, and caveats for each recommendation are provided. For each research area, a scoring matrix was developed to rate each data product against each evaluation criterion based on the needs of the specific research area. The sections summarizing the Big Data product information and research area information used to develop these tables are provided in **Appendix B**. From the independent review of each research area, a simplified recommendation summary matrix was developed to summarize which data products are recommended for which research areas. This summary table is provided in **Chapter 4 Recommended Options**.

Table 3-1 | Big Data Products Evaluated

Product Category	Big Data Product	Product Summary
Mobile Device-Based Travel Patterns Data	AirSage	This is a mobile device analytics product containing insights derived using cell tower triangulation, rather than location-based service/global positioning system (LBS/GPS); as such, spatial granularity is limited to the "polygon" level and not to individual roadway links. At this time, AirSage offers a Nationwide Trip Matrix product which allows users to purchase origin-destination (O-D) trip tables (extrapolated to the entire population) at the county level as well as products geared toward tourism and hospitality users to understand visitor origins and demographics.
	Google Travel Time	This product provides travel time data between specific points on the roadway network. Google provides credits for pulling (and storing) travel time data based on the frequency of data pulls and the number of locations being queried. This product may be best used on a corridor-level based on the pricing structure.
	HERE Technologies	This is a provider of traffic analytics such as speeds and travel times similar to INRIX. Link speeds are available in 5-minute granularity bins for passenger vehicles and trucks. HERE previously offered a Trip Data analytics product with O-D data, but this product has been discontinued.
	INRIX	INRIX is a provider of vehicle probe data for segment-level congestion analytics as well as O-D's for customized zones. The underlying data sources are a combination of LBS, GPS from local delivery fleets and long-haul trucks, and connected vehicles (e.g., Audi/BMW); INRIX has been increasing their data providers to increase their penetration rate for passenger vehicles as well as temporal resolution (e.g., pings every 3 to 5 seconds).
	Locus (Cambridge Systematics)	Locus is built through a partnership with PlaceIQ for the underlying LBS dataPset (mainly from cell phone apps). This product is essentially a combination of consultant services and software; it is a customizable/tailored solution. Analyses are not constrained by the available inputs and options on an online platform.
	Replica	Replica is an online platform providing travel patterns data "replicating" real-world travel through two modules – one at an aggregate level and one at a detailed, disaggregate level. Replica's first module, Trends, delivers weekly aggregate-level (state/city/census tract) trend data on mobility (including mode choice), consumer spend, and COVID-19-related information based on mobile location data integrated with underlying contextual data. The second module, Places, provides detailed travel pattern data simulated by their Activity-Based travel model built upon mobile device location data, publicly available datasets, select modeling geographies and various ground-truth data. The Places module is currently available for more than a dozen select regions, including the Metropolitan Washington Council of Governments (MWCOG) region, and is planned to be expanded to cover most of the 48 continental states in 2021. This module provides information down to individual network links/transit routes/transit stops, as well as individual census block groups or transportation analysis zones (TAZs).
	Strava Metro	Strava Metro is an aggregator of bicycle and pedestrian movements using data obtained from the Strava user mobile app (for tracking bicycle/running/walking trips). Strava Metro is their product offering for planners and provides the relative level of activity (separated into bicycle/pedestrian) along various facilities. Strava does not have an O-D product offering.
	StreetLight	StreetLight Data is an online platform for O-D or segment-based analytics based on mobile device data. StreetLight's underlying mobile data sources are mainly LBS-based (cell phone apps); INRIX is one of their underlying data sources. This mobile device data is integrated with underlying contextual data such as census demographics to provide additional insights
	Streetlytics	This is a mobile device analytics platform offered by Citilabs, the vendor for Cube software. According to their website, they offer O-D data, trip purpose and modal breakdown, volume and speed data broken down by hour and time of day, and routes between origins and destinations. This product appears to be a partnership with AirSage.
	Taxi/TNC Trip Data	The District of Columbia has aggregated data for private vehicles-for-hire, such as taxis and TNCs, including Uber and Lyft, as information is required to be reported to Department of For-Hire Vehicles (DFHV) and District Department of Transportation (DDOT) under D.C. law. A dashboard has been created to visualize the trip data, including the magnitude of TNCs trips, temporal distribution of TNCs trips, average wait-time (request-to-pickup), and number of unique operators.
	Teralytics	Teralytics is an online platform for O-D analytics based on cell phone tower triangulation data. Cell tower triangulation has a high sample rate (estimated at 15-35 percent of the population) but a lower spatial resolution. Thus, this data is not usable for route-level analysis (individual trips cannot be mapped to specific roads), but it is usable at the census tract or even TAZ level. Segment-level traffic count estimates (average annual daily traffic [AADTs] or turning movement counts) are not estimated via this platform.
	Uber Movement	This dataset provides aggregated zone-to-zone travel time data derived from Uber trips freely available for the metropolitan Washington area, but only at the TAZ (District of Columbia only) or census tract (roughly the area contained by the I-495 Capital Beltway) level. This online tool allows a user to select a date/time range and pick

Product Category	Big Data Product	Product Summary
		origin/destination zones to view travel times between those zones. Uber also recently began offering a beta “mobility heatmap” product showing color-coded links throughout the District based on the density of Uber-owned “mobility devices” (e.g., Uber Jump bikes and scooters).
	Unacast	This is a mobile device analytics provider generally focused on location-specific insights and marketed toward private-sector/“specific place” applications. The research team’s current understanding of this product is that its transportation planning-focused applications have been in the aviation sector (e.g., understanding ground access travel to/from an airport).
	Verizon Data	Verizon has begun offering a series of smart cities-related products, including a product for traffic analytics. Their traffic data product involves “mapping” the cellular network for a given roadway or area to create a “digital signature” and capture a very high percentage of devices traveling along a corridor anonymously.
	Waze	Partner cities in Waze’s Connected Citizens Program (CCP; free for a city to join) have access to Waze data, including a historical feed, via Google Cloud. This feed provides the instances of “Waze jams” and incidents along roadways.
Socioeconomic or Location-based Data	CoStar Data	CoStar is a database of commercial real estate transactions. CoStar data can be used to help understand the real estate development patterns in the region, provide input and parameters to land use modeling, and derive employment type and activities.
	Disaggregate Census Data	The US Census Bureau publishes large demographic datasets on travel behavior and population demographics. Decennial Census data (e.g., 2020 Census) is usually made available by the Census Bureau less than a year after the survey. Various demographic/socioeconomic estimates, including the land use data used by the regional travel demand models, are usually benchmarked against decennial Census data once it becomes available. The census bureau also makes American Community Survey (ACS) data at various geographic levels and disaggregate Public Use Microdata Sample (PUMS) data available for public use. Census Transportation Planning Products (CTPP) is another variation of ACS products that is customized using multiple years of ACS data for transportation planning/modeling applications.
	InfoUSA (Data Axle USA) and Dun & Bradstreet Business Listings	InfoUSA (Data Axle USA) and Dun & Bradstreet business listing data are among the most commonly used data products to support the development of employment databases for both regional land use, travel demand modeling, and economic development activities. Both InfoUSA (Data Axle USA) and Dun & Bradstreet data products provide detailed establishment location-based business information including employment size and industry sector.
	Google Places	Google Places provides information on location and type of places such as offices, parks, restaurants, and transit stops. This product offers an alternative approach to verify and supplement the business listing data acquired from the other sources (e.g., InfoUSA (Data Axle USA), Dun & Bradstreet, and Quarterly Census of Employment and Wages (QCEW), particularly about the presence, location, and type of businesses to better inform regional land use and travel demand models.
	Household Travel Survey	The Household Travel Survey (HTS) reveals insightful information with behavioral characteristics of travelers and the relationship between travel decisions and travelers’ demographic backgrounds. Traditionally, the HTS collects the data that pertains to household, person, vehicle, and trip characteristics on selected travel day(s) and usually is conducted every 5 to 10 years by metropolitan planning organizations (MPOs), state DOTs, and the Federal Highway Administration (FHWA) through the National Household Travel Survey (NHTS) program.
	MapBox	Mapbox offers three types of data products to support map-based illustration and analytics: MapBox Boundaries, MapBox Traffic Data, and MapBox Movement.
	Others	Other federal agencies also provide a variety of socioeconomic data products, such as the series of National Longitudinal Survey of Youth as well as many downloadable data tables in the areas of inflation and prices, employment, unemployment, employment projections, pay and benefits, spending and time use, productivity, workplace injuries, occupational requirements, regional resources, and international from the Bureau of Labor Statistics (BLS). However, with the exception of the QCEW dataset, which may provide location-based employment and wage information in a disaggregate fashion for almost any customized geography through special agreement with the state Department of Economic Security (DES), most of the data products made available by BLS and other federal agencies (e.g., Bureau of Economic Analysis [BEA]) are only intended to help analyze the national trend and by no means could be seen as a source of Big Data for local and regional applications and studies.
Data from Public Infrastructure	Automated Traffic Signal Performance Measures (ATSPM)	ATSPM data products contain performance measures for traffic signals based on high resolution (up to 1/10th second) event data. This data, which includes detector calls, phase changes, and transit signal priority/preemption, can be used to provide analytics on operations at individual signals or along entire arterial corridors.
	Continuous Traffic Count Station Data	Local and state DOTs typically have a series of permanent devices set up for collecting traffic counts., These devices also may collect vehicle classifications and spot speeds. From this data, agencies can estimate the average daily number of vehicles traversing roadway segments.

Product Category	Big Data Product	Product Summary
	Transit Data from On-Board Intelligent Transportation Systems (ITS) Devices	Transit agencies employ a variety of ITS devices for monitoring the status of their system and improving system performance. These products typically consist of devices on board the transit vehicles, communications from the vehicles to a central system (sometimes in real-time), and a backend server to store and process data from these devices.
End-User Platforms for Data Analytics	Regional Integrated Transportation Information System (RITIS)	RITIS from the Center for Advanced Transportation Technology (CATT) Lab at the University of Maryland is an “automated data fusion and dissemination system that provides an enhanced overall view of the transportation network. Participating agencies are able to view transportation and emergency management information through innovative visualizations and use it to improve their operations and emergency preparedness”.
	Moonshadow/DB4IoT	Moonshadow is a provider of connected vehicle data analytics. Through a partnership with underlying providers such as INRIX and wejo, they have developed a database and visualization suite called DB4IoT. This product serves as both a backend (data storage/processing engine for raw data from vehicles and infrastructure) and a frontend (dashboard for visualizing data).
	Moovit	Moovit is a mobility software company that offers both user-facing products (e.g., rider-trip planning tools) and analytics for transit agencies. They offer two main products: Moovit Urban Mobility Analytics (MUMA) and Moovit Mobility as a Service (MaaS). Due to their multimodal focus, they have as much experience with MPOs as they do with transit agencies.
	Swiftly	Swiftly is a vendor of transit analytics using feeds from agency systems; they claim to be “hardware-agnostic and software-centered” for consuming and analyzing feeds from various automatic vehicle location (AVL) providers. They assist public transit agencies in enhancing their transit service by analyzing on-time performance and identifying operational issues.
Emerging Big Data Sources	CAV Data	CAVs are anticipated to have major impacts on the transportation industry; however, the current stage of testing and deployment are currently limited as technology and policy mature. CAV data streams are currently coming from original equipment manufacturers (OEMs), scenario-based planning platforms, and pilot and test facilities.
	Micromobility Data	During the past few years, a variety of private companies have deployed fleets of scooters and bikes that can be rented via mobile apps. Transportation planners across the country have desired to have data on the use of these micromobility services to better understand their use and interactivity with other transportation modes. While many of the companies behind these services are unwilling to share much data, viewing this information as a trade secret, many jurisdictions have implemented contracts with these services that require sharing of some trip information, such as total numbers of trips, aggregated origins and destinations, or even information on individual trips.

Table 3-2 | Evaluation Criteria

Evaluation Criteria	Scoring Mechanism	
	Positive	**Negative**
Applicability*	The Big Data product can adequately serve the TPB's needs and interests within this research area. It is suitable for answering important questions asked by this research area. It can be used to address routine concerns over time.	The Big Data product is currently unable to serve the TPB's needs and interests within this research area. It is likely unsuitable for answering important questions required of this research area. It may address a few current concerns, but its application would be in a one-time effort rather than routine usage.
Data Reliability and Validity*	The Big Data product has well documented data validation methodology and the data is proven statistically and widely used in research and industry applications. The product has been well-vetted as a mature data source. Data integrity is ensured over time.	The Big Data product does not have a well-documented data validation methodology due to the proprietary nature of data or due to it being a new maturing product. There is an uncertain time-integrity: product quality and availability could change over time.
Data Coverage*	The Big Data product has comprehensive coverage in the analysis region. This can be measured in terms of proportion of population covered in sample, time periods offered, and frequency of data collection (seconds, minutes, hours, real-time). If data coverage is less extensive, the Big Data product can demonstrate significant generalizability (the results there apply everywhere) or transferability (the results there also apply here).	The Big Data product lacks comprehensive coverage in the analysis region. Such a product may cover only a small percentage of the analysis region, it may have few time periods offered for analysis, or the data may be too infrequent or not offer enough granularity. The Big Data product may not lead to generalizable or transferable conclusions.
Resource Requirements (Technical Staff Competencies and Training Requirements)**	The Big Data product is user-friendly. The product vendor may provide its own software or web interface for manipulating the data. Additional data manipulation can be done in basic data analytics software, such as Microsoft Excel or GIS. Learning to use the data product would not require significant staff augmentation and training.	The Big Data product is not easily used without specialized understanding of the data, analytical skills, and software knowledge. Complexities with the data may require data analytics be completed in more complex data analytics software. Learning to use the data product may require significant staff augmentation and training or specialized knowledge.
Resource Requirements (Information Technology)**	The Big Data product can be used with existing information technology (IT) infrastructure. Usage of the product would not require consistent IT upkeep and maintenance.	The Big Data product requires significant computation power above existing IT capabilities. Usage of the product data would require consistent maintenance for data upkeep and storage. The data product may require the adoption of in-house databases and/or cloud hosting services.
Data Sharing Restrictions**	The Big Data product can be easily shared with TPB stakeholders and partners. License restrictions do not prevent sharing raw data or access to the data product's web-interface.	The Big Data product cannot be easily shared with the TPB stakeholders. License restrictions prevent the sharing of raw data and access to the data product's web-interface. Only findings or processed results from data analytics can be shared externally.
Data Cost**	The Big Data product has relatively low costs associated with procurement, subscriptions, and data/software downloads. The data may be available for little or no cost via agreements with local transportation agencies.	The Big Data product has relatively high costs associated with procurement, subscriptions, and data/software downloads. The pricing structure is opaque and may change frequently.
Overall Recommendation*	**Yes | Recommendation that the TPB consider this Big Data product.**	**No | Recommendation that the TPB <u>not</u> consider this Big Data product.**

** Specific to each unique research area.*
*** Specific to each data product, consistent between all research areas.*

From the insight gleaned in the evaluation, a menu of potential Big Data product procurement options was developed and provided to the TPB for feedback. With this feedback, an implementation plan was developed. This plan includes specific investment strategies, resource requirements (i.e., technical staff competencies, training requirements, and IT needs), and procurement methods specific to the recommended vendors and products. Additional recommendations are provided to summarize additional non-procurement next steps.

The findings of the independent evaluation of Big Data products for the TPB research areas are provided in the following sections. These findings are presented in multiple ways to meet the needs of TPB and the agency partners.

- A summary of all evaluated Big Data products is provided in **Table 3-1.**
- The comprehensive research area tables with all evaluated Big Data products alongside the score for evaluation criteria are provided in **Appendix B.**
- **Section 3.2: Specifics of Big Data Products** provides a summary of the most promising data products based on the criteria for this evaluation, and a brief summary of the corresponding applicable research areas. Reference this section for detailed product-specific information including an overview of the product; validation documentation and case studies; cost and pricing structure; data licensing and sharing, and data storage; and processing and analysis considerations.
- **Section 3.3: Applicable Big Data Products for the TPB Research Areas** provides a summary of the needs of each research area alongside the corresponding most promising data sources for that research area. Reference this section for the consolidated list of applicable Big Data products for each research area and to understand what options exist for a specific research need. Product details can then be referenced in **Section 3.2**.
- **Section 3.4: Summary of Evaluation Findings** summarizes the findings from all products and research areas, while **Section 3.4.1: Opportunities and Limitations** provides a comprehensive discussion of Big Data opportunities and limitations within each research area.

3.2 SPECIFICS OF BIG DATA PRODUCTS

This section provides an overview of all Big Data products reviewed by the research team. The intent of this section is to provide the reader with an understanding of the variety of products available and their specific applicability and limitations.

Based on the findings related to the various products, the research team organized these into different groups:

- **Mobile Device-Based Travel Patterns Data** | vendor products inferring information on travel patterns using aggregated, anonymized data from mobile devices. From this pool of products, four products were specifically identified that are potentially applicable to a large number of the TPB's research needs. For these four products, a detailed evaluation is provided in this section identifying various applicable research areas, examples of other agency use cases, validation documentation, costs/pricing structure, data sharing/licensing concerns, and data storage, processing, and analysis requirements. Other mobile device-based products also were reviewed that were deemed to have a more limited applicability; those products are described in less detail.
- **Socioeconomic or Location-based Data** | products that provide information about specific locations or various socioeconomic data elements for a region. These products may not necessarily reside in the domain of "Big Data" by definition but do have applicability to the TPB for understanding the underlying demographics and economic generators of travel in the region. These data sources also may be applicable to one of or a few specific TPB research areas.
- **Data from Public Infrastructure** | data derived from physical infrastructure deployed by an agency in the field, such as traffic count sensors, traffic signals, and transit ITS devices.
- **End-User Platforms for Data Analytics** | products that feature dashboards geared toward an end-user of Big Data; some of these platforms ingest data produced by external sources and claim to be "data-agnostic". Thus, these products are used in combination with data from other sources to produce analytics and summary metrics.
- **Emerging Big Data Sources** | finally, these are products that are considered to be in the "emerging" stage by the research team; any commercial products for transportation planning purposes are very novel or being developed for integration with end-user products. These represent areas where the research team anticipates significant progress throughout the next several years but does not currently recommend that the TPB invest in a solution. Rather, the research team recommends that the TPB take a proactive approach to monitoring the development and progress of these data sources and ensuring regional coordination on their applications.

3.2.1 MOBILE DEVICE-BASED TRAVEL PATTERNS DATA

This group of Big Data products represents a suite of available tools that infer information on travel patterns (for example, link speeds or O-D travel between zones) using aggregated, anonymized data from mobile devices. These mobile devices could be cell phones (including data from the phone's GPS, device applications using LBS, or cell tower triangulation of a device's location) as well as on-board devices within a vehicle (such as an OnStar system or freight fleet management systems). Various vendors typically aggregate many underlying raw mobile device datasets, such as a set of LBS data from a mobile app. These resulting datasets are then often processed to provide additional context, such as mapping pings from a device to a specific route or synthesizing demographic information such as income levels, race, etc. Various levels of metrics and customization are provided within each platform.

The research team has identified several products with a potential wide range of applicability to the TPB's research needs. These products are described in detail in this section. The research team also identified several products that were considered in the product evaluation process but ultimately deemed to have a more limited applicability; these products are described at the end of this section in less detail.

INRIX

Product Overview

- INRIX is a provider of vehicle probe data for segment-level congestion analytics as well as O-Ds for customized zones.
- The underlying data sources are a combination of LBS, GPS from local delivery fleets and long-haul trucks, and connected vehicles (e.g., Audi/BMW); INRIX has been increasing their data providers to increase their penetration rate for passenger vehicles as well as temporal resolution (e.g., pings every 3 to 5 seconds).
- INRIX has various product offerings: probe data (segment-level speed/travel time/AADT estimates), O-D summaries (trip starts/ends), trip paths (includes waypoints along routes and is a very large dataset), and volume profile, which provides average segment volumes by direction, day of week, and time of day across an entire year (e.g., the average segment volume at 7:15 AM on Wednesdays in 2019).
- Agencies can purchase directly (raw data) or contract with the CATT Lab at University of Maryland for providing analytics tools. Note that the RITIS platform for data analytics is described in the **Section 3.2.4**.
- INRIX is, at the time of this report, the underlying data source for the National Performance Management Research Data Set (NPMRDS), which provides region-wide travel speeds and volume estimates specifically for National Highway System (NHS) facilities. NPMRDS provides average travel speeds using the same segment definitions as INRIX and uses the underlying INRIX data, although the dataset does not impute data for missing time periods as INRIX does. It is the understanding of the research team that TPB has access to INRIX data beyond the NPMRDS dataset, including secondary roads, based on existing data agreements in place with INRIX from the Virginia Department of Transportation (VDOT), Maryland Department of Transportation (MDOT), and DDOT.
- Real-time probe (segment congestion) data is available in addition to the historic archive.

Applicable Research Areas

- **Travel Demand Forecasting - O-D Information |** O-D information is available through the INRIX Trip Analytics tools for O-D matrices (generally custom zones), select link, and route analysis (analogous to StreetLight's "Top Routes" tool); a screenline analysis tool is planned to be released soon. No modal breakdown is provided other than passenger cars/trucks and trip purpose data is not provided.
- **Travel Demand Forecasting - Regional Speeds and Volumes** | the INRIX vehicle probe dataset is being used by a large number of agencies for real-time and historic speed/congestion data (INRIX Roadway Analytics and INRIX Speed); INRIX now offers directional segment AADTs and subhourly volume estimates for intervals as granular as 15 minutes (volume profile), although these volume estimates represent averages for a day of week across an entire year and are not available for individual days or selections of days.
- **Travel Demand Forecasting - Regional Signal Delay Estimation and Queue Length** | INRIX, in conjunction with the CATT Lab, has been developing a new signal analytics tool based on connected car data (pings every 3 to 5 seconds - a "sub-segment" of the trips data that doesn't include commercial vehicles and most of the consumer vehicles that ping at lower temporal intervals). This is an emerging data source and a very new tool that has only been tested in a few locations so far (e.g., Michigan).
- **Travel Demand Management |** O-D data, while limited to vehicular trips, could be explored to understand opportunities for multimodal shifts.
- **System Performance and Congestion Management Process:** INRIX, at the time of this report, is the underlying data source for NPMRDS, which provides region-wide travel speeds and daily volume estimates for NHS facilities (free to most state DOTs and MPOs); even if an agency does not use the post-processed NPMRDS data, this data can be purchased directly from INRIX, and additional data for

other facilities is available for purchase. Event monitoring and post-hoc analysis is available through real-time feeds or through analyses using custom dates.

- **Traffic Count Variability Analysis (e.g., day-by-day) |** INRIX Volume Profile allows for segment-level volumes to be queried for individual days and intervals throughout the day.

Agencies Using Product

- The TPB has been using INRIX historic speed data for its CMP for freeway and arterial performance monitoring; the most recent CMP Technical Report from 2018[29] noted concerns about the validity of arterial data and the need for staff to continue to monitor the quality of this data. The TPB is not using any speed data for validation of the regional travel demand model at this time, noting a preference to validate to traffic volumes rather than speeds.
- A large number of agencies have been using INRIX's speed data for several years[30] [31] [32], most notably the Eastern Transportation Coalition (ETC; formerly I-95 Corridor Coalition)[33], which partnered with the CATT Lab to provide a data analytics platform for the data (RITIS).
- There are currently several users of INRIX O-D Trip Analytics (through the CATT Lab/RITIS platform), including Maryland, Massachusetts, Pennsylvania, Rhode Island, and New York City. Maryland has purchased O-D data covering all of 2018.
- Volume Profile is a very new product being marketed to public agencies and the research team is unaware of any current public sector users.

Validation Documentation and Case Studies

- The ETC has conducted extensive validation of various INRIX products, most notably their validations of travel speeds, including a validation report for arterial speeds in September 2019[34]. The ETC is currently sponsoring ongoing research for validation of ubiquitous traffic volume data from INRIX[35].

Cost and Pricing Structure

- The TPB can currently freely access INRIX real-time and historic archive data for segment speeds and congestion for the entire region through the RITIS platform; this data is paid for by the state DOTs and was purchased through the ETC. This includes the NPMRDS probe data archive, which is freely available but technically a separate speed dataset.
- Pricing for other analytics, such as the Volume Profile or Trips O-D data, is typically based on the analytics being provided and geographic area for which data is being purchased (for example, the Trips dataset includes any trip starting, ending, or passing through that area).
- Data can be purchased as a subscription (typically providing 1 month historical and 12 months moving forward) or a discrete time period (e.g., all data from 2019) for a set geographic area; for example, Maryland bought Trips analytics statewide for all of 2015 and 2018.
- For the Volume Profile dataset for 2019, the research team obtained quotes from INRIX with pricing for Maryland, Virginia, and the District of Columbia based on population, with the understanding that it may be possible to purchase a selection of data from all three entities covering the region. Without accounting for any discounts, the research team estimates a cost of around $60,000 for the region.

[29] www.mwcog.org
[30] Using INRIX for CMP Data Collection www.dvrpc.org
[31] Using INRIX to Map Congestion drcmetro.maps.arcgis.com
[32] storymaps.arcgis.com
[33] tetc.aostudios.co
[34] inrix.com
[35] I-95 Coalition Phase 2

- Purchases of $100,000 or more from INRIX include access to RITIS analytics at no additional cost[36]. Data purchased through the ETC is subject to a 10 percent (or more) discount. "All-In" states that are part of the Vehicle Probe Project (VPP), of which Maryland and the District of Columbia are included, are subject to further discounts.
- The TPB could purchase data specific to their region (not full states); alternatively, they could freely access data purchased by a parent state DOT (such as the data Maryland has already purchased), but they would need to coordinate among all three DOTs to ensure that data time periods and granularity are consistent.

Data Licensing and Sharing

- In general, INRIX allows access to be shared between a DOT and underlying MPOs and jurisdictions as well as consultants serving those agencies.
- Maryland O-D data could be granted to the TPB (for its member jurisdictions); this data could be viewed in RITIS at the TAZ or subcounty level (Maryland provided these geometry files to the University of Maryland/CATT Lab), or a user could look at the raw data.
- DDOT purchased O-D data for 2015 for the District of Columbia; that data also could be shared with the TPB. This would contain any trips that started/ended in the District but is only O-D's (no intermediate waypoints). DDOT also doesn't have access to the analytics tools and it is a much smaller dataset. Essentially, this O-D dataset couldn't sync with Maryland's dataset.

Data Storage, Processing, and Analysis

- Raw data, especially the O-D (raw trip paths) dataset, is enormous and requires significant data processing, analysis, and storage expertise. These individual device pings are simply a latitude/longitude/timestamp and are not mapped to a specific facility.
- Data conflation to model networks is time consuming.
- The more common and recommended use is going through RITIS/University of Maryland; they house all the data, and users have access to their suite of user-friendly tools depending on what tools they have paid for (both the RITIS O-D tools as well as the traditional RITIS tools for congestion analytics using the speed data). This is done through an online graphical user interface (GUI); results can be downloaded in summary image or Excel files.
- The Volume Profile dataset, which provides estimated counts for segments by direction, day of week, and time of day in 15-minute intervals, is provided as a large comma separated values (CSV) file and currently has no interface with RITIS. Volumes are provided for INRIX's XD (most granular) segments.

STREETLIGHT DATA

Product Overview

- StreetLight Data is an online platform for O-D or segment-based analytics based on mobile device data.
- StreetLight's underlying mobile data sources are mainly LBS-based (cell phone apps) as well as navigation-GPS data; INRIX is one of their underlying data sources. This mobile device data is integrated with underlying contextual data such as census demographics to provide additional insights. As of July 2019, StreetLight's data repositories process analytics for approximately 79 million devices, or approximately 28 percent of the adult U.S. and Canadian population and about 13 percent of commercial truck trips. Sample sizes vary regionally, historically, and by type of analysis conducted[37].

[36] Interview with INRIX sales representative.
[37] learn.streetlightdata.com

- Trips are tied to an underlying OpenStreetMap network.
- StreetLight's online interface allowed for user-generated analyses using completely custom zones that can be either geographic areas or individual roadway segments.
- Similar to INRIX, StreetLight offers a suite of analysis tools: AADT estimation, O-D, O-D with middle filter (through a midpoint location), O-D to preset geography (e.g., TAZs, zip codes, census block groups), top routes between O-D's, and a congestion diagnostics tool for auto-generating insights.
- Coupled with these tools, StreetLight offers Trip Attributes (data related to trip characteristics, such as trip length, duration, and circuity) as well as Traveler Attributes (inferred trip purpose and demographic data such as race, income, education level, etc.).
- StreetLight recently began offering O-D analytics for bicycles and pedestrians.

Applicable Research Areas

- **Travel Demand Forecasting - O-D information**: O-D data is available for vehicle trips as well as bicycles and pedestrians (at an additional cost) for custom zones. Metrics had previously been provided via a normalized index, although StreetLight recently began offering estimates of actual trip volumes. Trip purpose (available as home-based work, home-based other, non-home-based) is inferred through underlying contextual data; additional information is available for traveler attributes such as race, income, etc.
- **Travel Demand Forecasting - Speeds and Volumes**: speeds/volumes can be estimated for any roadway links; StreetLight can provide region-wide coverage if desired.
- **Travel Demand Management**: Trip attributes (e.g., trip length distribution, trip circuity) and traveler attributes (e.g., income levels, trip purpose) can potentially be used to identify trips amenable to a potential mode shift; analytics provided for individual days of the week can be used to understand travel pattern differences across the work week.
- **System Performance and Congestion Management Process**: speeds and volumes can be estimated for any roadway link (region-wide coverage). StreetLight can be queried for specific dates and times (additional cost) for post-hoc analyses of event traffic data, but data is not available in real-time. Outputs can be broken out into medium and heavy trucks.
- **Transit and Active Travel**: bicycle and pedestrian O-D's for custom user-defined zones are now available; the available data is not yet broken down into further modes such as scooters. Non-motorized travel data is tied to the OSM bicycle/pedestrian network. This data is currently only available for a few select months in 2019, and the validation efforts to date are not as extensive as vehicular trip data validation efforts. Furthermore, the research team is concerned about sample size adequacy, especially in areas with low pedestrian and bicycle usage. Data is not available for transit modes (bus and rail).
- **Transportation Network Companies Evaluation**: custom analyses (not through the online platform) are available at additional cost to identify trips such as gig (e.g., TNC) travel (StreetLight provides an example case study in downtown Miami[38]). These trips are inferred from underlying data but are not explicitly provided to StreetLight as TNC-specific trips.
- **Traffic Counts**: AADT estimates are available for 2017 through 2019 using StreetLight's proprietary algorithm; AADTs can be calibrated if a user supplies data for various links (zones). Hourly turning movement counts can be estimated given the flexibility of the tool; for example, a user can run an analysis that provides data for a specific hour for specific movements. Outputs can be broken out into medium and heavy trucks.

[38] www.StreetLightdata.com

Agencies Using Product

- Several state DOTs currently have a statewide subscription (including VDOT) in which MPOs are granted access.[39] [40] [41] VDOT and MPOs across Virginia have used StreetLight for a variety of multimodal planning purposes.
- Met Council (Twin Cities MPO) used StreetLight for their system-wide CMP in 2020.
- StreetLight is frequently used for individual planning projects by public agencies for a variety of applications to understand local or regional travel patterns (e.g., Alexandria Transit Vision Plan, various VDOT traffic studies).

Validation Documentation and Case Studies

- Various whitepapers are available on StreetLight's website for several applications (e.g., validation of AADTs, turning movements, and Multimodal Data product).[42] [43]
- VDOT recently completed an evaluation of various products offered by StreetLight, including AADT, O-D estimates, traffic link volumes, turning movement volumes, and truck traffic. Evaluation results were mixed depending on the product used and volume levels being measured, with larger errors often associated with lower volumes and shorter time periods. This evaluation also provides a literature review of other recent validation efforts of StreetLight[44]. This VDOT report includes set guidelines for using StreetLight products, including general guidance for data extraction and preparation, typical application scenarios, quality issues and calibration, and techniques and tools for working with the metrics.
- A third-party validation recently completed by Fehr and Peers for hourly turning movement counts showed that 90 percent of the locations were effectively replicated by StreetLight.[45]

Cost and Pricing Structure

- Data can be purchased via an annual subscription or for individual projects.
- Subscriptions for DOTs/agencies are based on population of coverage area (e.g., VDOT subscription to StreetLight Data is estimated to cost more than $500,000).
- Individual project level purchases are priced by number of zones and types of analysis tools to be accessed; the minimum is 10 zones or less for the Essentials package (approximately $5,000); the most expensive product is Multi-Mode.
- Essentials package tools include: AADT, O-D, zone activity, trip attributes, traveler attributes, O-D with preset geography.
- Advanced package tools include: Essentials plus segment analysis, top routes between O-D's, custom specific dates, commercial vehicle metrics, and a traffic diagnostics tool for identifying sources of congestion, mode shift potential, etc.
- Multi-mode package tools include: Advanced plus bicycle/ped metrics.

Data Licensing and Sharing

- A DOT with a subscription (such as VDOT) can grant access to underlying MPOs covered by that DOT; TPB access via VDOT's subscription would not provide data for trips in Maryland and the metropolitan

[39] Maricopa used StreetLight Data for Commercial Truck Movement Tracking: www.azmag.gov
[40] VDOT Guidelines for using StreetLight Data for Planning Tasks www.virginiadot.org
[41] NYMTC Used StreetLight Data to evaluate local impact of freight demand and activity in the CBD tmip.org
[42] www.StreetLightdata.com
[43] StreetLight Multimodal Methodology White Paper learn.streetlightdata.com
[44] www.virginiadot.org
[45] www.fehrandpeers.com

Washington region, that do not touch Virginia (note that VDOT's subscription includes a 15-mile buffer around the state).

Data Storage, Processing, and Analysis

- StreetLight's online platform (StreetLight InSight) has a visualization feature for exploring and summarizing data.
- No raw data (individual trip data) is available; CSVs and shapefiles can be downloaded from the online platform. Excel and geographic information system (GIS) (if desired) are typically sufficient for further analysis, although larger or more complex queries from StreetLight may require large CSVs to be processed via a scripting tool such as R or Python in order to be usable in Excel.

TERALYTICS

Product Overview

- Teralytics is an online platform for O-D analytics based on cell phone tower triangulation data.
- Cell tower triangulation has a high sample rate (estimated at 15-35 percent of the population) but a lower spatial resolution. Thus, this data is not usable for route-level analysis (individual trips cannot be mapped to specific roads), but it is usable at the census tract or even TAZ level. Segment-level traffic count estimates (AADTs or turning movement counts) are not estimated via this platform.
- Teralytics claim to have a less biased sample than LBS-based analytics providers as they sit behind the firewall of all major mobile phone carriers; they claim sufficient market share among all demographics, ethnic groups, income levels, and age groups.
- Generally, Teralytics is less expensive than products providing route-level analytics (e.g., INRIX and StreetLight) while providing a robust sample for understanding larger-scale travel patterns.

Applicable Research Areas

- **Travel Demand Forecasting O-D Information |** Teralytics provides O-D trip estimates between census tracts or customized to a more granular level such as TAZs (for an additional cost); trip purpose is provided (default is to home and to work but could be customized to home-to-work). Income/age data can be overlaid using census data.
- **Travel Demand Management** | Teralytics can examine trip frequency (e.g., the number of trips being made between O-D pairs for different days of the week).
- **System Performance and Congestion Management Process**| while this product does not provide data for individual facilities, it has been used for modeling travel patterns for events (e.g., stadium traffic).
- **Transit and Non-Motorized Travel** | Teralytics can separate out trips for specific modes such as long-distance rail, subway, and planes; it is not able to separate out trips for bus, bicycle, pedestrian, etc.
- **Transportation Network Companies Evaluation** | Teralytics staff noted that it is possible to filter the data to look at commercial trips (TNCs and delivery vehicles) inferred based on the movement pattern of a device.

Agencies Using Teralytics

- SANDAG (San Diego MPO) is a major client, using Teralytics for its regional transportation plan[46].
- This tool is being used on various studies in California for vehicle miles traveled (VMT) analyses (Senate Bill 743 regulations).

[46] www.sdforward.com

- Teralytics was recently used for an analysis of commercial travel done for the City of Seattle.
- This tool has been used for multiple event analyses (e.g., Dodger Stadium, Atlanta Braves).

Validation Documentation and Case Studies

- No validation whitepapers are provided on their website. Teralytics claims to have a high-level of accuracy given its deep penetration rate, but all benchmarking appears to be done internally.[47]
- One study was identified noting a limited validation effort of Teralytics "*because Teralytics relies on a single data source with a considerably large penetration rate*"[48]. This study showed a distribution of light rail transit (LRT) trips by time-of-day estimated by Teralytics to be consistent with a "general understanding of transit trip-making patterns" for the LRT system.

Cost and Pricing Structure

- Pricing is provided on an individual project basis or via subscription (e.g., one year of unlimited use and one year's worth of data).
- Pricing is based on population of coverage area.
- The estimated subscription cost for standard out of the box use of the platform for an area with a population approximately the size of the metropolitan Washington region the metropolitan statistical area (MSA) is approximately $50,000.
- Custom data sets for a population of the same size are likely in the $70,000-90,000 range; cost depends on the level of customization.

Data Licensing and Sharing

- No sharing of data is allowed outside of the purchasing agency except with consulting firms doing a project with that agency using the data.
- A purchasing agency can share derivatives (e.g., analysis results).
- A purchasing agency could negotiate a data sharing agreement if needed.

Data Storage, Processing, and Analysis

- Similar to StreetLight, data is accessed through an online visualization platform and viewed through the user interface (UI); data can also be downloaded via CSV and analyzed in Excel or GIS.

LOCUS (CAMBRIDGE SYSTEMATICS)

Product Overview

- Locus is a product of LBS-based travel analytics provided by Cambridge Systematics (consulting firm that has been the TPB's travel demand model developer).
- Locus is built through a partnership with PlaceIQ for the underlying LBS dataset (mainly from cell phone apps).
- This product is essentially a combination of consultant services and software; it is a customizable/tailored solution. Data can be sliced as needed, with analyses conducted on an as-needed basis in addition to the products described in the next bullet. Analyses are not constrained by the available inputs and options on an online platform.

[47] www.teralytics.net
[48] www.ssti.us

- Four separate products are offered in addition to the custom analyses: O-D tables (expanded/validated by travel purpose/time-of-day), a transit competitiveness dashboard, a geofence analysis of movement around activity centers, and survey assist to supplement traditional HTS data.
- At this time, this product is not providing speeds/travel times, count estimates, or breakdowns by modes/vehicle classes (trucks). This tool is focused on understanding regional travel patterns for all trips across all modes.

Applicable Research Areas

- **Travel Demand Forecasting O-D information** | This product provides O-D tables which can be expanded/validated by travel purpose and time of day. The trip purpose can be inferred from home/work/school, but the other category is much more nebulous. This product does not provide a breakdown of vehicle types/modes (essentially showing all multimodal trips) and is not providing speeds/travel times as of the publication of this report.
- **Travel Demand Management** | The analyses could feasibly be performed to understand travel patterns that are amenable to mode shift/telework.
- **Transit and Active Travel** | While trips shown represent all multimodal trips, data can be used to understand people movements.

Agencies Using Product

- Locus has an ongoing case study with Los Angeles (LA) Metro (transit agency) to understand transit riders and overhaul their bus system routing.

Validation Documentation and Case Studies

- Validation is a trade secret.

Cost and Pricing Structure

- Base product (trip tables), including consultant services, is approximately $150,000 to $200,000; data would be available at the census tract level. These trip tables would provide O-D flows by trip purpose and time of day.
- Consulting labor fees are the main driver of cost for additional custom analyses, including building custom dashboards.

Data Licensing and Sharing

- Data licensing and sharing is flexible; MPO sharing upward to a DOT level would be an additional fee.

Data Storage, Processing, and Analysis

- Limited data storage and processing are required on the agency end, as analyses are being conducted as part of consultant services and data accessed via dashboards.

REPLICA

Product Overview

- Replica is an online platform for aggregate-level mobility, economic activity, and COVID-19 trend data as well as detailed travel pattern data for select regions.
- Replica ingests a wide range of data inputs, including mobile location data similar to what INRIX and StreetLight Data use; publicly available datasets such as the US Census, land use regulations, and

real estate transaction data; observed ground-truth, such as traffic counts or transit ridership; and commonly used geographies, such as Census geographies.

- Replica has two modules: Trends and Places. Trends depicts weekly trends of aggregate-level travel patterns, provided at the national and statewide levels as well as for cities, MSAs, and census tracts. Places provides much more in-depth travel data for select U.S. metropolitan areas (including the metropolitan Washington region, with a model that covers all of Virginia and Maryland). The Places data that Replica provides for a metropolitan area is simulated data, based on an activity-based travel model that Replica has developed for the area using a synthetic population.
- The Trends module provides various metrics for an average weekday within a selected week (e.g. week of April 5, 2021), and these metrics can be compared for two separate selected weeks. This data is available at the census tract level or less granular levels such as cities/MSAs or states. Examples include the following:
  - O/D Data
    - Origin locations for trips ending in a select location, or destination locations for trips starting in a select location
  - Mobility Data
    - Trip volume
    - Travel to work and school (number of residents making a weekday trip to work or school)
    - Sheltering in place (number of residents not making any weekday trip outside of their home)
    - Mode split (private auto, carpool, transit, walking, biking)
    - Trip purpose (home, work, eating, social, shopping, recreation)
    - Intra-geography trips (i.e., the percentage of trips that start and end within the same selected geographic location)
    - Residential VMT
    - Trip start time broken out by hours of an average day
  - Spending data for six sectors, including retail, grocery store, gas stations, parking-taxis-tolls, restaurants and bars, airline-hospitality-car rental, and entertainment-recreation.
  - COVID-19 data for national/regional trends on COVID-19 cases.
- The Places module includes a much wider spectrum of travel metrics, and data can be filtered/partitioned using any of the following variables:
  - Trip origin and/or destination, down to individual census block groups or TAZs
  - Trip start time (individual hours of the day)
  - Trip purpose (work, home, eat, shop, school, social, recreation, errands, lodging, pass-through, commercial and other)
  - Primary mode (Driving, auto passenger, taxi/TNC, transit, walk, bike and commercial vehicles)
  - Distance
  - Duration (travel time)
  - Network link (e.g., Francis Scott Key Bridge, I-66, I-495, I-495 Express Lanes)
  - Transit route
  - Transit boarding and/or alighting station or stop
  - Household income group (in thousands $: 0-15, 15-25, 25-50, 50–75, 75–100, 100–150, 150–200, >200)
  - Age group (< 5, 5–11, 12–17, 18–34, 35–49, 50–64, > 65)
  - Auto availability (0, 1, 2, 3+ household vehicles)
  - Race and ethnicity (White, Black, Hispanic or Latino, Asian, multiple races, Other, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander)
  - Employment status (employed, unemployed, not in the labor force)

  - Work/School location (people with a workplace or school in a geography such as a census tract)
  - Home location (people living in a geography such as a census tract or zip code)
- The Places data may be downloaded for various geographic levels, including the following:
  - Block group
  - Census tract
  - Zip code
  - City
  - County
  - TAZ

  The downloaded Places data includes a table of each trip taken by each member of the synthetic population over the course of an average weekday or weekend day.
- All Trends data and Places insights are accessible in Replica's web-based interface. Replica also generates pre-packaged reports, such as access to jobs, transit ridership by demographic group, or seasonal comparisons. Also, complete trip table downloads are available as an option for those who need to run custom analyses outside of the web app.
- Replica is currently in the process of developing a scenario analysis component to help understand future impacts from alternatives defined across four dimensions — changes in demographics, land use, infrastructure, and behavioral preferences. This scenario tool is not yet available.

Applicable Research Areas

- **Travel Demand Forecasting - O-D information**: O-D data, represented by a synthetic population, is available for all major purposes (work, home, eat, shop, school, social, recreation, errands, lodging, pass-through, commercial and other) and modes (driving, auto passenger, taxi/TNC, transit, walk, bike and commercial vehicles). O-D data can be further filtered and partitioned based on several other variables, such as trip start time, distance and duration. The data for individual trips from the synthetic population data can be downloaded for post-processing outside of the Replica platform.
- **Travel Demand Forecasting - Speeds and Volumes**: Replica provides highway traffic volume estimates based on the OSM street network. Replica also classifies highway travelers based on demographic factors, such as household income, age, gender, and race/ethnicity. These variables offer additional insights to help understand the specific travel pattern and highway system usage of any targeted demographic groups on regional highway networks and selected highway links. However, Replica does not provide travel time/speed data for highway links in the manner of National Performance Management Research Data Set (NPMRDS) through its data platform.
- **Travel Demand Management**: The granularity of the synthetic trip data (purpose, mode, length and duration) and demographic information of travelers (income, age, gender and race/ethnicity) are particularly useful and valuable to identify trips amenable to understanding the impacts of travel demand management policies and strategies from both transportation system management and environmental justice perspectives.
- **System Performance and Congestion Management Process**: Replica provides estimated traffic volume by vehicle classes that can be used to help with system performance and congestion management process. At this time, the platform does not provide information related to congestion on individual links and facilities.
- **Transit and Active Travel**: Replica's synthetic population data address multimodal aspects of the transportation system and provide detailed travel pattern data and insights on transit and active travel modes. Transit, bicycle, and pedestrian O-Ds are also available for Census geographies, cities, counties and TAZs. However, the available data is not yet broken down by some "new" modes such as scooters. In addition to transit and active travel mode O-D data that can be partitioned based on

demographic information of travelers and other variables as mentioned above, Replica provides public transit passenger boarding/alighting counts at stations/stops and onboard passenger counts. Similarly, transit boarding data can be classified or filtered based on demographic variables. These variables offer additional insights to understand the transit travel patterns and transit system usage of any targeted demographic groups on the entire transit network and for selected transit routes/stations/stops.

- **Transportation Network Companies Evaluation**: Although Replica uses taxi/TNC as a separate mode for travel pattern data, the data accuracy and applicability may need to go through a more rigorous data validation from independent data users.
- **Traffic Counts**: AADT or intersection turning movement counts (TMC) are not available from Replica in the same manner as from StreetLight Data, although Replica provides model estimated traffic volumes at the link level.

Agencies Using Product

- Illinois DOT currently uses Replica data as its statewide Activity-Based Travel demand model.
- NJ Transit has a data subscription to help develop and manage regional pandemic transit plans.
- Sacramento Area Council of Governments (SACOG) purchased Replica data to support travel demand modeling and transportation planning.
- Other Replica data users include New York DOT, New York Metropolitan Transportation Council (MTC), New York City Metropolitan Transportation Authority (MTA), California Air Resources Board (CARB), and Caltrans.

Validation Documentation and Case Studies

- While Replica's platform provides potential in many application areas, its products are relatively new and in their early stage of market expansion. Due to this, it is difficult to locate any validation documentation and case study reports from Replica users. SACOG is one of the agencies that acquired Replica data to assist travel demand modeling and transportation planning. SACOG has reviewed and evaluated Replica data independently and their review results, although not formally documented, suggested Replica data reasonably represented the regional travel patterns. Replica has published a series of internal quality reports that compared the Replica data with ground truth mostly on demographics, highway traffic and transit ridership for several regions includingthe metropolitan Washington region., Baltimore, El Paso, Great Lake East, Kansas City, New York City, Northern/Southern California, Sacramento, and Portland, and several states including Illinois, Texas, New Mexico and Oklahoma.

Cost and Pricing Structure

- Replica offers a subscription service to MPOs that provides complete access to the user interface as well as unlimited seats and data downloads for all users. Inside the user interface, the MPO will have access to Trends and Places.
- There are two primary subscription tiers that are delineated by the amount of access the MPO wants to procure and how many local agencies they want to include in their subscription:
    - **MPO Single Access** – estimated $104,000 annual cost and two-year contract for MWCOG.
        - Access for all employees of the MWCOG MPO.
        - Cost breakdown is $0.15 per resident which is calculated off the largest city in the MPO's geographic footprint (in this case, District of Columbia).
    - **MPO Regional Access** – estimated $173,000 annual cost and two-year contract for MWCOG.

  - Access for all employees of MWCOG MPO as well as all employees of member organizations within the MPO's jurisdiction (for example, Arlington County and the City of Alexandria employees would also receive access through this subscription).
  - Cost breakdown is $0.25 per resident calculated off all residents in the MPO's footprint.

- MPO agencies can also benefit from a Replica data subscription from Replica's partner consulting firms if those firms are hired to work on MPO projects and studies. However, MPOs will not have direct user interface access without subscribing to one of the access structures above.

Data Licensing and Sharing

- An MPO with an annual subscription (e.g. MWCOG) can grant access to data users based on the type of user access specified in the subscription. The MPO Single Access subscription allows unlimited access for unlimited users only within the MPO, while the MPO Regional Access subscription also permits data access from all users from the member organization of the MPO. With either type of subscription, the MPO may grant temporary but full data access to its consultants/contractors helping them with projects and studies, viewing them as extended MPO staff.

Data Storage, Processing, and Analysis

- Replica's online platform has visualization and reporting features for exploring and summarizing both Trends and Places data and developing standard reports.
- Although no raw mobile location data is available, the detailed synthetic travel pattern data and associated demographic data can be downloaded from the online platform for customized analysis. The O-D data is downloaded as compressed CSV files with file extension '.gz'. Although there are limits on the number of rows of data that can be downloaded each time, such limits are subject to change and are usually high enough (e.g., more than 25 million) for most downloading needs. Highway network data can be downloaded as a CSV file for network attributes including estimated traffic volumes and a GeoJSON file for network geo-information and topology. The other data tables generated by the platform can also be easily downloaded.
- Although the size of some downloaded data files may be large, the files are easy to work with in Excel, ArcGIS and other commonly used software tools. Large files will require scripting tools (e.g. R, Python) and/or database tools for efficient processing.

OTHER MOBILE DEVICE DATA PRODUCTS EVALUATED

The research team also identified several mobile device-based products that were considered in the product evaluation process but ultimately deemed to have a more limited applicability than the previously described products. The following products were determined to either only be applicable for limited use cases given the TPB's research needs, or the product viability/use cases are unclear at this time.

- **Uber Movement**[49] | This dataset provides aggregated zone-to-zone travel time data derived from Uber trips freely available for the metropolitan Washington., area, but at the TAZ (District of Columbia only) or census tract (roughly the area contained by the I-495 Capital Beltway) level. This online tool allows a user to select a date/time range and pick O-D zones to view travel times between those zones. Uber also recently began offering a beta Mobility Heatmap product showing color-coded links throughout the District based on the density of Uber-

[49] movement.uber.com

owned "mobility devices" (e.g., Uber Jump bikes and scooters). The data is available for select cities, including Atlanta, Boston, Cincinnati, Los Angeles, Miami, New York, Orlando, Pittsburgh, San Francisco, Seattle, Tampa Bay, and Washington, D.C.

- **Transportation Network Companies Trip Data as Required by D.C. Law** | The District of Columbia has aggregated data for private vehicles-for-hire, such as TNCs, including Uber and Lyft, as information is required to be reported to DFHV and DDOT under D.C. law[50] (§ 50–301.29a. General requirements for private vehicles-for-hire). A dashboard[51] has been created to visualize the TNC trip data, including the magnitude of TNCs trips, temporal distribution of TNCs trips, average wait-time (request-to-pickup), and number of unique operators. More granular data may not be publicly available but is being currently reported to the District government for anonymous individual trips. The Washington Metropolitan Area Transit Authority (WMATA) and the TPB can request access to the data for specific purposes that are agreed upon between these agencies and the District government.
- **Strava Metro** | Strava Metro is an aggregator of bicycle and pedestrian movements using data obtained from the Strava user mobile app (for tracking bicycle/running/walking trips). Strava Metro is their product offering for planners and provides the relative level of activity (separated out into bicycle/pedestrian) along various facilities. Strava does not have an O-D product offering. While their dataset is derived from users who are primarily monitoring athletic activity with Strava's mobile app, they offer a whitepaper on their website claiming to have a representative sample of the overall population for counts/route choice[52]. This research team feels that this dataset could be useful for investment prioritization and other multimodal planning applications, such as identifying network gaps and barriers for bicycle and pedestrian planning—it is likely not very applicable for travel demand modeling unless future model versions aim to calibrate by individual modes.
- **Verizon Traffic Data[53]** | Verizon has begun offering a series of smart cities-related products, including a product for traffic analytics. Their Traffic Data product involves mapping the cellular network for a given roadway or area to create a digital signature and capture a very high percentage of devices traveling along a corridor anonymously. The research team's understanding of this product at this time is that it is mainly focused toward localized corridor-level analyses (due to this physical mapping effort); for example, a recent study was done using this tool for quantifying tourist versus local traffic through downtown Sedona, Arizona. Other smart cities-related products offered by Verizon include tools for intersection monitoring (including traffic counts and high-resolution analytics such as arrivals on red), lighting, video, and parking solutions. Verizon's website states that smart parking, traffic, and lighting products have all been deployed in the metropolitan Washington area.
- **HERE[54]** | This is a provider of traffic analytics such as speeds and travel times similar to INRIX. Link speeds are available in 5-minute granularity bins for passenger vehicles and trucks. HERE previously offered a Trip Data analytics product with O-D data, but this product has been discontinued. HERE has a contract with ETC similar to INRIX for purchasing data. At this time, only North Carolina, Georgia, and Maryland are using this data (others are using INRIX, or, in the case of Maryland, they have access to both datasets).

[50] code.dccouncil.us
[51] app.powerbi.com
[52] cdn2.hubspot.net
[53] enterprise.verizon.com
[54] www.here.com

- **Unacast**[55] | This is a mobile device analytics provider generally focused on location-specific insights and marketed toward private-sector/specific place applications. The research team's understanding of this product, at this time, is that its transportation planning-focused applications have been in the aviation sector (e.g., understanding travel to/from an airport). Raw data feeds (e.g., individual trips, dwell, and travel events) can be purchased and could potentially be a source for developing an activity-based model, but a data purchase would require significant raw data storage and processing. Unacast also offers a product specifically geared toward event analytics.
- **AirSage**[56] | This is a mobile device analytics similar to Teralytics in that insights are derived using cell tower triangulation, rather than LBS/GPS; as such, spatial granularity is limited to the polygon level and not to individual roadway links. At this time, AirSage offers a Nationwide Trip Matrix product which allows users to purchase O-D trip tables (extrapolated to the entire population) at the county level as well as products geared toward tourism and hospitality users to understand visitor origins and demographics. AirSage also claims to offer high-resolution traffic insights with O-D data down to the census block group level and information such as demographics, speed, and congestion on individual roadway links[57]; however, the research team did not identify any examples of this product being used, and the vendor did not respond to requests for further information.
- **StreetLytics (Citilabs)** | This is a mobile device analytics platform offered by Citilabs, the vendor for Cube, used for the TPB regional travel demand model. According to their website, they claim to offer O-D data, trip purpose and modal breakdown, volume and speed data broken down by hour and time of day, and routes between origins and destinations. This product appears to be a partnership with AirSage[58]. However, multiple attempts to reach out to the vendor for a demo did not receive a response. The research team was unable to find agency applications or projects using this product other than a single university study[59]. At this time, it is unclear to the research team if this product is a viable potential product to meet the TPB's needs.
- **Google Travel Time** | This product provides travel time data between specific points on the roadway network. Google provides credits for pulling (and storing) travel time data based on the frequency of data pulls and the number of locations being queried (i.e., they use a pay-per-credit framework). To use this data set at a large scale, a billing account must be created with Google (the data must be purchased), and then data must be pulled via an application programming interface (API) using third-party software or user-generated scripts. The data must then be stored and processed (unless a third-party software product is able to do this). This dataset allows for looking at customized travel time routes but requires the user to manually set up routes to analyze.
- **Waze**[60] | Partner cities in Waze's CCP (free for a city to join) have access to Waze data, including a historical feed, via Google Cloud. This feed provides the instances of Waze jams and incidents along roadways. This data is in a raw format (e.g., JSON) but can be visualized and analyzed using third-party software such as Trafmine or Carto.

[55] www.unacast.com
[56] www.airsage.com
[57] www.airsage.com
[58] www.citilabs.com
[59] www.ltrc.lsu.edu
[60] www.waze.com

3.2.2 SOCIOECONOMIC OR LOCATION-BASED DATA

The following products provide information about specific locations or various socioeconomic data elements for a region. These products may not necessarily reside in the domain of Big Data by definition, but they may have applicability to the TPB for understanding the underlying demographics and economic generators of travel in the region. These data sources also may be applicable to one of or a few specific TPB research areas.

DISAGGREGATE CENSUS DATA

Product Overview

- The US Census Bureau publishes large demographic datasets on travel behavior and population demographics. Decennial Census data (e.g., 2020 Census) is usually made available by the Census Bureau less than a year after the survey. Various demographic/socioeconomic estimates, including the land use data used by the regional travel demand models, are usually benchmarked against decennial Census data once it becomes available.
- Other major Census product uses are the PUMS from the CTPP. PUMS provides microdata samples of households and persons that are often used to synthesize the regional households and population that underlie activity-based models (ABM), such as the one that the TPB is currently in the process of developing. CTPP also is based on ACS but is a customized data product for transportation planning usage with the inclusion of detailed Journey-to-Work information that is linked with household and person characteristics.
- QCEW is a data product developed by BLS and used for a variety of purposes, such as economic security monitoring and labor statistic reporting. QCEW provides both detailed disaggregate records and summaries of employment count, wage information for more than 95 percent of U.S. jobs, available at the county, MSA, state, and national levels by industry. The data also reports the number of establishments by geographic location.

Applicable Research Areas

- **Economic Development** | this product provides employment and wage information by industry sector that can be used by the regional economic development task force to understand the employment structure, industry strength, and weakness and prepare the workforce to stimulate economic growth.
- **Travel Demand Forecasting** | the employment size and sector information in QCEW is critical for travel demand model zonal database development. PUMS records of household and person characteristic information are one of the most important data sources for household and population synthesis that is gaining popularity in travel demand modeling and forecasting.
- **Travel Demand Management** | Census Journey-to-Work data can potentially be used to identify high-frequency home-work pairs that may be amenable to a modal shift. QCEW data can potentially be used to understand detailed employment information and possible locations to target employer-based TDM policies on travel behavior.

Agencies Using Product

- Peer agencies such as Southern California Association of Governments (SCAG), Chicago Metropolitan Agency for Planning (CMAP), Atlanta Regional Commission (ARC), Maricopa Association of Governments (MAG) and many others use the QCEW data to help develop the employment databases, and PUMS data to synthesize disaggregate regional household and person database for their regional travel demand models.

Validation Documentation and Case Studies

- **NCHRP 08-36, Task 127, Employment Data for Planning** | A Resource Guide[61], includes detailed discussion of QCEW data, how it is compared with the other employment data products as well as how the quality and granularity of QCEW data may vary from state to state. Specifically, although QCEW includes a rich and reliable set of employment and wage information that is updated quarterly and serves its purpose well, it is subject to several limitations related to its applications in transportation planning and travel demand modeling.
- **Incomplete Employment Coverage** | The QCEW program produces employment and wage information only for workers covered by State UI laws and for federal workers covered by the Unemployment Compensation for Federal Employees (UCFE) program. Workers who are not subject to UI or UCFE programs and hence not covered by QCEW include self-employed workers, agricultural workers on small farms, uniformed members of the Armed Forces, elected officials in most states, most railroad workers, some domestic workers, most student workers at schools, and employees of foreign governments and certain small nonprofit organizations.
- **Multiple Site Employment Reporting |** In the QCEW program, not all employers are able or required to report employment by location, and state rules governing reporting can vary. This may result in data issues where employment is over reported for corporate headquarters or addresses associated with the administration of payroll records and underreported for actual workplace locations.
- **Industry Classification of Establishments |** QCEW also may classify employment at the establishment and within the establishment differently, particularly when classifying primary industry activities. For example, the QCEW program only reports the predominant economic activity at an establishment, while other data programs or commercial vendors (e.g., InfoGroup and Dun & Bradstreet) also may include non-predominant economic activity within counts.

Cost and Pricing Structure

- The disaggregate records of QCEW can only be acquired through special agreements with the state DES that may require a fee which is usually nominal to process the data.

Data Storage, Processing, and Analysis

- Relatively small number of data records that are easy to analyze using any conventional data tools.

HOUSEHOLD TRAVEL SURVEY

Product Overview

- HTSs reveal insightful information with behavioral characteristics of travelers and the relationship between travel decisions and travelers' demographic backgrounds. Traditionally, HTS collect data that pertains to household, person, vehicle and trip characteristics on selected travel day(s) and usually is conducted every 5 to 10 years by MPOs, state DOTs, and FHWA through the NHTS program.

Applicable Research Areas

- **Travel Demand Forecasting |** HTS provide the most critical travel behavior and traveler's characteristic information that underlies travel demand model estimations and calibration.
- **Travel Demand Management |** These traditional and long-standing surveys are valuable in understanding the makeup of a community to better define current and forecast future travel demand. Self-reporting travel diaries can be used to better understand telecommuting habits.

[61] onlinepubs.trb.org

- **Transportation Network Companies |** HTS data will reveal TNC rider demographic information and trip purpose information that is not accessible from the previously discussed private vehicle-for-hire data. It serves a specific research area of interest by TPB for estimating demographics for TNC riders in the region, considering factors such as income, race, ethnicity, and gender.

Agencies Using Product

- Starting in 2020, the NextGen household travel survey from FHWA will combine the core data, (e.g., household and person characteristics and self-reported trip purposes, collected through traditional household travel survey instruments) with O-D and other information inferred from Big Data sources to meet the ever growing and changing travel analysis needs.
- VDOT will participate in the add-on program of the first NextGen HTS to collect an additional 10,000 household samples throughout the state of Virginia, including areas of Northern Virginia in the metropolitan Washington region.
- All surveyed peer agencies stated the usage of HTS data to support travel demand modeling and transportation planning studies.
- The TPB completed its once-in-a-decade regional HTS in 2017 and 2018.

Validation Documentation and Case Studies

- NHTS[62] provides a wealth of information regarding national household travel survey through the published survey data, reports, and other documentations.
- MAG; Tampa, FL, Baton Rouge, LA; along with several others, recently conducted GPS-assisted household travel surveys. MAG's household travel survey report[63] describes in much detail how the survey was administered, data was collected, and processed for the Phoenix area in Arizona.
- TPB recently conducted a GPS-assessed follow-up survey for its 2017-18 HTS using cell-phone technology targeted to households in its regional activity centers.

Cost and Pricing Structure

- Household travel survey data can be collected either through participating in the add-on program of NHTS at a unit price of $200-$250 per completed household sample, or by other data collection contractors with the cost ranging in general between $150-$350 per sample.

Data Storage, Processing, and Analysis

- HTS data usually comes with the sample expansion factors for households, persons, and trips that should be applied to represent the regional travel pattern. Other data processing requirements are all analysis-specific. Most analyses can be performed with traditional data processing tools (e.g., Excel).

INFOUSA (DATA AXLE USA)/DUN & BRADSTREET BUSINESS LISTINGS

Product Overview

InfoUSA (Data Axle USA)/Dun & Bradstreet business listing data are among the most commonly used data products to support the development of employment databases for regional land use, travel demand modeling, and economic development activities. Both InfoUSA (Data Axle USA) and Dun & Bradstreet data products provide detailed establishment location-based business information, including employment size and industry sector.

[62] nhts.ornl.gov
[63] www.azmag.gov

- InfoUSA (Data Axle USA) and Dun & Bradstreet business listing data are very similar to QCEW data in terms of applicable research areas at MPOs. However, the business listing data do not provide as detailed wage information as the QCEW provides.
- The archived data from InfoUSA (Data Axle USA)/Dun & Bradstreet is a good data source to help understand the historical trend of local and regional employment and industries as well as the formation, closure, transformation, and relocation of individual businesses.

Applicable Research Areas

- **Travel Demand Forecasting** | Many peer agencies indicated usage of these products for employment data or underlying land use development.
- **Travel Demand Management** | The location-based data from these vendors may be beneficial for providing context to the TPB's Growth Management and Alternative Commute Programs. Business information could be used to identify regions with employers that are more likely to adopt teleworking policies.

Agencies Using Product

- Many peer agencies, including ARC, CMAP, Southeast Michigan Council of governments (SEMCOG), and Pima Association of Governments (PAG) indicated the usage of either or both of the data products for employment data development.

Validation Documentation and Case Studies

- NCHRP 08-36, Task 127, Employment Data for Planning—A Resource Guide[64], provides some coverage of both data products with the following highlights.
    - Both datasets include similar information as disaggregate QCEW, employment size, industry type (both primary and secondary types as opposed to only primary type in QCEW), location of employment, etc. that is critical for transportation planning and modeling.
    - Both databases are proprietary in nature and consolidate the records from multiple data sources, including administrative records, public filings, online and in-person verification and information imputation using a modeling approach.
    - Both InfoUSA (Data Axle USA) and Dun & Bradstreet collect data on a relatively continuous basis, but data compilation and verification is conducted in a frequency that is unknown.
    - Both datasets also may be subject to the employment coverage issue, but to a largely unknown extent which varies from area to area.
    - Both datasets can only be acquired through a one-time purchase and/or fixed-term subscription arrangements.
- No validation white papers are provided on their websites[66, 67]. Both InfoUSA (Data Axle USA) and Dun & Bradstreet data claim to have a high-level of data accuracy and coverage.

Cost and Pricing Structure

- Although the cost of either dataset is generally based on the number of business records and number of data attributes included in the regional dataset, discrete purchase vs. continuous subscription, data customization, and the overall size of the data, are among many other factors that contribute to the overall cost.

[64] onlinepubs.trb.org
[66] www.dataaxleusa.com
[67] www.dnb.com

Data Storage, Processing, and Analysis

- The data size is relatively small and easy to analyze in Excel or other similar tools.

COSTAR DATA

Product Overview

CoStar is a database of commercial real estate transactions. CoStar data can be used to help understand the real estate development patterns in the region, provide input and parameters to land use modeling, and derive employment type and activities.

Applicable Research Areas

- **Land Use Modeling** | agent-based land use modeling simulates both residential and commercial real-estate transactions. CoStar data provides detailed historical commercial real-estate transaction information, type, size, occupancy, price, etc. to help understand and estimate the mechanism of interactions among various simulation agents, developers, buyers, renters, etc. in land use modeling.
- **Travel Demand Forecasting** | Many peer agencies indicated usage of these products for employment data or underlying land use development.
- **Travel Demand Management** | The location-based data from these vendors may be beneficial for providing context to the TPB's Growth Management and Alternative Commute Programs. Business information could be used to identify regions with employers that are more likely to adopt teleworking policies.

Agencies Using Product

- The ARC used CoStar data to assist with regional modeling work, especially as a data source of average monthly rents by type of multifamily units for land use modeling. MAG also used CoStar data for socioeconomic modeling activities.

Validation Documentation and Case Studies

- Validation is a trade secret.

Cost and Pricing Structure

- Cost of CoStar data is subscription based and the usage of the data is limited to the designated personnel included in the subscription agreement which covers 1 year at a minimum. Detailed pricing information can only be acquired through providing many usage specifics to CoStar's sales team, including the term of subscription, number of assigned data users, geography to be covered as well as the nature of the purchase agency (government, nonprofit, etc.).

Data Storage, Processing, and Analysis

- A user interface is provided to query the information for specific commercial real-estate records and the query results can be downloaded into a format that can be easily handled by Excel and other tools.

OTHER SOCIOECONOMIC OR LOCATION-BASED DATA PRODUCTS EVALUATED

The research team also identified several socioeconomic or location-based data products that were considered in the product evaluation process but ultimately deemed to have a more limited applicability than the previously described products. The following products were determined to be applicable for limited use cases given the TPB's research needs, or the product viability/use cases are unclear at this time.

- **Google Places**[65] | Google Places provides information on the location and type of places such as offices, parks, restaurants, and transit stops. This product offers an alternative approach to verify and supplement the business listing data acquired from the other sources (e.g., InfoUSA (Data Axle USA), Dun & Bradstreet, and QCEW), particularly about the presence, location, and type of the businesses, to better inform regional land use and travel demand models. PAG uses Google Places data as a routine part of the standard processes to develop the socioeconomic dataset and statewide employment database. A few other agencies also use this product but in a less formal fashion. According to their website, the database is comprised of 150 million places and updates are performed regularly with 25 million updates each day to provide accurate real-time location information. The product currently has one billion monthly active users across all industries. The Google Places web-based interface and data are paired with the Google Maps and Google Routes interfaces within the Google Cloud Platform. A user-friendly web-based interface is provided on this platform that is similar to Google Maps. Acquisition of this product may include access to these applications as well, depending on the product selected. As advertised on their website, pricing for this service is pay-as-you-go; therefore, it is a function of use for each product. Cost is calculated differently for each type of query made within Google Places. For example, a query of Place Details (i.e., name and address) is approximately $17 per 1,000 requests. Further information on potential pricing can be estimated based on intended usage at the cited reference.[66]
- **SafeGraph** | SafeGraph provides data generally associated with points of interest (POI), most notably businesses in the US and Canada, covering locations for major retail chains, shopping malls, convenience stores, airports, and more. SafeGraph data also provides visitor analytics, foot-traffic counts, and demographic insights data for POI. SafeGraph Data can answer questions such as: How often do people visit stores? Where do they come from? Where else do they shop? SafeGraph data is provided in a CSV format with no user interface or dashboard provided for working with the data. A standard number of records-based pricing information is publicly available, but contacting the sales department also is needed for any custom data products. The Illinois statewide travel demand model, which is still under development, integrates SafeGraph data for a hybrid modeling approach that leverages the merits of Big Data sources for various modeling functions and capabilities (e.g., O-D distribution and destination choices). SafeGraph data is not currently used by the TPB.
- **MapBox** | MapBox offers three types of data products to support map-based illustration and analytics[67]: MapBox Boundaries, MapBox Traffic Data, and MapBox Movement. MapBox Boundaries includes administrative, legislative, locality, postal, statistical boundaries that come cartographically matched, georeferenced, and processed as MapBox vector tiles. This tool enables comprehensive geospatial analysis such as customized sales territory planning, supply chain optimization and risk management, real estate portfolio management, or any type of business intelligence analysis among many other use cases. MapBox Traffic Data can be added to any choice of maps from HERE, TomTom, or OpenStreetMap to build custom traffic routing engines, automotive platforms, or traffic analysis solutions for live and typical traffic. MapBox traffic data also can be used to power accurate estimated time of arrival (ETAs), and efficient dispatch for logistics and delivery vehicle fleets as well as to understand historical and real-time vehicle speeds for road analysis and geospatial insights. MapBox Movement is a large location database that is continually updated on a daily basis. It may be used to understand aggregate activity, density, and movement over

[65] cloud.google.com
[66] developers.google.com
[67] www.mapbox.com

time at any custom spatial scale. As a result, MapBox Movement data also can be used to support transportation planning, demand modeling, and impact analysis. In the state-of-practice review, no peer agencies indicated the usage of MapBox data of any type.

3.2.3 DATA FROM PUBLIC INFRASTRUCTURE

This section highlights sources of Big Data that involve physical infrastructure deployed by an agency in the field and data feeds coming from this field infrastructure. Data produced by these deployments may be stored and processed to some extent by the vendor of the field devices (or a third party), but this data is typically in a very raw format not feasible for straightforward use in transportation planning. Third-party software solutions may be available for further processing this data and producing end-user analytics (described in the next section). It is the understanding of the research team that the TPB's role with these datasets would be as an end user of summary performance metrics aggregated across several agencies or jurisdictions.

CONTINUOUS TRAFFIC COUNT STATION/SENSOR DATA

Local and state DOTs typically have a series of permanent or temporary devices set up for collecting traffic counts, and additionally, vehicle classifications and spot speeds. From this data, agencies can estimate the average daily number of vehicles traversing roadway segments and VMT. Counts along local or secondary roads are typically derived from short-term (e.g., 48-hour) counts that are factored to an annual estimate based on nearby continuous count data; additionally, as agencies typically do not have the capacity to count every single road in their system annually, short-term counts may be taken on a less-than-annual basis and factored to current-year conditions. In the absence of a Big Data solution for ubiquitous count data across a region, permanent count stations provide necessary control points for calibration and validating Big Data solutions.

Permanent count devices/sensors may have communications set up to allow for real-time transfer of count data to a central server. While this data as well as data from short-term counts, is stored and organized to an extent within each DOT's count data management systems, real-time feeds for traffic counts are not typically publicly available and publiclyavailable historic archive count data is typically in a post-processed format (e.g., PDF, Excel, Shapefile) that provides summary information such as AADT and daily truck percentages.

Aggregated traffic count data for the region is already being compiled by TPB and is provided on the Regional Transportation Data Clearinghouse (RTDC)[68].

AUTOMATED TRAFFIC SIGNAL PERFORMANCE MEASURES

An emerging field of Big Data analytics in the transportation industry involves performance measures for traffic signals based on high resolution (up to 1/10th second) event data. This data—such as detector calls phase changes, and transit signal priority/preemption—can be used to provide analytics on operations at individual signals or along entire arterial corridors. For example, ATSPM can be used to show the percentage of arrivals on green along a corridor or phases that are not being completely served during their allotted green time. ATSPMs are emerging as an alternative to the traditional traffic engineering practice of ad-hoc turning movement count data collection, traffic signal timing, and arterial operations analysis.

ATSPMs are typically an undertaking at the individual agency level for the traffic signals that the agency owns and maintains. An ATSPM system requires traffic signal controllers that can generate the high-resolution event logs, vehicle detection (for most measures), communications from the signal (e.g., fiber or cellular communications), a server to gather and store the data in a database, and software to convert the data into usable information and dashboards. Various software packages are available for storing and

[68] rtdc-mwcog.opendata.arcgis.com

analyzing the data, including open-source software (initially developed by Utah DOT) or private third-party software products, such as Miovision, Live Traffic Data, and Verizon. VDOT has deployed ATSPM systems in its northern region. It currently includes 12 signals on Route 50 in Fairfax County[69].

TRANSIT DATA FROM ON-BOARD ITS DEVICES

Transit agencies employ a variety of ITS devices for monitoring the status of their system and improving system performance. These products typically consist of devices on board the transit vehicles, communications from the vehicles to a central system (sometimes in real-time), and a backend server to store and process data from these devices. These devices may include:

- **Automatic vehicle location** | used to track the locations of vehicles and monitor on-time performance. When integrated with passenger information systems, AVL can be used to inform passengers of vehicle arrival times or show the locations of vehicles on interactive maps. In a planning context, AVL data can be used to understand transit travel speeds and hotspots, delay at stop locations, and, if combined with data from passenger counters, vehicle loading.
    - Many agencies make real-time feeds of vehicle locations from their AVL system publicly available via a standardized format (GTFS-Realtime) for use in third-party applications.
    - Most, if not all, local bus agencies in the metropolitan Washington region (in addition to WMATA Metrobus and Metrorail) have AVL systems deployed, but these systems vary by vendor and age in terms of interoperability of archived data and publicly available feeds.
- **Electronic farecard or farebox data** - used to automate collection of fare payment; in a planning context, farebox data can potentially provide locations of passenger boardings, passenger counts, and transit travel times (if timestamps are recorded for transactions).
    - Across the entire metropolitan Washington region, the SmarTrip card system allows riders to load value onto their card and pay for transit rides, including bus service that is not part of the WMATA system. Bus riders also can pay using cash.
    - In the case of WMATA Metrorail, a SmarTrip card is currently required for riding in the system, and riders use their farecard to check in and check out of the system (as opposed to bus trips, in which riders only check in to the system).
    - Therefore, across the metropolitan Washington. region, WMATA SmarTrip card data can be used to understand O-D patterns across the Metrorail system as well as some (but not all) bus rides, including transfers to/from Metrorail. Typically, bus fare/ridership data associated with SmarTrip is tabulated at the end of the day by each local bus agency and the information on SmarTrip-associated riders is transferred back to WMATA's central system.
    - WMATA's internal Trace model maps SmarTrip card tap data (from where a user enters/exits the Metrorail system or boards a Metrobus) to vehicle locations to analyze vehicle crowding[70].
- **Automated passenger counters (APCs) |** used to collect counts of passengers boarding and alighting vehicles at the stop level. APC's are most commonly seen on vehicles pertaining to medium to large transit agencies. Raw APC data can be unreliable and often need to go through a formal statistical validation process. In a planning context, APC data can be strengthened when joined to AVL data, allowing for calculations of vehicle load.

Typically, the data from these systems is available via the vendor software for agencies to process further and analyze. However, depending on the age of the software and/or investment level by an agency, the data from these systems may be in a very raw format and require a significant amount of processing and

[69] The consultant team is currently working with VDOT on ATSPM applications.
[70] mobilitylab.org

analysis to gain insights. For example, raw AVL data for a transit vehicle can consist of pings from the vehicle (e.g., latitude, longitude, timestamp) recorded every 15 seconds on average, with additional information when a vehicle is at a bus stop.

3.2.4 END-USER PLATFORMS FOR DATA ANALYTICS

This section highlights products featuring dashboards geared toward an end-user of Big Data. Several of the products previously described provide their own dashboards for setting up analyses and visualizing results through maps and summary charts and graphs, while other products ingest data produced by external sources and claim to be data-agnostic so long as data is provided in a workable, standard format. The research team did not include this second group of products in the evaluation as they are not sources of Big Data per se, but rather products used in combination with data from other sources to produce analytics and summary metrics.

PRODUCT-SPECIFIC DASHBOARDS

As noted, several of the products described in the section on mobile device-based travel patterns products contain their own in-house online dashboards for users to analyze and visualize the data. StreetLight (**Figure 3-2**), Teralytics (**Figure 3-3**), and Replica (**Figure 3-4**) provide dashboards for setting up analyses as well as visualizing results. These dashboards, which are included in the purchase of project data or a subscription, feature fairly user-friendly drop-downs for various filtering and display options. Both of these products allow for the user to export results to a CSV and further analyze in Excel. Locus also offers a Transit Competitiveness Dashboard as well as the option to build further custom dashboards for end users at an additional cost.

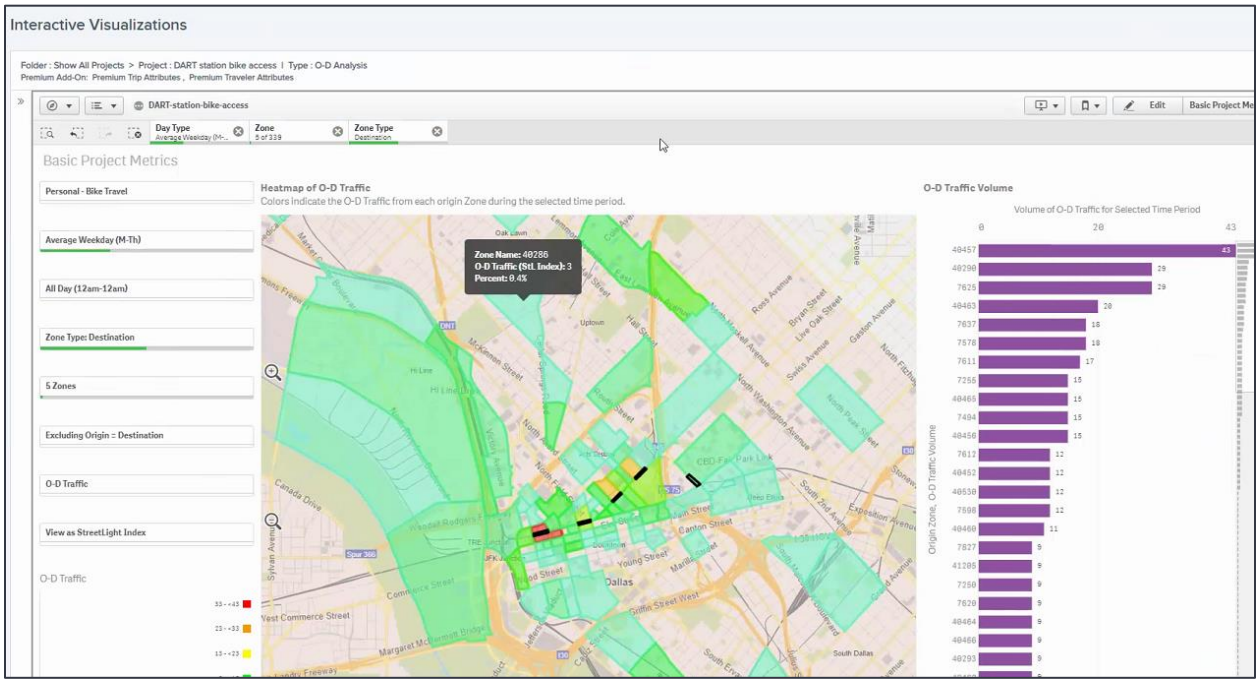


Figure 3-2 | StreetLight Dashboard (Demo)

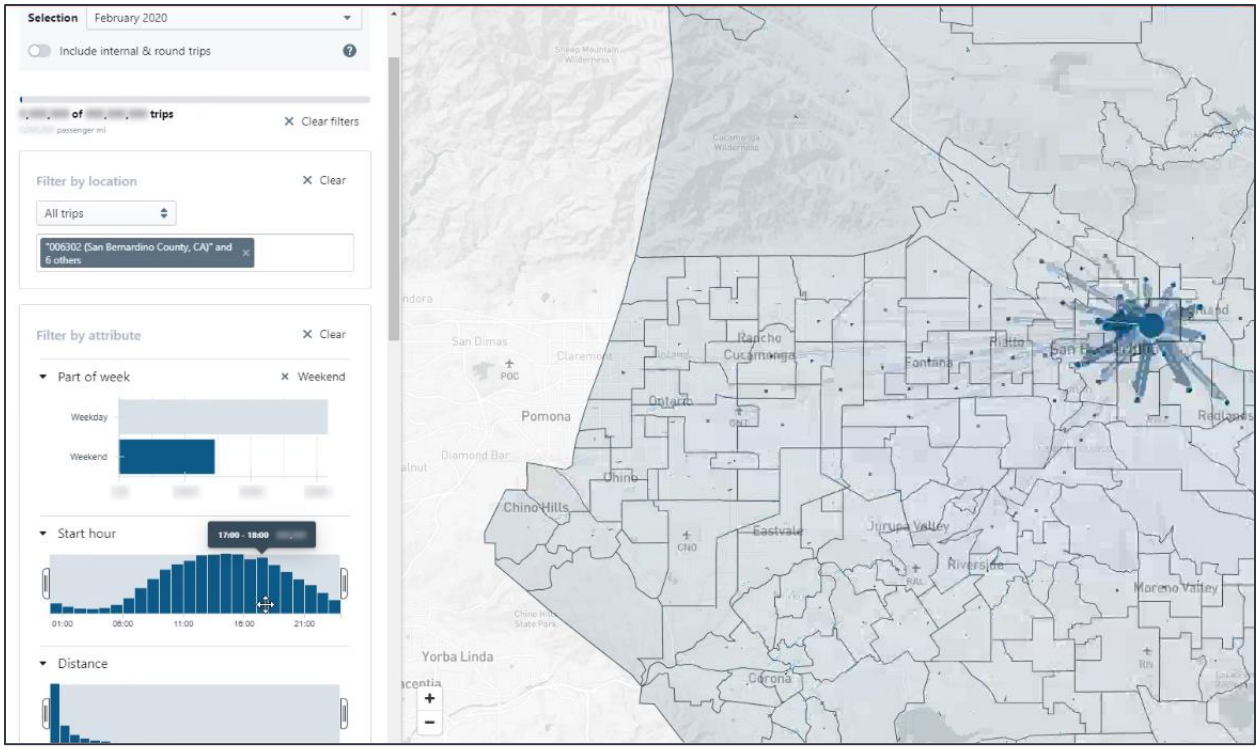


Figure 3-3 | Teralytics Dashboard (Demo)

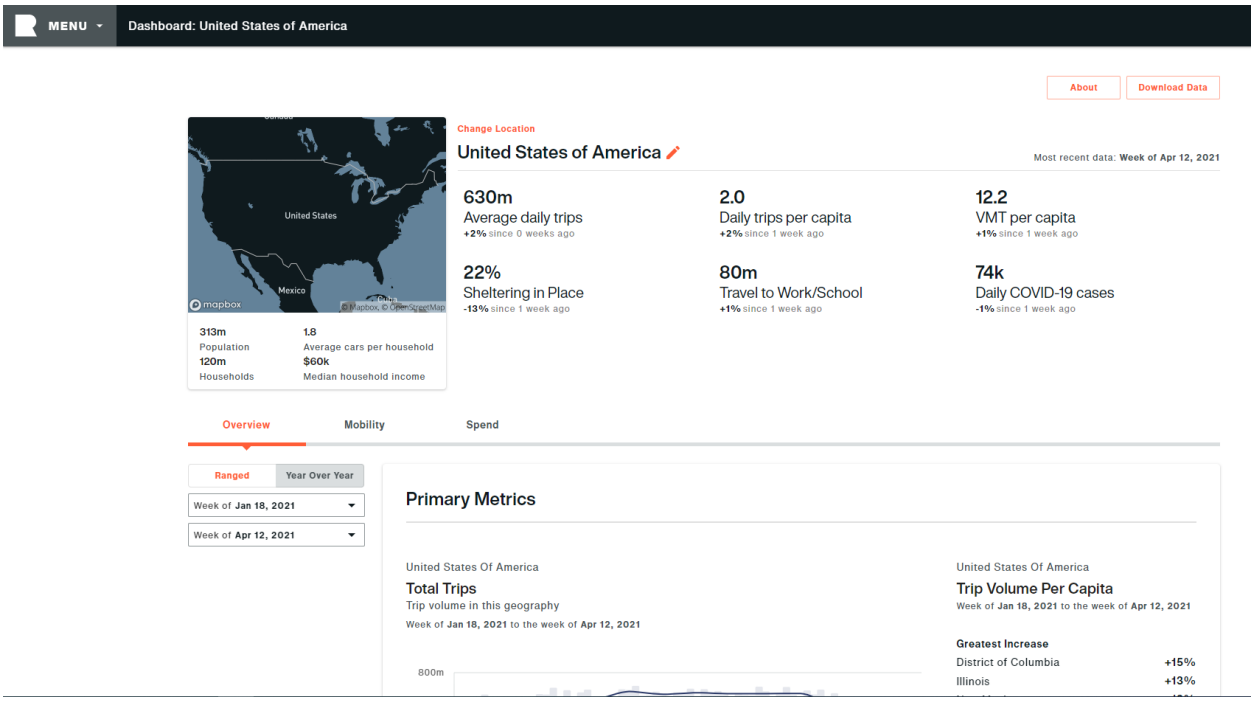


Figure 3-4 | Replica Dashboard (Demo)

REGIONAL INTEGRATED TRANSPORTATION INFORMATION SYSTEM (RITIS)

The Regional Integrated Transportation Information System (RITIS) from the CATT Lab at the University of Maryland is an "automated data fusion and dissemination system that provides an enhanced overall view of the transportation network. Participating agencies are able to view transportation and emergency management information through innovative visualizations and use it to improve their operations and emergency preparedness[71]." RITIS ingests standardized data from outside sources, including mobile device data from providers such as INRIX, HERE, and TomTom as well as data on weather, traffic incidents, and other sources of data such as agency closed-circuit television (CCTV) and detector feeds. RITIS contains a suite of real-time data feeds and situational awareness tools used by traffic operations centers and emergency management as well as a separate suite of historic archived data analysis tools.

For transportation planning purposes, the RITIS historic archive data is popular across many agencies for its suite of tools for visualizing trends. Access to various tools as well as the underlying data being processed by those tools, is dependent upon each state's or agency's contract with the CATT Lab.

- **Speed/congestion analytics |** many states, especially those on the east coast, contract with the CATT Lab to have RITIS host INRIX or HERE speed/congestion data (most clients use INRIX). This information is provided to end users at the segment level (segments are defined as traffic message channels or TMCs). Different levels of spatial coverage and tools are available depending on the contract; for example, many states only have access to the NPMRDS dataset, which contains data for only segments in the NHS system and only allows access to certain RITIS tools. Through the ETC the respective DOTs from Virginia, Maryland, and the District of Columbia all have additional access beyond the NPMRDS dataset/tools that allow these agencies to query data for many arterials and local roads and use the full RITIS suite of speed/congestion visualization tools. Currently, the TPB uses the RITIS probe data (speed/congestion) suite (and underlying INRIX data) for its CMP (most recently in 2018). This dataset also contains AADT estimates. The RITIS Probe Data Analytics Suite dashboard is shown in **Figure 3-5**.
- **O-D analytics** | separate from the probe data suite, RITIS has developed a suite of O-D analytics tools. The O-D tools currently are rooted in INRIX trajectory data (individual trip waypoints or pings) purchased by a handful of agencies, although RITIS could hypothetically ingest data from other providers into this suite. Agencies such as Maryland DOT contract with INRIX to purchase this separate dataset, which INRIX has already mapped to the roadway network, and then RITIS provides the data storage and visualization tools. The RITIS Trip Analytics suite includes a dashboard of tools similar to those offered by StreetLight, including O-D matrices, select link analyses (O-D's for trips passing through a selected segment), and an analysis of routes between two zones. Analysis zones can be defined by the agency purchasing the data (e.g., counties, TAZs) or shared via a shapefile, although unlike StreetLight, the RITIS Trip Analytics platform does not provide the user the option to manually draw custom zones within the platform. Agencies can purchase this additional dataset with pricing based on the spatial and temporal resolution desired. For example, Maryland DOT purchased Trip Analytics for all of 2018. The TPB could potentially partner with the respective state DOTs, or counties, to obtain Trip Analytics for the entire metropolitan Washington coverage area. The RITIS O-D Analytics Suite dashboard is shown in **Figure 3-6**.
- **Traffic signal analytics** | RITIS recently unveiled a Signal Analytics tool, also developed in conjunction with INRIX using trajectory data, to provide signal delay and queueing metrics. These metrics are essentially analogous to ATSPM applications, only using mobile device probe data instead of field sensor data. This platform uses only a select subset of the INRIX trajectory data in which waypoints are provided for a trip every few seconds or less (note that much of INRIX's underlying LBS data is

[71] i95coalition.org

providing a much less granular temporal resolution). The CATT Lab has thus far been working with commercializing this tool using data from Michigan. This platform offers the potential for providing traffic signal analytics without deployment of ATSPM but is very novel and not yet ready for widespread application according to the research team. Given the proliferation of data from connected vehicles (e.g., much higher-resolution data than the individual waypoint pings from GPS/LBS sources), the research team expects advances in this area throughout the next few years.

CATT Lab staff noted that there also are efforts underway to bring multimodal data, such as transit data, into the RITIS platform; for example, MDOT-MTA is paying for the CATT Lab to ingest AVL data from their buses (GTFS-RT feed). They have made efforts to form partnerships with TNCs, although company trade secrets have prevented much headway from being made and no TNC data is currently being processed in RITIS.

CATT Lab staff noted to the research team that they are willing to work with any agency to import an agency's own data into the RITIS platform and suite of tools. Pricing would be based on the format of the data and the ability to integrate within the existing RITIS suite; for example, custom applications that don't fit within the tools of the existing platform may be more expensive, but the option is there.

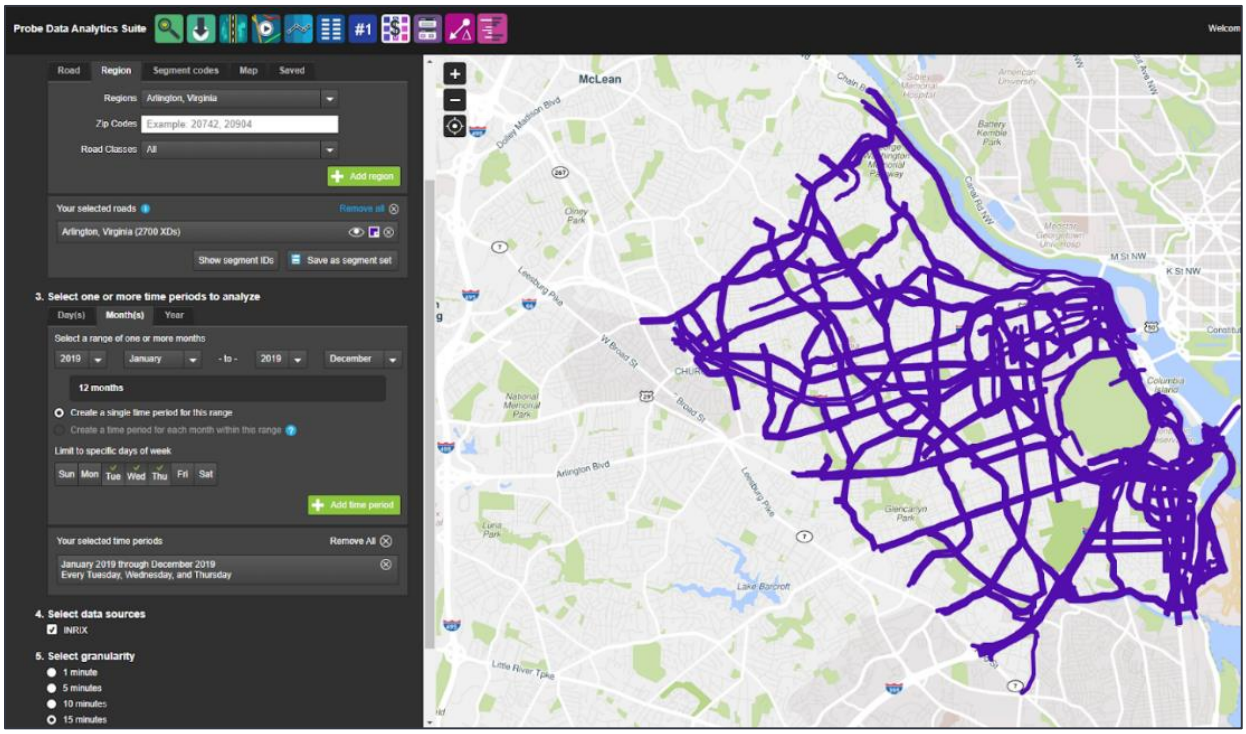


Figure 3-5 | RITIS Probe Data (Speed/Congestion) Analytics Suite Dashboard

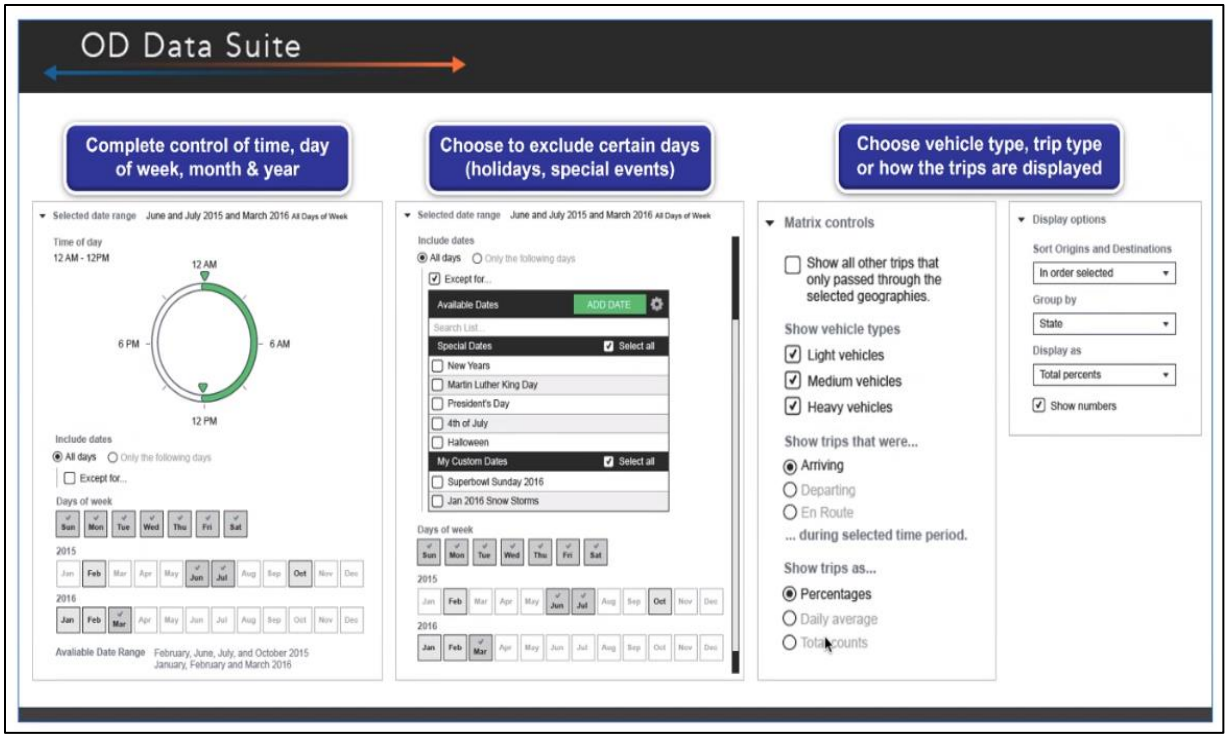


Figure 3-6 | RITIS O-D Data Analytics Suite Dashboard

MOONSHADOW/DB4IOT

Moonshadow is a provider of connected vehicle data analytics. Through a partnership with underlying providers such as INRIX and wejo, they have developed a database and visualization suite called DB4IoT. This product serves as both a backend (data storage/processing engine for raw data from vehicles and infrastructure) and a frontend (dashboard for visualizing data). DB4IoT currently incorporates data from a variety of sources:

- **Connected vehicle data** | underlying providers include INRIX and wejo (data points every 3 seconds or less). This data differs from LBS data not just in terms of temporal granularity but also in terms of information being provided; for example, wejo provides data such as hard braking/fast acceleration, which can be used for safety applications.
- **Mobile application LBS data** | underlying providers include Unacast and X-Mode Social
- **Customer data** | data from sources such as public transit feeds (e.g., vendors of AVL and transit on-board ITS systems, such as Clever and Trapeze), micromobility companies, WiFi or Bluetooth field data collection devices, and traffic counters.

DB4Iot includes a dashboard interface for users to set up analyses and visualize data. Its competitive advantage is its processing speed, which allows for queries and analyses to be run in a few seconds, rather than requiring the user to check back in to see when an analysis is done running (as currently needed with the StreetLight and RITIS tool suites). The platform allows for fairly custom analyses (e.g., filtering for specific dates/roadway segments/geographic areas as well as filtering to individual underlying datasets). It also allows for viewing of individual trip waypoints (note that INRIX anonymizes the start and end points of trips). Example use cases provided to the research team include mapping origins and destinations of trips through a given location (select link analysis) and before-and-after studies of travel through a work zone. The research team feels that this product currently offers the most potential to users who are very familiar with the underlying datasets and comfortable setting up custom analyses and viewing outputs as a heat map; its end-user analytics features are not as refined and developed as a platform such as RITIS. It appears to be most applicable for project-specific use cases, rather than system-wide performance monitoring at this time. No modal breakdown is provided at this time and data is provided as raw trip counts (i.e., not scaled up to estimate total traffic on a facility).

Agencies purchase separate underlying datasets and then pay Moonshadow a set-up fee plus a percentage of the data purchase cost (typically 30 percent, although the pricing would be structured differently if agencies are coming in with data that they have already purchased separately but need Moonshadow to process). Essentially, for INRIX and wejo, Moonshadow is a value-added reseller of their data. Pricing also is dependent on the amount of time an agency wants access to the data and the number of users accessing the data.

Moonshadow is headquartered in the state of Oregon and noted for having a large presence with clients in western Europe as well as having worked with some agencies in the New York City area. Their staff noted to the research team that current clients are using their visualization platform as a service rather than building separate downstream applications using the DB4IoT API.

SWIFTLY

Swiftly is a vendor of transit analytics using feeds from agency systems; they claim to be hardware-agnostic and software-centered for consuming and analyzing feeds from various AVL providers. They assist public transit agencies in enhancing their transit service by analyzing on-time performance and identifying operational issues. Swiftly works with agencies of all sizes, although larger agencies tend to have the biggest need for their type of product. If an agency doesn't already produce a real-time feed for certain vehicles, they also can choose to purchase the hardware through Swiftly. Furthermore, they can consume multiple agencies' feeds into one platform under a single login. Both the real-time and historical data can be accessed through their online dashboard and key-restricted API. A screenshot of the Swiftly dashboard is provided in **Figure 3-7**.

In Baltimore, Maryland, Swiftly helped MDOT-MTA improve on-time performance from around 65 percent averages in 2018 to 80 percent averages in early 2020. Their modules track transit headways, speeds, travel times, and runtimes by route, stop, day, and trip metrics that are monitored routinely or in real-time.

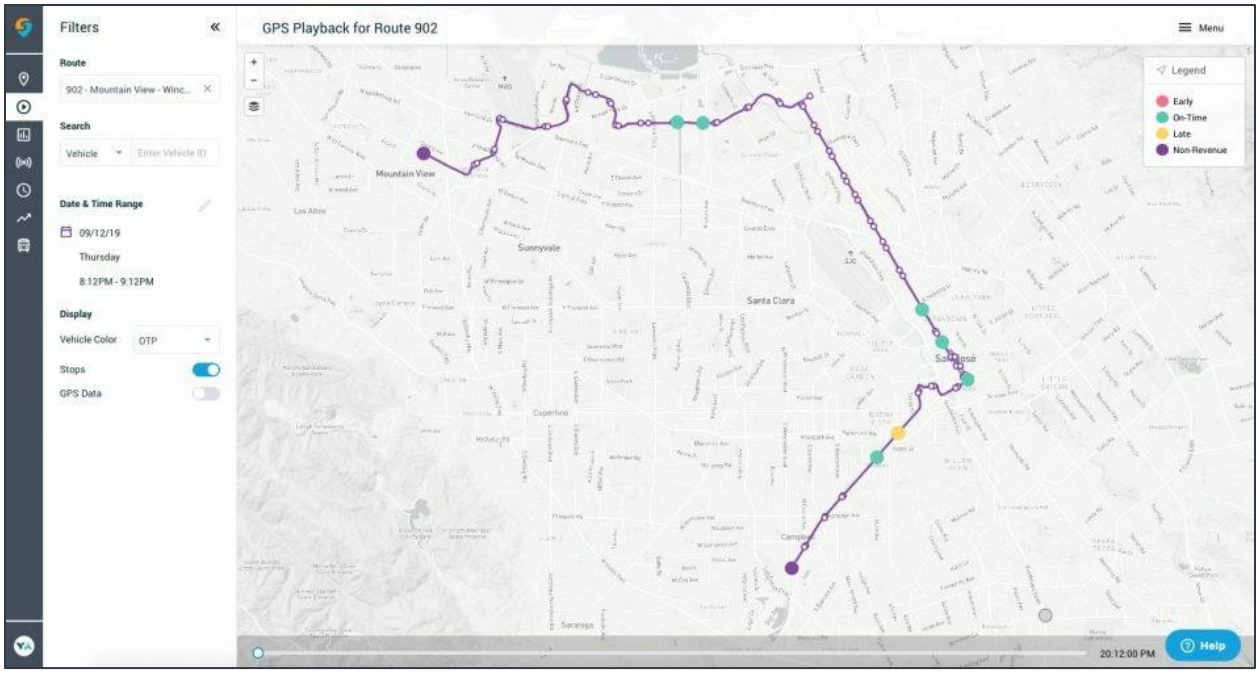


Figure 3-7 | Swiftly Dashboard (Demo)

MOOVIT

Moovit is a mobility software company that offers both user-facing products (e.g., rider trip planning tools) and analytics for transit agencies. They offer two main products: MUMA and MaaS. MUMA is an analytics tool that combines multiple data sets (e.g., transit AVL and APC data; data from their user app) to provide insights into multimodal mobility throughout a region. MUMA's list of products and visualizations include O-D matrices, travel time, and modal splits. MaaS is another software solution that attempts to integrate user-facing multimodal trip planning and aggregate the resulting movement data from its application for transit agencies. Their phone application is among the world's most downloaded in the urban mobility category, and while their user base is mainly located abroad, their presence is growing in the United States. A sample of the Moovit dashboard is provided in **Figure 3-8**.

Due to their multimodal focus, they have as much experience with MPOs as they do with transit agencies. Though they favor annual contracts as revenue streams, they have mentioned in the past that they are willing to do small pilot projects at no cost.

In early 2020, Moovit announced a partnership with Cubic Transportation Systems, a company providing ticketing, fare collection, and management solutions. Through this partnership, Moovit will incorporate its MaaS APIs with Cubic's Mobile Suite to offer an integrated traveler experience enabling Moovit's multimodal trip planning and Cubic's mobile payment and ticketing capabilities. They have been working with transit agencies to roll out this platform in Boston, Chicago, Los Angeles, Miami, New York, San Francisco, and the metropolitan Washington region.

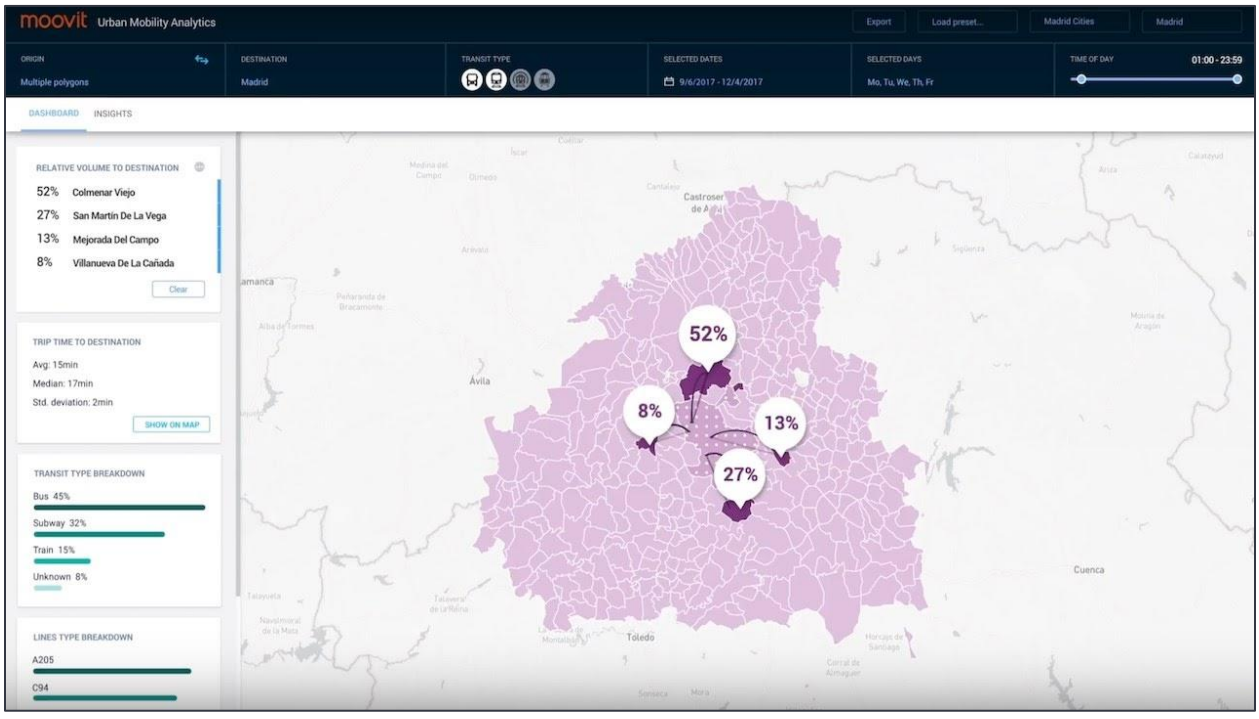


Figure 3-8 | Moovit Dashboard (Demo)

3.2.5 EMERGING BIG DATA SOURCES

The following transportation Big Data products are considered to be in the emerging stage by the research team—any commercial products for transportation planning purposes are very novel or being developed for integration with end-user products. These represent areas where the research team anticipates significant progress throughout the next several years but does not currently recommend that the TPB invest in a solution. Rather, the research team recommends that the TPB take a proactive approach to monitoring the development and progress of these data sources and ensuring regional coordination on their applications.

CONNECTED AND AUTOMATED VEHICLE DATA

Connected and automated vehicle (CAV) technologies are an area of active research and development. According to one source from 2018, companies at that point in time had invested more than $100 billion in driverless technology[72], mostly by automakers, suppliers, and technology companies. Partly based on this level of investment, most experts in the transportation industry expect autonomous vehicles will make significant changes to travel patterns and behaviors in our lifetime.

That being said, it is still unclear what those changes will be. In addition to technology readiness, consumer demand and public acceptance will depend on such unknowns as cost, track records of safety, and cybersecurity. Right now, autonomous vehicles on the road fall into one of two categories: limited autonomy, meaning the driver is still required to be engaged and take over control at any time; and fully autonomous within a very controlled environment. The former would include Teslas; the latter would be autonomous shuttles in campus settings, which still require an operator to take control in the environments in which they are being tested.

The research team categorizes current CAV data sets as follows:

- **Connected and automated vehicle Data from Original Equipment Manufacturers** | many automakers have invested in developing connectivity in vehicles and several automakers currently sell data from OEMs to vendors such as INRIX and wejo. This connected vehicle data provides information on the vehicle in much more granular time intervals than LBS data (e.g., every 3 seconds or less) and provides more information than just location and timestamp. For example, wejo, which currently has an exclusive contract with General Motors, obtains data from approximately 14 million GM vehicles in the United States and provides additional information such as speed, heating, seatbelt usage, and sharp braking/acceleration. This is resold to downstream product vendors such as Moonshadow or to engineering, construction, and consulting clients. The spatial and temporal granularity of this dataset, supplemented by the additional information about the status of the vehicle, provides opportunities for analytics not currently feasible using LBS data.
- **Scenario-Based Planning Platforms |** for the purpose of this evaluation, there is currently no data source to evaluate the effects of autonomous vehicles. Planning studies to date have simply placed various assumptions on traditional methods to account for projected changes to travel behavior. For corridor studies, setting simulation parameters to model self-driving vehicle behavior (e.g., low car-following distances) have shown some possible outcomes. However, each of these approaches fall within a type of scenario planning that does not make use of new data sources on autonomous vehicles.
- **Pilot and Test Data:** Some data has come out of pilot projects and operational tests, such as those funded by FHWA. These are yielding some data, but their applications are more for research rather than understanding the impacts of CAV on a region, as would be in the TPB's interest areas.

[72] www.leasingoptions.co.uk

MICROMOBILITY DATA/MOBILITY DATA SPECIFICATION

Throughout the past few years, a variety of private companies have deployed fleets of scooters and bikes that can be rented via mobile apps. Transportation planners across the country have desired to have data on the use of these micromobility services to better understand their use and interactivity with other modes. While many of the companies behind these services are unwilling to share much data, viewing this information as a trade secret, many jurisdictions have implemented contracts with these services that require sharing of some trip information, such as total numbers of trips, aggregated origins and destinations, or even information on individual trips. In 2019, Los Angeles (LA) DOT helped lead the creation of the Mobility Data Specification (MDS), a data standard for sharing scooter and micromobility service data. The goal of MDS is for agencies to be able to set up APIs to receive data from various private sector micromobility options. LADOT is currently exploring ways to incorporate information for additional new mobility options such as food and grocery delivery into this specification.

Separately, DDOT currently publishes downloadable files of Capital Bikeshare trip data each quarter. The data includes information on when a trip was taken, its duration, its start and end station (since this system requires docking of bikes), and type of member.[73] DDOT also has access to near real-time dockless MDS data from eight different operators (scooters and bicycles) in the region. This information is integrated with DDOT's performance management system as well as third-party software under the collaboration with the DDOT Performance Management Division, Micromobility program manager, micromobility operators, and providers of end-user platform for data analytics.

REGIONAL MULTI-MODAL MOBILITY PROGRAM (RM3P) DATA LAKE

VDOT is soliciting input from industry practitioners on existing capabilities and emerging trends and areas of innovation in the delivery of a cloud-based data lake/data store, AI-based decision support system, parking information system, data- and scenario-driven mobility analytic dashboard, and opportunities for dynamic incentivization. This data lake will provide a central repository for travel data of all modes in the Northern Virginia area.

[73] DDOT Capital Bikeshare Data www.capitalbikeshare.com

3.3 APPLICABLE BIG DATA PRODUCTS FOR TPB RESEARCH AREAS

This section provides an overview of the most applicable Big Data products for each of TPB's research areas. The intent of this section is to provide the reader with an understanding of which data products could be considered to address the needs of each research area.

3.3.1 TRAVEL DEMAND FORECASTING

The following data needs were identified for Travel Demand Forecasting by the study working group (SWG).

- Data representing O-D patterns by mode.
- Data representing O-D patterns for trips passing through a specific road segment.
- Data representing better information about vehicle travel speeds and estimated volumes.
- Data providing improved understanding of driver behavior related to dynamically-priced toll lanes and high occupancy toll (HOT) lanes. This would include information about which drivers are using them and at what rate are they being used each week.
- Data supporting improved understanding of the volume breakdown between adjacent general-purpose travel lanes and managed lanes.
- Data supporting improved understanding of the sensitivity of the aforementioned volume breakdown with dynamic changes to HOT lane price.
- Data supporting the identification of trip purpose and type of destination.
- Data providing estimates of through-travel versus external-travel (e.g., external-to-internal and internal-to-external) trips, including external transit travel, which is not currently included in the travel demand forecasting model.
- Data that can be used to support existing calibration and validation data sources, such as estimating signal delay and queue length.

APPLICABLE BIG DATA PRODUCTS

Of the evaluated Big Data products, the following products were identified as having the most potential for the travel demand forecasting research area. A comprehensive evaluation table of these projects is provided in **Appendix B**.

Mobile Device-Based Travel Patterns Data

- **INRIX** | INRIX is a provider of vehicle probe data for segment-level congestion analytics as well as O-D's for customized zones. INRIX has various product offerings: (1) segment-level probe vehicle data (speeds, travel times), (2) O-D summaries with trip start and end locations, (3) trip paths with way-points to identify routes, and (4) Volume profile providing estimated counts for each segment by direction, time of day, and day of week (averaged for a select time period). Real-time data and historic archived data can be purchased for certain product offerings.
    - **Applicability to Research Area** | INRIX is applicable to essentially all of TPB's research needs in this area, with a few limitations. Given the widespread applicability of this product for this research area, the applicability of this product is summarized for each data need in **Table 3-3**.
    - **Limitations to Research Area** | Limitations also are shown in **Table 3-3**. These include O-D data being limited to vehicular (car and truck) modes without trip purpose data. Additionally, while TPB is already using the INRIX vehicle probe dataset for speed and congestion data, the high-cost for O-D and sub-AADT Volume Profile data may be a barrier to entry.
    - Please reference product validation limitations provided in **Section 3.2.1** for INRIX data.

Table 3-3 | Applicability of INRIX to Travel Demand Forecasting Research Area

Data Need	Applicability/Limitations
O-D patterns by mode	INRIX can provide O-D data at the TAZ level for passenger vehicles and trucks. Other modes are not part of the INRIX dataset, at this time.
O-D patterns for trips passing through a specific roadway segment	The RITIS O-D tools for working with INRIX O-D data include options for understanding trips passing through a specific road segment, including select link (origins and destinations of trips passing through a segment) and route analysis (proportional breakdown of popular routes between points A and B). A screenline analysis tool is planned to be released soon as well.
Better information about vehicle travel speeds and estimated volumes	The INRIX vehicle probe dataset is being used by a large number of agencies for real-time and historic speed/congestion data (INRIX Roadway Analytics and INRIX Speed); INRIX now offers directional segment AADTs and sub-hourly volume estimates for intervals as granular as 15 minutes (INRIX Volume Profile).
Improved understanding of driver behavior related to dynamically-priced toll lanes and HOT lanes, including which drivers are using them and at what rate are they being used each week.	INRIX Volume Profile offers directional segment AADTs as well as volume by time of day and day of week. This dataset could potentially be used to understand volume trends in barrier-separated managed lane facilities, such as the I-95 and I-495 Express Lanes in Virginia, including trends by day of week and time of day. The spatial precision of underlying GPS-based and connected vehicle-based data is granular enough to support mapping of vehicle trips to parallel roadways, although the spatial precision of underlying LBS data is not and relies on inferences to map a trip to a specific facility. Currently, the research team is not aware of any validation efforts to match managed lanes volume data from this product to ground-truth data.
Improved understanding of the volume breakdown between adjacent general-purpose travel lanes and managed lanes	
Improved understanding of the sensitivity of the aforementioned volume breakdown with dynamic changes to HOT lane price	HOT lane pricing data is not available from this dataset. Volume data applicability and limitations are described above.
Identification of trip purpose and type of destination	Trip purpose data is not provided. Information about destinations is not part of this dataset.
Estimates of through-travel versus external-travel (e.g., external-to-internal and internal-to-external) trips, including external transit travel	This dataset would be very applicable to understanding or validating external trips. External transit travel could not be gleaned from this dataset at this time.
Estimating signal delay and queue length	INRIX, in conjunction with the CATT Lab, has been developing a new signal analytics tool based on connected car data (pings every 3 to 5 seconds - a subsegment of the Trips data that doesn't include commercial vehicles and most of the consumer vehicles that ping at lower temporal intervals). This is an emerging data source and a very new tool that has only been tested in a few locations so far (e.g., Michigan).

- **StreetLight** | This vendor provides a suite of analytic tools: (1) average annual daily traffic estimation, (2) O-D traffic counts, (3) O-D traffic counts with a middle filter (O-D's passing through a certain point), (4) O-D based on present geography, (5) prominent routes between an origin and destination, (6) congestion diagnostics, and (7) bicycle and pedestrian analytics.
    - **Applicability to Research Area** | StreetLight is applicable to essentially all of TPB's research needs in this area, with a few limitations. Given the widespread applicability of this product for this research area, the applicability of this product is summarized for each data need in **Table 3-4**.

- **Limitations to Research Area** | Limitations also are shown in **Table 3-4.** For StreetLight, the research team has noted the black box nature of their proprietary machine learning algorithms, which present validity questions especially for the most granular products offered (e.g., hourly turning movement counts, bicycle/pedestrian O-D data). Validation efforts and whitepapers are provided in **Section 3.2.1**. Additionally, the research team estimates a high cost of a subscription given the population of the region (likely greater than $500,000 unless overlapping state DOTs are able to purchase and share with the TPB).
- Please reference product validation limitations provided in **Section 3.2.1** for StreetLight data.

Table 3-4 | Applicability of StreetLight Data to Travel Demand Forecasting Research Area

Data Need	Applicability/Limitations
O-D patterns by mode	StreetLight can provide O-D data at the TAZ level (or any custom zones defined by a user) for passenger vehicles, trucks, bicycles, and pedestrians. Other transit modes such as bus and rail are not available.
O-D patterns for trips passing through a specific roadway segment	The StreetLight O-D tools include options for understanding trips passing through a specific road segment, including origin-middle filter-destination (origins and destinations of trips passing through a segment) and top routes analysis (proportional breakdown of popular routes between points A and B).
Better information about vehicle travel speeds and estimated volumes	Speeds/volumes can be estimated for any roadway links; StreetLight can provide region-wide coverage if desired. Volumes can be estimated at the individual hour level for roadway segments and turning movements. The sample size and quality of the data depends on the location, type of analysis being conducted, and granularity of data required (e.g., AADT vs. individual hour volumes)
Improved understanding of driver behavior related to dynamically-priced toll lanes and HOT lanes, including which drivers are using them and at what rate are they being used each week.	Roadway segment volumes are available as directional segment AADTs as well as volume by time of day and day of week. This dataset could potentially be used to understand volume trends in barrier-separated managed lane facilities, such as the I-95 and I-495 Express Lanes in Virginia, including trends by day of week and time of day. The spatial precision of underlying LBS data is not able to place a vehicle in an individual lane and relies on inferences to map a trip to a specific facility. Currently, the research team is not aware of any validation efforts to match managed lanes volume data from this product to ground-truth data.
Improved understanding of the volume breakdown between adjacent general-purpose travel lanes and managed lanes	
Improved understanding of the sensitivity of the aforementioned volume breakdown with dynamic changes to HOT lane price	HOT lane pricing data is not available from this dataset. Volume data applicability and limitations are described above.
Identification of trip purpose and type of destination	Trip purpose (available as HBW, HBO, and NHB) is inferred through underlying contextual data; additional information is available for traveler attributes such as race, income, etc. Given that all this information is inferred by StreetLight's algorithms, this represents a limitation of the dataset.
Estimates of through-travel versus external-travel (e.g., external-to-internal and internal-to-external) trips, including external transit travel	This dataset would be very applicable to understanding or validating external trips. The research team is aware of MPOs using StreetLight for developing external trip matrices for their regional demand models. External transit travel could not be gleaned from this dataset at this time.
Estimating signal delay and queue length	This dataset is not applicable for signal analytics at this time given the spatial and temporal resolution of the dataset.

- **Teralytics** | This vendor provides the following products: (1) O-D trip estimates with trip purpose, (2) capability to overlay socioeconomic and demographics data such as income and age within the analytic platform, and (3) limited differentiation of travel mode.
    - **Applicability to Research Area** | Teralytics is applicable to several of TPB's research need in this area relating specifically to O-D data. Given the widespread applicability of this product for this research area, the applicability of this product is summarized for each data need in
    - Additionally, this dataset at a regional level is much less expensive than a more granular dataset such as INRIX or StreetLight.
    - **Limitations to Research Area** | Limitations are shown in **Table 3-5.** For Teralytics, limitations are mainly a function of the spatial granularity of the data, given that cell tower triangulation data does not provide the ability to map trips to individual roadways.
    - Please reference product validation limitations provided in **Section 3.2.1** for Teralytics data.

Table 3-5 | Applicability of Teralytics to Travel Demand Forecasting Research Area

Data Need	Applicability/Limitations
O-D patterns by mode	O-D data is available at the census tract level or TAZ level (for an additional cost). Teralytics can separate out trips for specific modes such as long-distance rail, subway, and planes, claiming that these travel patterns are easier to distinguish via machine learning algorithms; it is not able to separate out trips for bus, bicycle, pedestrian, etc.
O-D patterns for trips passing through a specific roadway segment	N/A – Teralytics is not able to provide information for individual roadway segments but rather travel between geographic areas.
Better information about vehicle travel speeds and estimated volumes	
Improved understanding of driver behavior related to dynamically-priced toll lanes and HOT lanes, including which drivers are using them and at what rate are they being used each week.	
Improved understanding of the volume breakdown between adjacent general-purpose travel lanes and managed lanes	
Improved understanding of the sensitivity of the aforementioned volume breakdown with dynamic changes to HOT lane price	
Identification of trip purpose and type of destination	Trip purpose is provided (default is to home and to work but also can be customized to home to work). Income/age data can be overlaid using census data.
Estimates of through-travel versus external-travel (e.g., external-to-internal and internal-to-external) trips, including external transit travel	Teralytics could be used to understand external trip travel or validate external station data. External transit travel by rail (or plane) also could likely be identified. External transit travel by bus is not likely obtainable through this product.
Estimating signal delay and queue length	N/A – Teralytics is not able to provide information for individual roadway segments but rather travel between geographic areas.

- **Replica** | This vendor provides two modules of data: (1) Trends, which provides weekly updated data covering mobility, economic activity, and COVID-19 cases with census-tract-level fidelity, and (2) Places, which provides high-fidelity synthetic representation of travel-related outcomes that can improve the monitoring and planning of transportation and land use systems.

- **Applicability to Research Area** | Replica data are applicable to essentially all TPB's research needs in this area, with a few limitations. Given the rapidly growing applicability of this product for this research area, the applicability of this product is summarized for each data need in **Table 3-6.**
- **Limitations to Research Area** | Limitations also are shown in **Table 3-6.** For Replica, the research team has noted the black box nature of their modeling algorithms, which present validity questions especially for the most granular products offered (e.g. detailed demographic classification of travelers using transit and active transportation modes). Although data quality reports developed by Replica are available for many metropolitan areas and states, independent data validation efforts and case studies are hard to find.
- Please reference product validation limitations provided in **Section 3.2.1** for Replica data.

Table 3-6 | Applicability of Replica Data to Travel Demand Forecasting Research Area

Data Need	Applicability/Limitations
O-D patterns by mode	Replica can provide O-D data at the TAZ level (or Census geographies and jurisdictions) for driving, auto passenger, taxi/TNC, transit, walk, bike, and commercial vehicle modes. O-D transit data is not broken down into rail and bus modes.
O-D patterns for trips passing through a specific roadway segment	Replica supports select-link analysis using a polygonal lasso or selector to select links inside each megaregion.
Better information about vehicle travel speeds and estimated volumes	Replica provides estimated traffic volumes, daily and hourly, and by vehicle classes (private auto, on-demand auto, commercial and transit buses), at the link level. However, travel speed and congestion information is not available at this time.
Improved understanding of driver behavior related to dynamically-priced toll lanes and HOT lanes, including which drivers are using them and at what rate are they being used each week.	Although the data is synthesized, Replica offers detailed demographic and mode usage information for travelers using all types of transportation facilities. Replica provides a volume breakdown between adjacent general-purpose travel lanes and managed lanes, such as along I-495 or I-95 in Virginia. However, at this time, it is unclear to what extent Replica is accounting for tolling in its traffic assignment process.
Improved understanding of the volume breakdown between adjacent general-purpose travel lanes and managed lanes	
Improved understanding of the sensitivity of the aforementioned volume breakdown with dynamic changes to HOT lane price	Replica is in the process of developing a scenario analysis tool, but this tool is not yet available. It is anticipated that the granularity of the underlying dataset will be such that this tool could be used in such a manner. However, as noted above, at this time, it is unclear to what extent Replica is accounting for tolling in its traffic assignment process.
Identification of trip purpose and type of destination	Trip purpose (i.e., work, home, eat, shop, school, social, recreation, errands, lodging, pass-through, commercial and other) is inferred through the underlying contextual data used for the activity-based model. This contextual data includes underlying consumer spending data, land use regulations, commercial real estate transactions, and Census data. While the population is synthetic, Replica provides the most granular breakdown of trip purpose data of the products surveyed.
Estimates of through-travel versus external-travel (e.g., external-to-internal and internal-to-external) trips, including external transit travel	Pass-through is a separate trip purpose in Replica data for describing travel patterns. However, it is unclear at this time if external-to-internal and internal-to-external trips can be segregated from pass-through and/or other trip purposes.

Data Need	Applicability/Limitations
Estimating signal delay and queue length	This dataset is not applicable for signal analytics at this time given the spatial and temporal resolution of the dataset.

- **Locus** | This vendor provides four products: (1) O-D tables that can be expanded and validated by travel purpose and time of day, (2) transit-competitiveness dashboard, (3) geofence analysis of activity surrounding activity centers, and (4)"survey assist to supplement data received from traditional HTS.
    - **Applicability to Research Area** | Locus is applicable to several of TPB's research need in this area relating specifically to O-D data. Given the widespread applicability of this product for this research area, the applicability of this product is summarized for each data need in **Table 3-**. A potential advantage to this solution is the customizable, tailored nature of the product, including provided consultant support. Analyses are not constrained by the available inputs and options on an online platform.
    - **Limitations to Research Area** | Limitations are shown in **Table 3-.** For Locus, limitations include a lack of modal breakdown of trips as well as the lack of speed or volume data.
    - Please reference product validation limitations provided in **Section 3.2.1** for Locus data.

Table 3-7 | Applicability of Locus to Travel Demand Forecasting Research Area

Data Need	Applicability/Limitations
O-D patterns by mode	O-D patterns are available at customized levels of analysis; trips are not broken down by mode but represent all multimodal trips.
O-D patterns for trips passing through a specific roadway segment	Unclear at this time if this product allows for analysis of individual roadway facility O-D information.
Better information about vehicle travel speeds and estimated volumes	Currently, this product is not providing speeds/travel times or count estimates.
Improved understanding of driver behavior related to dynamically-priced toll lanes and HOT lanes, including which drivers are using them and at what rate are they being used each week.	
Improved understanding of the volume breakdown between adjacent general-purpose travel lanes and managed lanes	
Improved understanding of the sensitivity of the aforementioned volume breakdown with dynamic changes to HOT lane price	
Identification of trip purpose and type of destination	Trip purpose information is provided by time of day for home/work/school; the vendor noted that deriving the other category for trip purposes is much more nebulous.
Estimates of through-travel versus external-travel (e.g., external-to-internal and internal-to-external) trips, including external transit travel	Given the level of customization available for O-D flows, this product could likely be used to understand through-travel versus external-travel for regional model calibration and validation. This product is unlikely to be able to provide information about external transit travel.
Estimating signal delay and queue length	N/A

Socioeconomic or Location-based Data

- **Disaggregate Census Data** | The US Census Bureau publishes large demographic datasets on travel behavior and population demographics. Decennial Census data (e.g., 2020 Census) is usually made available by the Census Bureau less than a year after the survey. Various demographic/socioeconomic

estimates, including the land use data used by the regional travel demand models, are usually benchmarked against decennial Census data once it becomes available.

  - **Applicability to Research Area** | Many Census products are already in use in development of the TPB regional socioeconomic data forecasts, which provide the underlying land use inputs into the regional travel demand model. The TPB could consider using PUMS for synthesizing regional household and population data as it migrates the model to an ABM platform.
  - **Limitations to Research Area** | While Census products provide ground-truth data for underlying population and employment, information on trip-making patterns remains limited. For example, ACS Journey-to-Work today does not contain information on travel to and from federal government facilities, including military bases. Given the large presence of the federal government in the metropolitan Washington, region, this dataset is limited in that regard.
  - Please reference product validation limitations provided in **Section 3.2.2** for Disaggregate Census data.

- **Household Travel Survey** | The Household Travel Survey (HTS) reveals insightful information containing behavioral characteristics of travelers and the relationship between travel decisions and travelers' demographic backgrounds. Traditionally, the HTS collects the data that pertains to household, person, vehicle and trip characteristics on selected travel day(s). Starting in 2020, the NextGen household travel survey from FHWA will combine the core data, (e.g., household and person characteristics and self-reported trip purposes, collected through traditional household travel survey instruments) with O-D and other information inferred from Big Data sources to meet the ever growing and changing travel analysis needs.

  - **Applicability to Research Area** | The HTS provides the underlying O-D information used in calibration and validating travel patterns and trip distributions in the MWCOG regional travel demand model. All surveyed peer agencies stated the usage of household travel survey data to support travel demand modeling and transportation planning studies. The TPB just completed their 2018-2019 regional household travel survey.
  - **Limitations to Research Area** | The HTS represents a very small sample of trips in any region—much smaller than the potential number of trips represented by a Big Data product. One of the intents of this study is to identify sources that can be used to supplement or replace traditional survey data for use in a regional travel demand modeling context.
  - Please reference product validation limitations provided in **Section 3.2.2** for HTS data.

- **InfoUSA (Data Axle USA), Dun & Bradstreet, Google Places, CoStar** | InfoUSA (Data Axle USA) and Dun & Bradstreet offer databases comprised of location-based business information, including employment size and industry sector. Similarly, Google Places provides location and type information for offices, parks, restaurants, and transit stops. CoStar offers a database of commercial real estate transactions, which combines spatial location data of places with demographics information of individuals involved in the transactions. Other program areas within MWCOG and outside of the TPB are investing in similar data sources.[74] Therefore, there is potential for interagency collaboration to obtain greater benefit from investing in a location-based Big Data product.

  - **Applicability to Research Area** | Many peer agencies indicated usage of these products for employment data or underlying land use development.
  - **Limitations to Research Area** | These data sources do not provide any information about travel patterns and are strictly sources of underlying employment/land use information.

[74] www.mwcog.org

- Please reference product validation limitations provided in **Section 3.2.2** for InfoUSA (Data Axle USA) , Dun & Bradstreet, Google Place, and CoStar data.

Data from Public Infrastructure

Several large datasets are available or potentially available from field-deployed infrastructure for use in regional travel demand model development and validation/calibration. The TPB is already using several of these.

- **Continuous Traffic Count Station/Sensor Data** | Local and state DOTs typically have a series of permanent or temporary devices set up for collecting traffic counts, and likely additionally, vehicle classifications and spot speeds. From this data, agencies can estimate the average daily number of vehicles traversing roadway segments and VMT. In the absence of a Big Data solution for ubiquitous count data across a region, permanent count stations provide potential control points for calibration and validating Big Data solutions.
  - **Applicability to Research Area** | Traffic count data is used in validation of volumes at a screenline and individual link level as part of regional travel demand model highway assignment validation[75]. Current count data used in this process is derived from agency count data obtained by TPB.
  - **Limitations to Research Area** | Similar to HTS, continuous count station data represents a very small number of count locations in any region—much smaller than the potential number of traffic count locations which can be provided using a Big Data product. One of the intents of this study is to identify sources that can be used to supplement or replace traditional traffic count data.
  - Please reference product validation limitations provided in **Section 3.2.3** for Continuous Traffic Count Station/Sensor data.
- **Automated Traffic Signal Performance Measures** | ATSPMs are typically an undertaking at the individual agency level for the traffic signals that agency owns and maintains. An ATSPM system requires traffic signal controllers that can generate the high-resolution event logs, vehicle detection (for most measures), communications from the signal (e.g., fiber or cellular communications), a server to gather and store the data in a database, and software to convert the data into usable information and dashboards. Once an ATSPM system is deployed, it can be scaled to include more traffic signals.
  - **Applicability to Research Area** | ATSPMs are emerging as an alternative to the traditional traffic engineering practice of ad-hoc turning movement count data collection, traffic signal timing, and arterial operations analysis. They allow for continuous data collection and performance measurement for traffic signals and provide the ability to understand signal delay and queueing at a widespread scale.
  - **Limitations to Research Area** | As noted, ATSPMs are typically an undertaking at the individual agency level and require investment in physical hardware and backend software deployments. Currently, the research team is only aware of limited ATSPM use in the region. Furthermore, ATSPM requires signal controller upgrades to be capable of generating high-resolution event logs. The research team is unaware, at this time, what level of regional investment would be required to upgrade traffic signal controllers for various jurisdictions as well as facilitate the deployment of detection and communications to support ATSPMs.
  - Please reference product validation limitations provided in **Section 3.2.3** for ATSPM data.
- **Transit Data from On-Board ITS Devices** | Transit agencies employ a variety of ITS devices for monitoring the status of their system and improving system performance. The most applicable sources

[75] www.mwcog.org

of transit ITS data for travel demand modeling is O-D data from regional SmarTrip cards (WMATA maintains the SmarTrip farecard data for the region). Therefore, this data can be used to understand O-D travel patterns across the Metrorail system as well as some (but not all) bus rides, including transfers to/from Metrorail. Boarding and alighting data for bus systems in the region also can be applicable to model validation but do not provide O-D information.

- **Applicability to Research Area** | WMATA farecard data provides O-Ds for all Metrorail trips as well as any bus-to-rail or rail-to-bus transfers using a SmarTrip card. Thus, O-D data is available for trips within the Metrorail system. Note that this data does not provide where riders are ultimately starting or ending their trips (the first-mile/last-mile problem).
- **Limitations to Research Area** | O-D data for bus trips is much more challenging to obtain, as many agencies do not have APCs (or if they do, they do not have APCs calibrated/validated). Regardless, farebox and APC data provide information of where riders board (or also depart, in the case of APCs), but do not connect origins and destinations together; this data source does not provide first-mile/last-mile information.
- Please reference product validation limitations provided in **Section 3.2.3** for Transit Data from On-Board Devices.

End-User Platforms for Data Analytics

Several end-user platforms for data analytics are applicable to this research area given the underlying data they ingest, and the output analytics produced.

- **Product-specific dashboards** | StreetLight, Teralytics, and Locus have built-in dashboards. For StreetLight and Teralytics, the dashboards are how all data is queried and obtained.
- **RITIS** | This platform is already being used regionally to obtain link-level speed data and could be used to obtain speeds/travel times for model validation, if desired. If INRIX O-D data is procured for the region, this platform would presumably be used to set up the analyses. INRIX purchases of $100,000 or more are given free access to the platform. If the TPB wanted to bring other data into RITIS, such as transit AVL data, that would incur an additional fee.
- **Moonshadow/DB4IoT** | This platform ingests INRIX waypoint (individual trip) data and can be used for a variety of custom analyses. Hypothetically, TPB could purchase INRIX O-D data and choose to use DB4IoT for analytics (for an additional 30 percent of the data purchase cost). This platform also could ingest other data sources such as transit AVL data (again at an additional cost). The research team notes that given its novelty, the end-user analytics features are not as refined and developed as a platform such as RITIS.
- Please reference product validation limitations provided in **Section 3.2.4** for End-User Platform for Data Analytics.

Emerging Big Data Sources

The research team does not currently recommend that TPB invest in a solution for the following data sources but anticipates significant progress during the next several years.

- **CAV data**, especially data from OEMs, is likely to become more widespread as vehicle fleets turn over. Given the spatial and temporal resolution of this data source, opportunities are available for analytics not currently feasible using LBS data, such as traffic signal delay and queueing analytics. As this dataset becomes more widespread, it is possible that probe data analytics providers (such as INRIX) will continue to enhance their underlying data sets with connected vehicle data.
- **Micromobility data** may rapidly evolve given data sharing requirements from jurisdictions, proposed data standards, and a wide variety of mobility platforms (e.g., scooters, bicycles) and uses (e.g., transportation, food delivery, package delivery). This could lead to the availability of O-D data for various new modal options for incorporation into the regional travel demand model.

- Please reference product validation limitations provided in **Section 3.2.5** for Emerging Big Data Sources.

3.3.2 TRAVEL DEMAND MANAGEMENT

The following data needs were identified for travel demand management by the TPB SWG.

- Estimating the share of the workforce that is telecommuting regularly and forecasting the share of employees that may engage in telecommuting in the future.
- Evaluating the influence of employer-based travel demand management policies on travel behavior. For example, obtaining more insight on the influence of teleworking policies on trip generation.

A number of the strategies that fall under travel demand management and potential Big Data products that may enhance these research areas are discussed at length in their corresponding sections of this report (i.e., Travel Demand Modeling, Transit, Micromobility, and TNCs). The following section focuses on Big Data sources that may be used to enhance TPB's Alternative Commute Programs and Growth Management research areas.

APPLICABLE BIG DATA PRODUCTS

Of the evaluated Big Data products, a selection of products was identified as having the most potential for the travel demand management research area. These products are summarized below and the comprehensive evaluation table is provided in **Appendix B**.

Mobile Device-Based Travel Patterns Data

- **INRIX** | INRIX is a provider of vehicle probe data for segment-level congestion analytics as well as O-D's for customized zones. INRIX has various product offerings: probe data (segment-level speed/travel time/AADT estimates), O-D summaries (trip starts/ends), and trip paths (includes waypoints along routes and is a very large dataset).
    - **Applicability to Research Area |** Improved understanding of prominent travel patterns within the region could be used to target high-density regions for Alternative Commute Programs. INRIX O-D data could be explored to highlight opportunities for multimodal shifts.
    - **Limitations to Research Area |** INRIX O-D data is limited to vehicular trips. The O-D data provided by INRIX is a newer product that has not undergone as extensive external validation procedures as other INRIX products (e.g., travel speeds and traffic volume data). While the TPB has access to INRIX speed and congestion data through the parent state DOTs, INRIX O-D data would need to be purchased separately. The TPB could leverage partnerships with Maryland DOT and DDOT for access to their existing O-D data; however, these data sets do not cover the same temporal timeframe and could not be used together.
    - Please reference product validation limitations provided in **Section 3.2.1** for INRIX data.
- **StreetLight** | StreetLight Data is an online platform for O-D or segment-based analytics based on mobile device data. Similar to INRIX, StreetLight offers a suite of analysis tools: AADT estimation, O-D, O-D with middle filter (through a midpoint location), O-D to preset geography (e.g., TAZs, zip codes, census block groups), top routes between O-D's, and a congestion diagnostics tool for auto-generating insights.
    - **Applicability to Research Area |** Trip attributes (e.g., trip length distribution, trip circuity) and traveler attributes (e.g., income levels, trip purpose) can potentially be used to identify trips amenable to a potential mode shift. Analytics provided for individual days of the week can be

used to understand travel pattern differences across the work week. This data stream could be used to inform the TPB's Alternative Commute Programs by highlighting areas with high-travel demand. Compared to INRIX, StreetLight provides additional traveler information and is now providing O-D data for bicycles and pedestrians. This O-D data can be calibrated by supplying user counts if available. StreetLight has an online platform with visualization features that can be used to explore and summarize their data.

  - **Limitations to Research Area |** Acquisition of this data product is very expensive. Similar to INRIX, data sharing agreements could be developed to share StreetLight data between partner agencies (e.g., VDOT has an active subscription). However, the subscription from VDOT would not cover the rest of the region in Maryland, the metropolitan Washington region, and West Virginia.
  - Please reference product validation limitations provided in **Section 3.2.1** for StreetLight Data.

- **Teralytics** | Teralytics is an online platform for O-D analytics based on cell phone tower triangulation data. The company claims to have a less biased data sample than LBS-based analytics providers and they claim sufficient market share among all demographics, ethnic groups, income levels, and age groups.

  - **Applicability to Research Area |** Since Teralytics uses cell phone tower triangulation data, they are able to sustain a longer-term trace of individuals to ascertain trip purpose and trip frequency. This data would be valuable for identifying work-related trips as well as identifying the number of trips being made between O-D pairs on different days of the week. Thus, this could provide data to inform the TPB's Alternative Commute Programs. Generally, Teralytics is less expensive than products providing route-level analytics (e.g., INRIX and StreetLight) while providing a robust sample for understanding larger-scale travel patterns. Teralytics is being used by the San Diego MPO and various other agencies across the United States.
  - **Limitations to Research Area |** Cell tower triangulation has a high sample rate (estimated at 15 to 35 percent of the population) but a lower spatial resolution. Thus, this data is not usable for route-level analysis (individual trips cannot be mapped to specific roads), but it is usable at the census tract or even TAZ level. While Teralytics claims to have a high level of accuracy given its deep penetration rate, limited validation efforts exist to confirm data validity as most benchmarking appears to be done internally.
  - Please reference product validation limitations provided in **Section 3.2.1** for Teralytics.

- **Replica** | Replica is an online platform for trend analysis on mobility, economic activities and COVID-19, and in-depth travel data synthesized by an activity-based model. The in-depth travel data includes O-D data for a rich set of modes and trip purposes that can be further partitioned based on several other variables, such as trip length/duration and start time. Replica's activity-based model also converts O-D data to highway traffic volumes and transit ridership.
  - **Applicability to Research Area |** Trip attributes (e.g., modes, trip length/duration distribution and start time) and traveler attributes (e.g., income levels, trip purpose, age, gender and race/ethnicity) can potentially be used to identify trips amenable to a potential mode shift and evaluate environment justice for travel demand management policies and strategies. Synthetic population data is available at the individual trip level. The project team notes that transit and active transportation mode travel data provided by Replica tends to reasonably represent reality based on user feedback, and this level of detail is not available from many other similar datasets.
  - **Limitations to Research Area |** Most Replica data is synthetic in nature, including all individual person trip data. The accuracy of the data relies heavily on the availability of ground truth data to train and calibrates the underlying activity-based model which, to a large extent, is a "black box" to the data users. However, this is also true for the other vendors that use their proprietary

data algorithms to process and impute the data for their platforms. Analytics are not currently provided for individual days of the week for understanding travel pattern differences across the work week.
    - Please reference product validation limitations provided in **Section 3.2.1** for Replica Data.

- **Locus** | Locus is a product of LBS-based travel analytics product provided by Cambridge Systematics (the consulting firm that has been MWCOG's travel demand model developer). Four separate products are offered in addition to the custom analyses: 1) O-D tables (expanded/validated by travel purpose/time-of-day), 2) a transit competitiveness dashboard, 3) a geofence analysis of activity around activity centers, and 4) a survey assist to supplement traditional HTS data.
    - **Applicability to Research Area |** Locus provides O-D trip tables which could be analyzed to better understand travel patterns that are amenable to mode shifts and telework programs. Trip tables provide O-D traffic patterns based on trip purpose and time of day. This tool is focused on understanding regional travel patterns for all trips across all modes. Cambridge Systematics also provides consulting support by developing customized dashboards and analytics for specific client needs.
    - **Limitations to Research Area |** O-D trip tables are not separated by mode. Validation is a trade secret. The base product and corresponding consultant services are relatively expensive compared with the other O-D products. The primary driver of the high costs is consulting labor fees required to develop project-specific dashboards.
    - Please reference product validation limitations provided in **Section 3.2.1** for Locus.

Socioeconomic or Location-based Data

- **Disaggregate Census Data** | The US Census Bureau publishes large demographic datasets on travel behavior and population demographics. Decennial Census data (e.g., 2020 Census) is usually made available by the Census Bureau less than a year after the survey. Various demographic/socioeconomic estimates, including the land use data used by the regional travel demand models, are usually benchmarked against decennial Census data once it becomes available.
    - **Applicability to Research Area |** Census Journey-to-Work data can potentially be used to identify high-frequency home-work pairs that may be amenable to a modal shift. QCEW data can potentially be used to understand detailed employment information and possible locations to target for employer-based travel demand management policies on travel behavior.
    - **Limitations to Research Area |** Both CTPP and QCEW have the issue that multi-site businesses and some job types are not reported consistently by employers or employees, and as a result are difficult to geocode and likely to show variability from one source to another. Additionally, in CTPP, while most workers have only a single work location, there are industries where the majority of jobs do not follow this pattern. In a survey of workers with these jobs, some people will give the address of their current assignment, some will give the headquarters' address appearing on their mail or paycheck, and some may give no answer.
    - Please reference product validation limitations provided in **Section 3.2.2** for Disaggregate Census Data.

- **Household Travel Survey** | The HTS reveals insightful information containing behavioral characteristics of travelers and the relationship between travel decisions and travelers' demographic backgrounds. Traditionally, the HTS collects the data that pertains to household, person, vehicle, and trip characteristics on selected travel day(s). Starting in 2020, the NextGen household travel survey from FHWA will combine the core data, (e.g., household and person characteristics and self-reported trip purposes, collected through traditional household travel survey instruments) with O-D and other

information inferred from Big Data sources to meet the ever growing and changing travel analysis needs.

  - **Applicability to Research Area |** This traditional and long-standing survey is valuable in understanding the makeup of a community to better define current and forecast future travel demand. Self-reporting travel diaries can be used to better understand telecommuting habits. As highlighted in the description, there is great potential for merging these survey results with new Big Data products to ascertain more information, while maintaining the baseline information from this traditional source.
  - **Limitations to Research Area |** The HTS is expensive to conduct; for this reason, agencies are not able to conduct these surveys frequent enough to document dynamic changes in travel behavior; the TPB's Regional Travel Survey (RTS) is performed once every decade.[76]
  - Please reference product validation limitations provided in **Section 3.2.2** for HTS data.

- **InfoUSA (Data Axle USA), Dun & Bradstreet, Google Places, CoStar** | InfoUSA (Data Axle USA) and Dun & Bradstreet Business offer databases comprised of location-based business information, including employment size and industry sector. Similarly, Google Places provides location and type information for offices, parks, restaurants, and transit stops. CoStar offers a database of commercial real estate transactions, which combines spatial location data of places with demographics information of individuals involved in the transactions. Other program areas within MWCOG and outside of the TPB are investing in similar data sources.[77] Therefore, there is potential for interagency collaboration to obtain greater benefit from investing in a location-based Big Data product.

  - **Applicability to Research Area |** The location-based data from these vendors may be beneficial for providing context to the TPB's Growth Management and Alternative Commute Programs. Business information could be used to identify regions with employers that are more likely to adopt teleworking policies.
  - **Limitations to Research Area |** These datasets were not developed explicitly for use by transportation agencies; therefore, their application is less directly applicable than the HTS and many mobile device-based travel patterns data sources.
  - Please reference product validation limitations provided in **Section 3.2.2** for InfoUSA (Data Axle USA), Dun & Bradstreet Google Places, and CoStar data.

Data from Public Infrastructure

- **Transit Data from On-Board Devices** | WMATA maintains the SmarTrip farecard data for the region, which is a database that records O-D data for passengers of the WMATA transit network. Therefore, these data can be used to understand O-D travel patterns across the Metrorail system as well as some (but not all) bus rides, including transfers to/from Metrorail.

  - **Applicability to Research Area |** This database could be used to support travel demand management program areas by providing reliable transit O-D data that could be paired with other Big Data products (e.g., mobile device-based travel pattern data) to identify the tendency for modal shifts. Metrorail O-D trips are fully captured as passengers and must use their farecard to enter and exit the Metrorail station.
  - **Limitations to Research Area |** O-D data for regional bus systems is not fully captured. Passengers may use their farecard data to access the bus, but there is not an indication from this system when the passenger leaves the bus. Furthermore, passengers can pay to ride a bus without using a farecard. Finally, for both bus and Metrorail data, the transit agency

[76] www.mwcog.org
[77] www.mwcog.org

datasets do not provide where riders are ultimately starting or ending their trips (the first-mile/last-mile problem).

- Please reference product validation limitations provided in **Section 3.2.3** for Transit Data from On-Board Devices.

End-User Platforms for Data Analytics

Several end-user platforms for data analytics are applicable to this research area given the underlying data they ingest, and the output analytics produced.

- **Product-specific dashboards:** StreetLight, Teralytics, and Locus have built-in dashboards. For StreetLight and Teralytics, the dashboards are how all data is queried and obtained.
- **Regional Integrated Transportation Information System:** If INRIX O-D data is procured for the region, this platform would presumably be used to set up the analyses. This O-D data could be used to identify major O-D pairs contributing to the most constrained locations within the highway system as locations to target for travel demand management.
- **Moonshadow/DB4IoT:** This platform ingests INRIX waypoint (individual trip) data and can be used for a variety of custom analyses. Hypothetically, TPB could purchase INRIX O-D data and choose to use DB4IoT for analytics (for an additional 30 percent of the data purchase cost). This platform also could ingest other data sources such as transit AVL data (again at an additional cost). The research team notes that given its novelty, the end-user analytics features are not as refined and developed as a platform such as RITIS. However, this platform does appear tailored toward project-specific applications.

3.3.3 SYSTEM PERFORMANCE AND CONGESTION MANAGEMENT

The following data needs were identified for system performance and congestion management by the SWG.

- Data allowing improved monitoring and evaluation of transportation system performance.
- Data supporting regional and project-specific efforts related to congestion management.
- Data supporting the evaluation (i.e., tracking the impact) of congestion management strategies.
- Data that can be used to identify trends in regional traffic and congestion and leveraged to develop new congestion management strategies.
- Data that can be used to understand the impacts of unscheduled non-recurring events, such as major traffic incidents and severe weather events.
- Data that can be used to understand the impacts of scheduled non-recurring events, such as sporting events and rallies. Similarly, data supporting an assessment of the success of intended congestion management strategies deployed for these planned events.

APPLICABLE BIG DATA PRODUCTS

Of the evaluated Big Data products, a selection of products was identified as having the most potential for the system performance and congestion management research area. These products are summarized below, and the comprehensive evaluation table is provided in **Appendix B**.

Mobile Device-Based Travel Patterns Data

- **INRIX** | INRIX is a provider of vehicle probe data for segment-level congestion analytics as well as O-Ds for customized zones. INRIX has various product offerings, the most applicable to system performance and congestion management is its segment-level probe vehicle dataset. INRIX data, through RITIS, can be queried to provide data for all facilities falling within a jurisdiction(s) or matching

a specific facility type, allowing for region-wide performance measurement. INRIX data is currently used in the TPB's biannual CMP process.

  - **Applicability to Research Area** | INRIX is applicable to all of the data needs for this research area. Data can be queried at a regional or corridor-specific level in support of performance measurement or project-specific effort, including data for multiple time periods to track the impact of congestion management strategies. Trends in regional traffic and congestion can be obtained by looking at annual performance compared to previous years. Certain tools within RITIS, such as the Bottleneck Ranking or User Delay Cost tools, can be used to target areas for congestion management. Data can be queried for individual days to understand the impacts of specific non-recurring events and compare to "average" days. Currently, the TPB can freely access INRIX real-time and historic archive data for segment speeds and congestion for the entire region through the RITIS platform; this data is paid for by the state DOTs and was purchased through the ETC. This includes the NPMRDS probe data archive, which is freely available but technically a separate speed dataset.
  - **Limitations to Research Area |** The TPB has been using INRIX historic speed data for its CMP for freeway and arterial performance monitoring; the most recent CMP technical report from 2018 noted concerns about the validity of arterial data and the need for staff to continue to monitor the quality of this data.
  - Please reference product validation limitations provided in **Section 3.2.1** for INRIX data.

- **Replica** | Replica is an online platform for trend analysis on mobility, economic activities and COVID-19, and in-depth travel data synthesized by an activity-based model. The in-depth travel data includes O-D data for a rich set of modes and trip purposes that can be further partitioned based on several other variables, such as trip length/duration and start time. Replica's activity-based model also converts O-D data to highway traffic volumes and transit ridership.
  - **Applicability to Research Area |** Replica provides traffic volume and other highway performance metrics, such as travel time between origin and destination.
  - **Limitations to Research Area |** Replica does not currently provide any congestion-related data, such as speeds or travel times, for individual links. Most, if not all, metrics are synthesized by the activity-based model Replica uses. This makes the data not always suitable for system performance and congestion management, which usually requires raw or lightly processed data collected in the field.
  - Please reference product validation limitations provided in **Section 3.2.1** for Replica Data.

- **StreetLight** | StreetLight Data is an online platform for O-D or segment-based analytics based on mobile device data. Similar to INRIX, StreetLight offers a suite of analysis tools, the most applicable of which are its AADT and segment analysis tools. Similar to INRIX, StreetLight can be queried to provide data for all facilities within a certain geographic area, allowing for region-wide performance measurement.

  - **Applicability to Research Area |** StreetLight is applicable to all the data needs for this research area. Data can be queried at a regional or corridor-specific level in support of performance measurement or project-specific effort, including data for multiple time periods to track the impact of congestion management strategies. Trends in regional traffic and congestion can be obtained looking at annual performance compared to previous years. Certain tools within RITIS, such as the congestion diagnostics tools, can be used to target areas for congestion management. Data can be queried for individual days (at an additional cost) to understand the impacts of specific non-recurring events and compare to average days.

- **Limitations to Research Area |** Compared to INRIX, the TPB currently cannot freely access region-wide speed and travel time data from StreetLight. This data could be obtained as part of a region-wide subscription, which also would include access to StreetLight's O-D data.
- Please reference product validation limitations provided in **Section 3.2.1** for StreetLight Data.

Data from Public Infrastructure

- **Continuous Traffic Count Station/Sensor Data |** The regional CMP includes congestion monitoring and data consolidation activities such as aggregation and consolidation of regional traffic count data. Aggregated traffic count data for the region is already being compiled by TPB and is provided on the RTDC.
- **Automated Traffic Signal Performance Measures |** The regional CMP includes an assessment of delay at signalized intersections. This assessment currently accounts for regional surveys of the status of signal optimization activities. As agencies migrate to upgraded signal controllers and improved communications, ATSPMs offer the opportunity for a more in-depth assessment of arterial delay and performance. Georgia DOT is currently using ATSPMs to assess corridor, zonal, and regional arterial performance in the Atlanta region.
- **Transit Data from On-Board ITS Devices |** The regional CMP includes an assessment of congestion on transit systems. Currently, this is assessed by looking at travel speeds on transit-significant roads (using INRIX data for those roads) as well as average bus travel speeds from agency AVL data. However, likely given the processing challenges with combining AVL data from several sources, the most recent region-wide bus speeds data cited in the CMP is from 2011-2012. The CMP also assesses crowding on transit vehicles, including buses and rail. WMATA is able to provide this data for the Metrobus and Metrorail system using passenger boarding data and their own in-house analyses (for example, WMATA has an internal tool called the Line Load App to monitor passenger loads and crowdedness at Metrorail stations). Data from transit ITS systems is likely to continue to feed this region-wide analysis.

End-User Platforms for Data Analytics

Several end-user platforms for data analytics are applicable to this research area given the underlying data they ingest and the output analytics produced.

- **Product-specific dashboards |** StreetLight has a built-in dashboard and is how all data is queried and obtained.
- **RITIS |** this platform is already being used regionally to obtain link-level speed data for the CMP process. Should TPB continue to use INRIX data for this process, this data will likely be accessed through RITIS.
- **Transit analytics platforms such as Swiftly or Moovit** | tools that aggregate transit data across multiple agencies could be used to streamline analyses such as the CMP process for assessing region-wide bus speeds.

Emerging Big Data Sources

The research team does not currently recommend that TPB invest in a solution for the following data sources but anticipates significant progress during the next several years.

- **CAV data**, especially data from OEMs, is likely to become more widespread as vehicle fleets turn over. Given the spatial and temporal resolution of this data source, opportunities are available for analytics not currently feasible using LBS data, such as traffic signal delay and queueing analytics. As this dataset becomes more widespread, it is possible that probe data analytics providers (such as INRIX) will continue to enhance their underlying data sets with connected vehicle data.

- **Micromobility data** may rapidly evolve given data sharing requirements from jurisdictions, proposed data standards, and a wide variety of mobility platforms (e.g., scooters, bicycles) and uses (e.g., transportation, food delivery, package delivery). This could lead to the availability of trip totals and other information quantifying travel using these new modal options, including an understanding of the amount of trips being taken, when they are taken, and travel times for these trips.

3.3.4 TRANSIT AND ACTIVE TRAVEL

The following data needs were identified for transit and active travel by the TPB SWG.

- Data representing intercity bus travel.
- Data supporting improved understanding of traffic, congestion, and bus operations.
- Data to monitor real-time traffic, transit, and rail demand.
- Data to monitor active (e.g., bicycle, scooter, and pedestrian) travel demand.
- Data supporting the understanding of how alternative commuting modes (e.g., bike, walk, transit, rideshare, car/vanpool, and teleworking) affect the overall network. For example, how the level of service (LOS) of major highways and arterials change based on commuting behavior.

A number of Big Data products that may enhance these research areas are discussed at length in their corresponding sections of this report (i.e., TDM, TNCs). The following section focuses on Big Data sources that may be used to enhance TPB's ability to compile transit and non-motorized travel data to identify needs and locations for public transportation improvements and to monitor, update, and implement new active transportation facilities to enhance multimodal travel within the region.

APPLICABLE BIG DATA PRODUCTS

Of the evaluated Big Data products, a selection of products was identified as having the most potential for the transit and active travel research area. These products are summarized below and the comprehensive evaluation table is provided in **Appendix B**.

Mobile Device-Based Travel Patterns Data

- **StreetLight** | This vendor does not provide transit analytics currently but does provide a suite of tools for bicycle and pedestrian analytics. Many of the tools that are applicable for vehicular travel also can be used for bicycle and pedestrian travel, including zone activity, O-D, O-D through middle filters, and O-D to or from pre-set geography.
    - **Applicability to Research Area |** StreetLight is applicable to TPB's research need for monitoring active travel demand. This product can be used to estimate the amount of bicycle or pedestrian activity at a given location as well as the origins and destinations of that activity. This data can be calibrated within the StreetLight algorithm via user-supplied count data.
    - **Limitations to Research Area |** For StreetLight, the research team has noted the black box nature of their proprietary machine learning algorithms, which present validity questions especially for the most granular products offered (e.g., hourly turning movement counts, bicycle/pedestrian O-D data). This data is currently only available for a few select months in 2019, and the validation efforts to date are not as extensive as vehicular trip data validation efforts. Furthermore, the research team is concerned about sample size adequacy, especially in areas with low pedestrian and bicycle usage.
    - Please reference product validation limitations provided in **Section 3.2.1** for StreetLight Data.
- **Replica** | Replica provides analytics for transit and active transportation modes similar to those it provides for driving modes.

- **Applicability to Research Area |** Replica is applicable to TPB's research need for monitoring transit and active travel demand. This product can be used to estimate the amount of transit, bicycle or pedestrian activity at a given location as well as the origins and destinations of that activity. Replica can also offer insights about how the transit system and walk/bike paths are used by travelers classified demographically, providing pedestrian/bike volumes on streets, and transit ridership at stops/stations and by routes.
- **Limitations to Research Area |** The accuracy of the transit and active transportation data relies on the quality of the activity-based model used for data development. Active transportation ground truth data are often difficult to collect to calibrate the model, especially for non-vehicular modes.
- Please reference product validation limitations provided in **Section 3.2.1** for Replica Data.

- **Teralytics** | Teralytics can separate out trips for specific modes such as long-distance rail, subway, and planes; it is not able to separate out trips for bus, bicycle, pedestrian, etc.
  - **Applicability to Research Area |** Teralytics is applicable to TPB's research need of understanding transit and rail demand in that it can separate out trips for modes such as long-distance rail, subway, and planes.
  - **Limitations to Research Area |** O-D data is provided on a zone level (e.g., TAZ), and travel data for bus transit is not available at this time. Given the tap-in/tap-out system that WMATA has in place for the Metrorail system, the SmarTrip card data is a much more comprehensive tool for understanding subway/heavy rail travel in the region.
  - Please reference product validation limitations provided in **Section 3.2.1** for Teralytics.

- **Strava Metro** | Strava Metro is an aggregator of bicycle and pedestrian movements using data obtained from the Strava user mobile app (for tracking bicycle/running/walking trips). Strava Metro is their product offering for planners and provides the relative level of activity (separated out into bicycle/pedestrian) along various facilities.
  - **Applicability to Research Area |** Separate datasets are provided for bicycle and pedestrian activity. Strava claims to provide a representative sample of the overall population for counts/route choice. This dataset could be useful for investment prioritization, identifying gaps and barriers in bicycle and pedestrian networks, and understanding bicycle and pedestrian demand.
  - **Limitations to Research Area |** Data is qualitative and shows relative level of activity on facilities. O-D data is not available (as compared to StreetLight).
  - Please reference product validation limitations provided in **Section 3.2.1** for Strava Metro.

- **Uber Movement** | Uber recently began offering a beta Mobility Heatmap product showing color-coded links throughout the District based on the density of Uber-owned mobility devices (e.g., Uber Jump bikes and scooters).
  - **Applicability to Research Area |** The Mobility Heatmap can help provide an improved understanding of the concentration and relative magnitude of mobility devices within the District. Mobility Heatmap data could be explored to highlight demand of micromobility activities and opportunities for improved multimodal facilities if overlaying with land use and other transportation network data.
  - **Limitations to Research Area |** Data is qualitative and shows relative level of activity on facilities. This source is still in beta testing and is limited solely to the District of Columbia.
  - Please reference product validation limitations provided in **Section 3.2.1** for Uber Movement.

Socioeconomic or Place-Based Data

- **Household Travel Survey** | The HTS reveals insightful information containing behavioral characteristics of travelers and the relationship between travel decisions and travelers' demographic backgrounds. Traditionally, the HTS collects the data that pertains to household, person, vehicle, and trip characteristics on selected travel day(s). Starting in 2020, the NextGen household travel survey from FHWA will combine the core data, (e.g., household and person characteristics and self-reported trip purposes, collected through traditional household travel survey instruments) with O-D and other information inferred from Big Data sources to meet the ever growing and changing travel analysis needs.
    - **Applicability to Research Area |** This traditional and long-standing survey is valuable in understanding mode choice and self-reported travel patterns. While the survey data would not provide real-time metrics of transit ridership, it may highlight communities in which alternative modes are more frequently used and provide more detailed information related to the first-last mile decisions.
    - **Limitations to Research Area |** The HTS is expensive to conduct; for this reason, agencies are not able to conduct these surveys frequent enough to document dynamic changes in travel behavior; the TPB's RTS is performed once every decade.[78]
    - Please reference product validation limitations provided in **Section 3.2.2** for Household Travel Survey.

Data from Public Infrastructure

Transit Data from On-Board ITS Devices are highly relevant to this research area. A number of applications under this product type exist that would be applicable. These specific applications are described below.

- **Automatic vehicle location** | AVL is used to track the locations of vehicles and monitor on-time performance. In a planning context, AVL data can be used to understand transit travel speeds and hotspots, delay at stop locations/road segments, and, if combined with data from passenger counters, vehicle loading.
    - **Applicability to Research Area |** Many agencies make real-time feeds of vehicle locations from their AVL system publicly available via a standardized format (GTFS-Realtime) for use in third-party applications. This data provides insight to how traffic congestion affects bus operations. The TPB has used aggregated AVL data in the past to understand regional bus travel speeds and hotspots as part of its CMP.
    - **Limitations to Research Area |** Most, if not all, local bus agencies in the Washington, D.C., region (in addition to WMATA Metrobus and Metrorail) have AVL systems deployed, but these systems vary by vendor and age in terms of interoperability of archived data and publicly available feeds. Raw AVL data is not easy to process due to the large size. For example, raw AVL data for a transit vehicle can consist of pings from the vehicle (e.g., latitude, longitude, timestamp) recorded every 15 seconds on average, with additional information when a vehicle is at a bus stop. The region-wide bus speeds used in the regional CMP were last estimated in 2011-2012, likely given the level of effort required to compile and aggregate AVL data from various agencies across the region. Third-party processing tools could help streamline this process.

[78] www.mwcog.org

- **Electronic farecard or farebox data** | This data is used to automate collection of fare payment; in a planning context, farebox data can potentially provide locations of passenger boardings, passenger counts, and transit travel times (if timestamps are recorded for transactions).
    - **Applicability to Research Area |** Across the metropolitan Washingtonregion, WMATA SmarTrip card data can be used to understand O-D patterns across the Metrorail system as well as some (but not all) bus rides, including transfers to/from Metrorail.
    - **Limitations to Research Area |** Complete transit O-D data is only available for the Metrorail system, as for bus trips, riders only check in to the system. Note that, regardless, this data does not provide where riders are ultimately starting or ending their trips (the first-mile/last-mile problem). Also, this data would not include transit riders who evade paying the fare. WMATA post-processes the farecard data to produce ridership numbers that match real-world demand.
- **Automated passenger counters** | APC data is used to collect counts of passengers boarding and alighting vehicles at the stop level. APCs are most commonly seen on vehicles pertaining to medium to large transit agencies.
    - **Applicability to Research Area |** APC data can be used to monitor bus demand, including intercity bus ridership. In a planning context, APC data can be strengthened when joined to AVL data, allowing for calculations of vehicle load.
    - **Limitations to Research Area |** Raw APC data can be unreliable and often need to go through a formal statistical validation process. It also can be considered quite sensitive due to personally identifiable information (PII), particularly in areas with low ridership or residential density. Typically, the data from these systems is available via the vendor software for agencies to process further and analyze. However, depending on the age of the software and/or investment level by an agency, the data from these systems may be in a very raw format and require a significant amount of processing and analysis to gain insights.

End-User Platforms for Data Analytics

- **RITIS |** RITIS is a provider of transportation analytics that currently provides a suite of tools mainly suited toward vehicular speed/travel time and O-D data. However, CATT Lab staff noted that there also are efforts underway to bring multimodal data, such as transit data, into the RITIS platform; for example, MTA is paying for them to ingest AVL data from their buses (GTFS-RT feed).
    - **Applicability to Research Area |** RITIS could potentially be used to ingest transit data from the region; while the research team is unaware of any transit-specific tools within the platform, many users throughout the region have familiarity with the platform.
    - **Limitations to Research Area |** As noted, the research team is unaware of any transit-specific tools within the platform. Additional costs would be incurred to ingest this data and develop tools to view and analyze the data. These tools would be less proven than those from transit-specific analytics providers.
- **Swiftly** | Swiftly is a vendor of transit analytics who consume and analyze feeds from various AVL providers. They assist public transit agencies in enhancing their transit service by analyzing on-time performance and identifying operational issues. Furthermore, they can consume multiple agencies' feeds into one platform under a single login. Both the real-time and historical data can be accessed through their online dashboard and key-restricted API.
    - **Applicability to Research Area |** Their modules track transit headways, speeds, travel times, and runtimes by route, stop, day, and trip—metrics that are monitored routinely or in real-time.
    - **Limitations to Research Area |** It is a software platform that mostly serves public transit agencies and does not directly serve TPB's core functions.

- **Moovit** | Moovit is a mobility software company that offers both user-facing products (e.g., rider trip planning tools) and analytics for transit agencies. Their analytics tool, MUMA, combines multiple data sets (e.g., transit AVL and APC data; data from their user app) to provide insights into multimodal mobility throughout a region. MUMA's list of products and visualizations include O-D matrices, travel time, and modal splits.
    - **Applicability to Research Area |** This data could enhance the understanding of how alternative commuting modes or changes in commuting behavior (e.g., bike, walk, transit, rideshare, car/vanpool, and teleworking) affect the overall network with their focus on integrating real-time traffic, transit, and rail demand to one platform to provide insights into multimodal mobility. They have as much experience with MPOs as they do with transit agencies. Though they favor annual contracts as revenue streams, they have mentioned in the past to be willing to do small pilot projects at no cost.
    - **Limitations to Research Area |** Their apps are widely used in Europe and starting to grow in the United States. Their platform focuses on trip planning and MaaS, while the planning applications are largely untested.

Emerging Big Data Sources

- **Micromobility Data/Mobility Data Specification** | During the past few years, a variety of private companies have deployed fleets of scooters and bicycles that can be rented via mobile apps. Transportation planners across the country have desired to have data on the use of these "micromobility" services to better understand their use and interactivity with other modes. While many of the companies behind these services are unwilling to share much data, viewing this information as a trade secret, many jurisdictions have implemented contracts with these services that require sharing of some trip information, such as total numbers of trips, aggregated origins and destinations, or even information on individual trips.

3.3.5 TRANSPORTATION NETWORK COMPANIES

The following data needs were identified for transportation network companies by the TPB SWG.

- Data estimating the temporal distribution of TNC ridership based on rush hours, weekday/weekend, monthly, and quarterly.
- Data quantifying the magnitude of TNC trips and the O-D patterns in the region.
- Data supporting efforts to identify areas of the region in which TNCs are influencing VMT.
- Data supporting an assessment of the existing and future relationship between TNCs and transit. In other words, understanding where TNCs support and enhance transit use and where TNCs compete with transit mode choice.
- Data providing an estimate of demographics for TNC riders in the region, considering factors such as income, race, ethnicity, and gender.

APPLICABLE BIG DATA PRODUCTS

Of the evaluated Big Data products, a selection of products was identified as having the most potential for the TNCs evaluation research area. These products are summarized below and the comprehensive evaluation table is provided in **Appendix B**.

Mobile Device-Based Travel Patterns Data

- **StreetLight** | This vendor provides customized analysis of ride-hailing trips (not through the online platform). StreetLight provides an example case study in downtown Miami. These trips are inferred

from underlying data but are not explicitly provided to StreetLight as TNC-specific trips. The pricing structure for such customized analysis is unclear.

  - **Applicability to Research Area |** Customized analysis could provide data to infer the magnitude and the O-D patterns of TNC trips. Such information may be used to understand how TNCs are influencing VMT and the relationship with transit use. Underlying demographic information such as income and race could be inferred from census data.
  - **Limitations to Research Area |** This is not a common application offered by StreetLight. The TNC trips are inferred from underlying data. The research team has noted the black box nature of their proprietary machine learning algorithms, which present validity questions for customized analysis. Validation efforts and whitepapers are provided in **Section 3.2.1**. Additionally, customized analysis bears an unclear pricing and is usually more expensive than typical in-the-platform applications offered by StreetLight.
  - Please reference product validation limitations provided in **Section 3.2.1** for StreetLight Data.

- **Replica** | Replica provides analytics for a combined taxi/TNC mode similar to those it provides for other modes.
  - **Applicability to Research Area |** Taxi/TNC is a separate transportation mode used by Replica to model O-D demand down to the TAZ/census block group level. Highway traffic volumes can be filtered by this mode, and individual person trip data is available at the most granular level for this mode.
  - **Limitations to Research Area |** The accuracy of the TNC data relies on the quality of the activity-based model in representing TNC travel. TNC data are notoriously difficult to collect for model calibration. Also, TNCs are combined with taxis as a single mode.
  - Please reference product validation limitations provided in **Section 3.2.1** for Replica Data.

- **Uber Movement** | This dataset provides aggregated zone-to-zone travel time data derived from Uber trips freely available for the metropolitan Washington, area but only at the TAZ (District of Columbia only) or census tract (roughly the area contained by the I-495 Capital Beltway) level. This online tool allows a user to select a date/time range and pick origin/destination zones to view travel times between those zones. The data is available for select cities, including Washington, D.C. Data can be downloaded as a CSV for further analysis.

  - **Applicability to Research Area |** This dataset currently has limited applicability for meeting the research needs but is able to provide average zone-to-zone travel times for TNCs by time of day and day of week.
  - **Limitations to Research Area |** The data currently is not able to meet the research needs for TNC evaluation; however, it provides a source of travel time and speed data as TNCs are an active transportation mode in the metropolitan Washington area.
  - Please reference product validation limitations provided in **Section 3.2.1** for Uber Movement.

- **TNC Trip Data as Required by D.C. Law** | The District of Columbia has aggregated data for private vehicles-for-hire, such as TNCs, including Uber and Lyft, as information is required to be reported to DFHV and DDOT under D.C. law. A dashboard has been created to visualize the trip data, including the magnitude of TNCs trips, temporal distribution of TNCs trips, average wait-time (request-to-pickup), and number of unique operators. More granular data may not be publicly available but is being currently reported to the District government for anonymous individual trips.

  - **Applicability to Research Area |** This dataset can provide an improved understanding of the magnitude and temporal distribution of TNC trips within the District of Columbia (e.g., either end of the trip falls within the District). Other analyses could potentially be done using the more granular dataset that consists of trip distance, fare, pick-up/drop-off points etc. that is reported

to DFHV and DDOT. WMATA and TPB can request access to the data for specific purposes that are agreed upon between these agencies and the District government.

- **Limitations to Research Area |** Data access and sharing is limited based on confidentiality agreement. Data validity is unknown until agencies start to use this data. The dataset is limited to TNC trips that either start or end in the metropolitan Washington region. Without having seen this dataset, the research team assumes that it is likely provided in a raw format requiring data processing skills to aggregate and summarize.
- Given the widespread applicability of this product for this research area, further details are provided in **Table 3-7**.

Table 3-7 | Applicability of Taxi and TNC Trip Data to TNC Research Area

Data Need	Applicability/Limitations
Is Temporal distribution of TNC ridership	A subset of TNC data in 2019 is available in the dashboard to explore the temporal distribution of TNC trips. TNC occupancy ridership is not available in the dashboard and the more granular dataset.
Data quantifying the magnitude of TNC trips and the O-D patterns	A subset of TNC data in 2019 is available in the dashboard to explore the magnitude of TNC trips; however, TNC O-D data is not available in the dashboard and may be inferred from the more granular dataset (e.g., pick-up and drop-off points for each trip).
Identification of areas of the region in which TNCs are influencing VMT by investigating land use	Such information may be inferred from the more granular dataset (e.g., travel distance of TNC trips, pick-up and drop-off points for each trip) when overlaid with transportation network data and traffic volume data.
Understanding where TNCs support and enhance transit use and where TNCs compete with transit mode choice	Such information may be inferred from the more granular dataset (e.g., travel distance of TNC trips, pick-up and drop-off points for each trip) when overlaid with transit network data and transit ridership data.
Estimate of demographics for TNC riders in the region, considering factors such as income, race, ethnicity, and gender	Such information is not available.

Socioeconomic or Location-based Data

- **Household Travel Survey** | The HTS reveals insightful information containing behavioral characteristics of travelers and the relationship between travel decisions and travelers' demographic backgrounds. Based on initial results from the once-per-decade TPB regional HTS[79], the survey will reveal TNC travel characteristics that are not available from other sources, such as demographic information of TNC riders, the purpose of TNC trips, etc. Travel pattern data such as O-Ds will be available for a sample of the population. Starting in 2020, the NextGen household travel survey from FHWA will combine the core data, (e.g., household and person characteristics and self-reported trip purposes, collected through traditional household travel survey instruments) with O-D and other information inferred from Big Data sources to meet the ever growing and changing travel analysis needs. The NextGen household travel survey from FHWA will touch on emerging modes such as ride-hailing that were not captured in the previous HTS.

    - **Applicability to Research Area |** HTS data will reveal TNC rider demographic information and trip purpose information that is not accessible from the previously discussed private vehicle-for-hire data. It serves a specific research area of interest by TPB for estimating demographics for TNC riders in the region, considering factors such as income, race, ethnicity, and gender.

[79] www.mwcog.org

- **Limitations to Research Area |** The sample size of the data should be closely monitored to provide support or caveats for any conclusions drawn upon the survey data.
- Please reference product validation limitations provided in **Section 3.2.2** for Household Travel Survey.

3.3.6 TRAFFIC COUNTS

The following data needs were identified for traffic counts by the TPB SWG.

- Data containing permanent, temporary, or hourly traffic counts.
- Data that could represent longitudinal traffic counts; that is, hour-to-hour and day-to-day season-to-season variability of traffic counts. These data could then be used to better define the impact of land development and lane management policies over time.
- Data to supplement and validate existing traffic counts.
- Data containing vehicle classifications to supplement existing traffic counts without this distinction.

APPLICABLE BIG DATA PRODUCTS

Of the evaluated Big Data products, a selection of products was identified as having the most potential for the traffic counts research area. These products are summarized below and the comprehensive evaluation table is provided in **Appendix B**.

Mobile Device-Based Travel Patterns Data

Two mobile device products were identified by the research team that claim to offer traffic count data at daily and more granular intervals, including hourly. These traffic counts are synthesized using proprietary algorithms that scale sampled mobile device trip totals to population totals, often using public agency count data as control or calibration points. A significant body of research is ongoing to validate the applicability of using probe data for widespread, ubiquitous count data either in place of or supplemented by traffic counts from field sensors.

- **INRIX** | INRIX is a provider of vehicle probe data for segment-level congestion analytics as well as O-Ds for customized zones. INRIX has various product offerings, the most applicable to traffic counts is its Volume Profile dataset. INRIX data, through RITIS, can be queried to provide data for all facilities falling within a jurisdiction(s) or matching a specific facility type, potentially allowing for region-wide count estimates.
  - **Applicability to Research Area |** INRIX's Volume Profile data can hypothetically provide counts for all facilities in a region. This dataset allows for segment-level volumes to be queried for individual days and intervals throughout the day. INRIX also provides AADT estimates for roadway segments as part of the NPMRDS dataset. The TPB can currently freely access the NPMRDS probe data archive, including the associated AADT data. Further coordination with the state DOTs is needed to understand access to the Volume Profile data, and if any state DOTs are able to access this information.
  - **Limitations to Research Area |** As noted above, further coordination is needed to understand agency access to the Volume Profile dataset. A significant body of research is ongoing to validate the applicability of using probe data for widespread, ubiquitous count data either in place of or supplemented by traffic counts from field sensors.
  - Please reference product validation limitations provided in **Section 3.2.1** for INRIX data.
- **StreetLight** | StreetLight Data is an online platform for O-D or segment-based analytics based on mobile device data. Similar to INRIX, StreetLight offers a suite of analysis tools, the most applicable of

which are its AADT and segment analysis tools. Similar to INRIX, StreetLight can be queried to provide data for all facilities within a certain geographic area, allowing for region-wide performance measurement.

- **Applicability to Research Area |** StreetLight's AADT tool can hypothetically provide AADTs for all facilities in a region. Counts for segments or turning movements are available at intervals as granular as a single hour. It is unclear at this time if any agencies are using StreetLight for widespread traffic count data collection, especially at a sub-daily level.
- **Limitations to Research Area |** As noted above, further coordination is needed to understand agency access to understand the ability to query for widespread traffic counts, especially at the sub-daily level. A significant body of research is ongoing to validate the applicability of using probe data for widespread, ubiquitous count data either in place of or supplemented by traffic counts from field sensors.
- Please reference product validation limitations provided in **Section 3.2.1** for StreetLight Data.

- **Replica** | Replica is an online platform that provides analytics for individual roadway segments based on an activity-based model of a synthetic population.
  - **Applicability to Research Area |** The individual roadway segment data does include traffic volume estimates broken down by mode. Data is available at the segment level and can be filtered down to individual hours. Turning movement count estimates are not available at this time.
  - **Limitations to Research Area |** The segment volume data represents the modeled outputs from an activity-based model of a synthetic population, as opposed to sampled trips along a facility that have been scaled up to a "count" level by an algorithm. Internal validation reports are provided that note that passenger vehicle volumes are compared to averaged seasonal daily volumes collected at sensor locations in the region, similar to how volume estimates are validated from other Big Data sources.
  - Please reference product validation limitations provided in **Section 3.2.1** for Replica Data.

Data from Public Infrastructure

- **Continuous Traffic Count Station/Sensor Data |** From this data, agencies can estimate the average daily number of vehicles traversing roadway segments and VMT. Counts along local or secondary roads are typically derived from short-term (e.g., 48-hour) counts that are factored to an annual estimate based on nearby continuous count data; additionally, as agencies typically do not have the capacity to count every single road in their system annually, short-term counts may be taken on a less-than-annual basis and factored to current-year conditions.
  - **Applicability to Research Area |** In the absence of a Big Data solution for "ubiquitous" count data across a region, permanent count stations provide potential control points for calibration and validating Big Data solutions. Aggregated traffic count data for the region is already being compiled by TPB and is provided on the RTDC.
  - **Limitations to Research Area |** Agencies do not have the capacity to count all roads in their systems, even for short-term counts. It is the intention of this study to explore Big Data solutions to supplement or even replace counts taken by agencies. Continuous count stations also require continuous maintenance, upkeep, and replacement.

- **Automated Traffic Signal Performance Measures |** ATSPMs are emerging as an alternative to the traditional traffic engineering practice of ad-hoc turning movement count data collection, traffic signal timing, and arterial operations analysis.

- **Applicability to Research Area |** ATSPMs offer the opportunity for collection of turning movement counts at signalized intersections based on high-resolution detector data.
- **Limitations to Research Area |** ATSPMs are typically an undertaking at the individual agency level for the traffic signals that agency owns and maintains. These traffic counts would be limited to locations in which ATSPM signal controller configuration, communication, and software have been deployed; additionally, further backend processing is required to derive count data from raw ATSPM data.

End-User Platforms for Data Analytics

Several end-user platforms for data analytics are applicable to this research area given the underlying data they ingest, and the output analytics produced.

- **Product-specific dashboards |** StreetLight has a built-in dashboard are how all data is queried and obtained.
- **RITIS |** this platform is already being used regionally to obtain link-level speed data. Should TPB seek access INRIX Volume Profile data, this data will likely be accessed through RITIS.

3.3.7 CONNECTED AND AUTOMATED VEHICLES

The following data needs were identified for CAVs by the TPB SWG. The caveat was added that these needs will become more critical to the TPB when CAVs represent a more significant share of motor vehicles.

- Data containing O-D information for individually-owned and fleet CAVs.
- Data supporting research aiming to identify the impact of vehicle occupancy levels (e.g., the impact of zero-occupancy vehicles) and shared mobility services with the advent of highly-automated CAVs.
- Data supporting estimations of the impact CAVs may have on commuter behavior and mode selection (e.g., walking, biking, transit, rideshare, and teleworking).
- Data supporting estimations of the impact CAVs may have on growth management, specifically considering employer and residential location considerations.
- Data supporting forecasting efforts to estimate the impact of CAVs on parking policies and revenue generation.
- Data containing CAV performance information.

Connected and automated vehicle (CAV) technologies are an area of active research and development, as discussed in **Section 3.2.5 Emerging Big Data Sources**. Based on this understanding, it is the research team's recommendation to adopt a wait-and-see approach to this area. Rapid evolution is expected in the adoption of this technology. Additionally, given the amount of sensor data that needs to be processed to make this technology work, the vehicles themselves may yield new Big Data products that are useful for many applications in the future. While the research team does not currently recommend that the TPB invest in a solution, it does recommend that the TPB take a proactive approach to monitoring the development and progress of these data sources and ensuring regional coordination on their applications.

3.3.8 OTHER RESEARCH AREAS

The following data needs were identified for other research areas by the TPB SWG.

- Data supporting the estimation of current and future freight and commercial vehicle travel within the region.

- Data that can be used to validate findings from household travel surveys (e.g., NHTS and Washington Regional Travel Survey).
- Data that can be integrated into HTSs to more frequently generate long-distance and local travel behavior.
- Data that can be used to research travel patterns and mode choice based on community demographics.
- Identifying how new sources of Big Data could be used to inform transportation-related project selection in the TPB.

APPLICABLE BIG DATA PRODUCTS

Mobile Device-Based Travel Pattern Data

- **INRIX** | INRIX is a provider of vehicle probe data for segment-level congestion analytics as well as O-D's for customized zones. INRIX has various product offerings: probe data (segment-level speed/travel time/AADT estimates), O-D summaries (trip starts/ends), and trip paths (includes waypoints along routes and is a very large dataset).
    - **Applicability to Research Area |** INRIX O-D data could be used to validate findings from the HTS and potentially could be integrated within the HTS to expand the temporal and geographic coverage of the traditional responses. INRIX O-D data can be integrated in the RITIS platform to visualize and analyze data. INRIX data is available for medium and heavy trucks as separate datasets.
    - **Limitations to Research Area |** INRIX O-D data is limited to vehicular trips. The O-D data provided by INRIX is a newer product that has not undergone as extensive external validation procedures as other INRIX products (e.g., travel speeds and traffic volume data). While the TPB has access to INRIX speed and congestion data through NPMRDS, INRIX O-D data would need to be purchased separately. The TPB could leverage partnerships with Maryland DOT and DDOT for access to their existing O-D data; however, these data sets do not cover the same temporal timeframe and couldn't be used together.
    - Please reference product validation limitations provided in **Section 3.2.1** for INRIX data.
- **StreetLight** | StreetLight Data is an online platform for O-D or segment-based analytics based on mobile device data. Similar to INRIX, StreetLight offers a suite of analysis tools: AADT estimation, O-D, O-D with middle filter (through a midpoint location), O-D to preset geography (e.g., TAZs, zip codes, census block groups), top routes between O-D's, and a congestion diagnostics tool for auto-generating insights.
    - **Applicability to Research Area |** StreetLight O-D data could be used to validate findings from the HTS and could potentially be integrated within the HTS to expand the temporal and geographic coverage of the traditional responses. The additional information StreetLight provides over INRIX (e.g., income levels, trip purpose, and bicycle/pedestrian O-Ds) could be further advantageous for understanding travel patterns and mode choice based on community demographics. Additionally, StreetLight O-Ds can be extracted for heavy and medium trucks, which may be used for estimating current and future freight and commercial vehicle travel within the region. StreetLight has an online platform with visualization features that can be used to explore and summarize their data.
    - **Limitations to Research Area |** Acquisition of this data product is very expensive. Similar to INRIX, data sharing agreements could be developed to share StreetLight data between partner agencies (e.g., VDOT has an active subscription). However, the subscription from VDOT would not cover the rest of the region in Maryland, the metropolitan Washington region., and West Virginia.

- Please reference product validation limitations provided in **Section 3.2.1** for StreetLight data.

- **Teralytics** | Teralytics is an online platform for O-D analytics based on cell phone tower triangulation data. The company claims to have a less biased data sample than LBS-based analytics providers and they claim sufficient market share among all demographics, ethnic groups, income levels, and age groups.
  - **Applicability to Research Area |** Since Teralytics uses cell phone tower triangulation data, they are able to sustain a longer-term trace of individuals to ascertain trip purpose and trip frequency. These data would be valuable for identifying trip purpose as well as identifying the number of trips being made between O-D pairs on different days of the week. These data could be used to validate findings from the HTS and could potentially be integrated within the HTS to expand the temporal and geographic coverage of the traditional responses.
  - **Limitations to Research Area |** Cell tower triangulation has a high sample rate (estimated at 15 to 35 percent of the population) but a lower spatial resolution. Thus, this data is not usable for route-level analysis (individual trips cannot be mapped to specific roads), but it is usable at the census tract or even TAZ level. While Teralytics claims to have a high level of accuracy given its deep penetration rate, limited validation efforts exist to confirm data validity as most benchmarking appears to be done internally.
  - Please reference product validation limitations provided in **Section 3.2.1** for Teralytics.

- **Replica** | Replica is an online platform for trend analysis on mobility, economic activities and COVID-19, and in-depth travel data synthesized by an activity-based model. The in-depth travel data includes O-D data for a rich set of modes and trip purposes that can be further partitioned based on several other variables, such as trip length/duration and start time. Replica's activity-based model also converts O-D data to highway traffic volumes and transit ridership.
  - **Applicability to Research Area |** Trip attributes (e.g., modes, trip length/duration distribution and start time) and traveler attributes (e.g., income levels, trip purpose, age, gender and race/ethnicity) can potentially be applied to several of the other research areas noted in this section. Synthetic population data is available at the individual trip level or at larger census-area levels. Commercial vehicle trips are included as a mode for analysis. In addition, because the Places model size extends well beyond the MWCOG planning area, the platform could be used to understand external travel.
  - **Limitations to Research Area |** Most Replica data is synthetic in nature, including all individual person trip data. The accuracy of the data relies heavily on the availability of ground truth data to train and calibrate the underlying activity-based model which, to a large extent, is a "black box" to the data users. However, this is also true for the other vendors that use their proprietary data algorithms to process and impute the data for their platforms.
  - Please reference product validation limitations provided in **Section 3.2.1** for Replica

- **Locus** | Locus is a product of LBS-based travel analytics product provided by Cambridge Systematics (consulting firm that has been MWCOG's travel demand model developer). Four separate products are offered in addition to the custom analyses: 1) O-D tables (expanded/validated by travel purpose/time-of-day), 2) a transit competitiveness dashboard, 3) a geofence analysis of activity around activity centers, and 4) a survey assist to supplement traditional HTS data.
  - **Applicability to Research Area |** Locus O-D trip tables show multimodal trips including trip purpose and time of day, which could be used to validate findings from the HTS and could potentially be integrated within the HTS to expand the temporal and geographic coverage of the traditional responses. Due to their customizable business model, where Cambridge Systematics is offering both a Big Data product as well as consulting services, it is possible

that the resulting data could be designed specifically for comparison and integration with the HTS.

- **Limitations to Research Area |** O-D trip tables are not separated by mode. Validation is a trade secret. The base product and corresponding consultant services are relatively expensive compared with the other O-D products. The primary driver of the high costs is consulting labor fees required to develop project-specific dashboards.
- Please reference product validation limitations provided in **Section 3.2.1** for Locus.

3.4 SUMMARY OF EVALUATION FINDINGS

The research team identified several high-level trends throughout the course of evaluating the various Big Data products:

- **There is no one size fits all solution to all of the TPB's research area needs**. Each of the Big Data products evaluated had unique advantages and disadvantages. For example, several products offer O-D data for various modes (e.g., rail transit, bicycle and pedestrian trips, truck trips), and several products provide trip purpose in some fashion. However, all of the products are limited in some capacity in their offerings in regard to TPB's research needs—such as only offering data for certain modes or certain trip purposes. In particular, the research team feels that mobile device-based Big Data sources face challenges that go beyond differentiating between all multimodal trips and trips by various sub modes, with perhaps the exception of truck trips. Additionally, while Big Data algorithms lend themselves to being able to identify home and work locations for trips, it becomes much more challenging to parse out other trips. Ultimately, the various products evaluated often differ in data resolution and level of detail (e.g., traffic patterns by mode, daily versus hourly versus sub-hourly counts, spatial accuracy in terms of adjacent road facilities of general purpose, HOV, express lanes) as well as data validity and transparency. Some products have more transparent validation procedures whereas others maintain their benchmarking internally.
- **The landscape of Big Data products is constantly evolving**, even throughout the course of this study. The products evaluated generally fall somewhere along the a spectrum as shown in **Figure 3-9**. Typically, products from novel or emerging datasets, such as CAV data or micromobility trip data, fall toward the left side of the spectrum. Data from these products is often in a very raw format requiring significant data storage and processing abilities in order to glean information. As various products mature and gain more widespread use, dashboards and visualization tools for more easily interpreting data are offered. These end-user analytics platforms can be built within a vendor's product offering or may be provided by third parties that ingest data from other sources (such as RITIS). Vendor products are continuously evolving across this spectrum. For example, during the course of this study, several new tools were deployed from various vendor offerings, such as StreetLight's Top Routes tool or the Signal Analytics offering from RITIS. FHWA is developing open source algorithms[80] to analyze raw mobile device-based travel pattern data for O-D information by mode in support of the Next Generation travel survey. Maryland Transportation Institute (MTI)/CATT Lab/University of Maryland has developed an end-user platform[81] that analyzes mobile data provided by the major teleproviders, report social-distancing index, and other TDM metrics during the COVID-19 pandemic. Certain datasets are much less mature in their evolution, such as CAV data and micromobility data. The research team expects that the landscape for these products will evolve during the next several years.

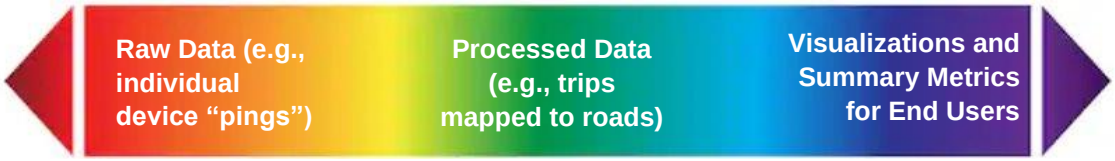


Figure 3-9 | Big Data Product "Spectrum"

- **Opportunities for collaboration and data sharing exist** to aid in procurement and downstream processing of data. For data sharing, opportunities exist for expanding a dataset's coverage from current state agency agreements with vendors, such as VDOT's current agreement with StreetLight or MDOT's current agreement with INRIX. The TPB plays an important role as a central repository of

[80] www.fhwa.dot.gov
[81] data.covid.umd.edu

planning data for the region, such as its current traffic count data warehouse. This could be expanded to incorporate other sources of data; in particular, TPB can aid in regional transit planning given the numerous providers of service in the region, and end-user analytics platforms could potentially be used. The TPB also certainly faces challenges given the multijurisdictional and multistate makeup of the region; for example, a dataset being purchased by one jurisdiction or state will be limited to that jurisdiction or state. Datasets derived from agency deployments, such as ATSPMs or transit data, will vary region-wide by vendor, level of investment, and agency expertise. Finally, opportunities for collaboration and data sharing also exist within MWCOG itself, such as among the various boards and working groups as well as the committees within TPB.

- **Agency-deployed field infrastructure is still needed for ground-truth data**. While this study sought to examine the potential for Big Data to validate field data, the findings showed that, at this stage, field data is overwhelmingly used as ground-truth data to validate Big Data. For this reason, continued investment in upkeep and maintenance of field infrastructure such as traffic count stations, traffic signal infrastructure, and transit ITS systems is warranted.
- **Certain datasets likely exist but are not yet available for public agency use**. These datasets include tolling and pricing data from privately-owned and operated managed lanes facilities in Northern Virginia, data on TNC trips (outside of those reported to the District of Columbia), and similar data on micromobility trips (e.g., scooters).
- **The TPB can take a proactive approach in monitoring the development and progress of various emerging data sources and ensuring regional coordination in their applications**. Emerging data sources that the TPB should focus on include taxi/TNC data, micromobility data, and CAV data. The more that the jurisdictions can coordinate to encourage common data standards and reporting, the more useful this data will be to planners in the region.

3.4.1 OPPORTUNITIES AND LIMITATIONS

This independent evaluation of Big Data products was focused on finding solutions to the TPB's designated research areas. Therefore, the products with the most direct applicability to those research areas were further investigated and additional detail provided for each. In some cases, additional product data was not available from vendors without signing a non-disclosure agreement; this was especially true for pricing information.

Table 3-8 provides a summary of the opportunities and limitations of Big Data for each of the TPB's SWG research areas.

Table 3-8 | Opportunities and Limitations by Research Area

Research Area		Description of Big Data Opportunities and Limitations
1	Travel Demand Modeling	• Several platforms exist for providing vehicular O-D data at various levels of granularity: for example, some products provide this data at the TAZ or census tract level, while other platforms offer O-D data for specific roadway segments (e.g., INRIX, StreetLight, Replica). The most common agency applications for O-D data for travel demand modeling is for validating model survey data, especially external travel. • O-D information for full set of transportation modes has become available from some (e.g. Replica) but not all data vendors • Trip purpose data is inferred based on device movement characteristics; as such, ascertaining a location other than home or work to a device for other trip purposes remains a challenge. However, Replica seems to have some success in this regard using a complex data integration and synthetic population modeling approach. • The accuracy and availability of O-D patterns by route and road segment is subject to the resolution of data needs; for example, there is currently no cost-effective means to collect accurate O-D data for facilities that are closely-spaced (e.g., Express lanes and general purpose lanes that are only separated by lane markings or barriers). O-D data for lower-volume facilities may suffer from low sample rates. • While daily and sub daily traffic count data is offered by several Big Data vendors; such information is less readily available or valid for lower-volume roadway facilities at this time. • Region-wide signal delay estimation/queue length is not available for incorporating into a travel demand model at this time but may evolve in the near future. This information can be potentially obtained from ATSPMs, but the deployment of ATSPM infrastructure and systems would be on an individual agency level and either has not started or is still in its early stages in the region. Data from CAVs with high temporal resolution (less than 3 seconds) offers the potential for obtaining these measures but is unproven at this time.
2	Travel Demand Management	• Evaluating the influence of telework policies on travel behavior relies on present and historical travel patterns data; it is not real-time as it takes time for travel behavior to change and the data needs to be archived over time. • The travel behavior information, depending on the resolution desired—mode of travel, home/work locations—may have limitations and caveats as outlined above under travel demand forecasting. For example, home/work location data by TAZ is much more reliable than route or mode data. • Additional contextual data and customized analysis may be needed to differentiate visitors versus home-based workers.

Research Area		Description of Big Data Opportunities and Limitations
3	System Performance/Congestion Management	• Big Data is widely used for system performance/congestion management for vehicular travel on freeways and higher-volume arterials. The validity and accuracy of data for non-NHS roads is subject to further evaluation, particularly for rural areas, because of sample rates of mobile device-based data and the number of road sensors deployed on these facilities. • Multimodal system performance data using mobile device-based travel pattern data is currently unattainable or very novel due to difficulties to chain the trips by different modes as well as to differentiate between certain modes (e.g., bus as mentioned above). • Monitoring network-wide congestion and performance can be achieved using a combination of mobile device-based travel pattern data and data from public infrastructure to produce different congestion and performance metrics; several end-user platforms are widely used for this purpose. This requires agencies to continue to invest and maintain this public infrastructure including traffic count sensors, traffic signals, and ITS devices, including those that are part of transit systems.
4	Transit and Active Travel	• Monitoring real-time transit performance is feasible using AVL data, and some end-user platforms for data analytics provide both hardware and software to report real-time performance; however, integrating data across transit agencies can be challenging, as they may use different end-user platforms and may not share standardized data formats for easy exchange of information. • Monitoring active travel demand and delay (e.g., for bicycles and pedestrians) using Big Data for these modes is fairly novel, and validation is limited at this time, in part due to a lack of validation data available from public infrastructure.
5	Transportation Network Companies (TNCs)	• There are no proven and large-scale methods to understand usage and demand for TNCs given the proprietary nature of these datasets; some Big Data vendors offer custom analyses attempting to infer TNC trips based on device movement characteristics • TNCs are required by D.C. law to report detailed trip data to the D.C. government, from whom TPB could potentially request access for agreed-upon internal analysis purposes. The details for how to access the data are contained in the agreement between TPB/WMATA and the D.C. government. • Socioeconomic and demographics of TNC riders are only available from regional or NHTSs. • The relationship between TNCs and transit (e.g., first/last mile problem) could be inferred using the datasets mentioned above but would not be straightforward at this time • Emerging end-user platforms for data analytics, such as SharedStreets, allow for summarizing TNC and micromobility data; DDOT is currently using this platform.
6	Traffic Counts	• Mobile device-based Big Data counts are still being validated by counts collected using traditional methods, not the other way around. Continuous count data from road sensors plays a critical role in providing control points for mobile device-based Big Data solutions. and is anticipated to continue to do so in the future. These control points are needed to validate Big Data solutions as they are continuously being refined to estimated ubiquitous counts. • Variability analysis of traffic counts (e.g., day-by-day) is obtainable using several Big Data products, subject to the caveat of limited sample sizes on

Research Area		Description of Big Data Opportunities and Limitations
		facilities for individual days, particularly for lower functional class roads and facilities in rural or less dense areas. • Vehicle classification data is heavily dependent upon data from road sensors; Big Data products do not provide sufficient information about the vehicle classification that is commensurate with road sensors (e.g., continuous count stations). • Big data solutions currently cannot collect vehicle occupancy data; the road sensor technology for collecting vehicle occupancy data is available but the accuracy is not there yet, especially for tolling industry deployment. • Mobile device-based Big Data products could be used to support HPMS for certain data types, but information such as roadway conditions and heavy vehicle percentages are difficult or unattainable via Big Data.
7	Connected and Automated Vehicles	• Implementable CAV data for widespread understanding of travel patterns and traffic operations is not yet available. The research team feels that the most promising near-term data sets are likely coming from third-party solutions integrating high-resolution data from vehicle OEMs, which report data in very short time intervals (3 seconds or less) and provide additional data relating to vehicle trajectory (e.g., braking/acceleration).
8	Other Research Areas	• Regional freight and commercial vehicle travel O-D data is offered by Big Data vendors for medium and heavy trucks. • Household survey data validation and supplementation can rely on mobile device-based travel pattern data. For example, FHWA is embarking on the Next Generation NHTS using mobile device-based Big Data products. • Big Data vendors that provide O-D patterns based on population demographics are inferring this data from underlying census information.

CHAPTER 4

RECOMMENDATIONS

4 RECOMMENDATIONS

4.1 INITIAL OPTIONS AND PARTNER AGENCY INPUT

Following the State-of-Practice review described in **Chapter 2** and the Big Data Product Evaluation described in **Chapter 3**, the project team developed an initial set of procurement options and conducted another round of outreach to staff at DDOT, MDOT, VDOT, Virginia DRPT, and WMATA. A detailed overview of the initial options that were considered prior to further agency input (noted in **Section 2.3**) is provided in **Appendix D**. Ultimately, these options were narrowed based on agency feedback of lessons learned and TPB staff indicating that the highest-priority research area needs required a customizable, granular mobile-device-based dataset that could be used for analyses of multiple travel modes.

*Therefore, the project team recommends that the TPB move forward with a **pilot procurement** of either **StreetLight Data** or **Replica**, taking into consideration the procurement options and strategies laid out in the following sections.*

STREETLIGHT DATA
Big Data for Mobility

REPLICA

Figure 4-1 summarizes the overall evaluation and outreach process that led to the recommended procurement options.

Figure 4-1 | Formulation of Recommendations

4.2 IMMEDIATE PROCUREMENT PLAN

4.2.1 PILOT PROGRAM APPROACH

Both StreetLight and Replica offer enterprise subscription models in which a large number (or potentially unlimited number) of users can access their products and conduct unlimited analyses. Both services offer pricing based off of the population of a region. Both products offer the option for extending access to partner agencies under an agency's geographic umbrella, although each product is unique in how this may affect pricing. Both of these products would represent a significant investment for the TPB or the region.

For the initial procurement, it is recommended that a **pilot program approach** be taken to allow the TPB to trial one or two products prior to the agency making a large investment. Three pilot procurement options were developed that meet the greatest number of TPB data needs.

- **Option 1** | Informal Trial of StreetLight
- **Option 2** | Formal Trial of Replica
- **Option 3** | Informal Trial of StreetLight & Formal Trial of Replica

OPTION 1: STREETLIGHT

As described in the report, StreetLight was rated well during the independent evaluation for offering numerous products that met the data needs identified by the TPB. Additionally, partner agencies in the Metropolitan Washington region have experience using StreetLight and have made substantial investments in the product.

Informal Trial: As an established company in the industry, StreetLight is not likely to offer a free data trial. Therefore, it is recommended that the TPB explore options for conducting an informal trial of StreetLight product offerings to determine whether a larger long-term investment of the data is warranted.

Evaluation Strategies: In order to evaluate the applicability of StreetLight products for the TPB's research areas and needs, it is recommended to select one or more specific projects with clear applications and objectives to focus on during the trial period. The scope of these projects should be used to determine how the data is acquired, which data is most appropriate for supporting the purpose and need of the project(s), and how much data is needed to achieve a valuable result.

Data Acquisition: The team identified two primary options for procuring a sample of StreetLight Data to support this initial pilot procurement.

- First, the TPB can request free access to the VDOT StreetLight subscription (subscription allows free access to partner MPOs). The VDOT subscription will not provide data for trips in Maryland and Washington, D.C., that do not touch Virginia, but it includes a 15-mile buffer around the commonwealth boarder. Therefore, TPB staff could set up analyses within with VDOT subscription in this constrained geographic area to gain familiarity with the latest StreetLight product offerings and further understand the strengths and limitations of the tool. The TPB could also use this data to perform validation analyses on the data provided to determine whether a larger investment in the product would provide sufficient value to the TPB.
- Second, the TPB can define a series of zones within the Metropolitan Washington region and purchase a finite amount of data for those zones directly from StreetLight. For example, 50 zones can be defined within the region and data could be requested for each season of the year (i.e., spring, summer, fall, winter). This data set should be available for approximately $25,000 based on preliminary conversations with StreetLight sales, assuming the TPB will want to obtain data for various non-vehicular modes (e.g. pedestrians, bike, bus, rail).

OPTION 2: REPLICA

The second option is completing a formal trial of a new Big Data product, Replica, which also was identified during the evaluation as being a promising product for addressing the TPB's data needs. Replica is a new online platform that models "synthetic population" of a given area at a very detailed level for a wide range of modes. Replica was formed as part of Alphabet, Google's parent company, in 2018, and became an independent company in 2019. The underlying dataset includes LBS and GPS data, consumer resident data, land use and real estate data, credit transaction data, and ground truth data (such as traffic counts and transit boarding). This online platform includes three main products:

- **Places |** a series of detailed, activity-based travel models for specific regions at specific points-in-time, with data down to the network-link level. Replica currently offers a Baltimore-Washington "Places" model, which includes the entirety of Maryland, Virginia, and the District of Columbia.
- **Trends |** a higher-level dashboard comparing mobility patterns over time, with metrics including: total trip counts, mode split, the percentage of people commuting to work, the percentage of people sheltering-in-place, and consumer spend, as well as COVID-19 data. Trends data is available for the entire US and can be obtained at the census tract, city, county, MSA, or state level.
- **Scenarios (not yet available) |** "what-if" analysis for changing network links/land use.

Additional detail on Replica product offerings can be found in the final report.

Formal Trial: Due to their new entry into the Big Data market, Replica is offering agencies trial packages to increase exposure to their products. The TPB could take advantage of these formal trials offered by Replica to learn more about their product offerings and perform validation analyses on the data provided to determine whether a larger investment in the product would provide sufficient value to the TPB.

Evaluation Strategies: The recommended evaluation strategy for Replica is similar to that suggested for StreetLight; however, because an actual trial version of the Replica platform is likely available, the TPB may have less ability to define the exact data acquired and will likely need to perform the evaluation at a more rapid pace. For these reasons, it is recommended to open conversations with Replica to define the scope of the trial and determine which products will be included and for what duration. Next, prior to the trial beginning, the TPB should define a handful of specific projects or research questions that will be evaluated as part of the trial. This way, when the data becomes available, the TPB will be able to make efficient use of this pilot evaluation.

Data Acquisition: The TPB can likely reach out directly to Replica to initiate a formal trial of the platform. The project team is aware of other agencies, including DDOT, being introduced to or conducting trials of Replica, but it is unclear at this time if those trials are still ongoing. All data must be obtained through the online platform but can be downloaded for further analysis if needed. The project team has explored the synthetic population data from the Places model and notes that the individual person-trip tables are very large files.

OPTION 3: STREETLIGHT & REPLICA

The third option combines the first two options, offering the TPB the opportunity to create an informal trial of StreetLight, while also conducting a formal trial of Replica. This option would allow the TPB to conduct similar analyses on both data sets to compare their effectiveness for addressing data needs and identify which platform may offer greater long-term benefits for the agency.

Evaluation Strategies: The recommended evaluation strategy for the combined acquisition of StreetLight and Replica is similar to those recommended for both independently. To achieve the greatest comparison of the two data products, it is recommended to select a few projects or research questions that can be evaluated with both vendors. Since the trial with Replica may be a formal trial, there is likely less flexibility

in which data can be provided; therefore, it may be advantageous for the TPB to identify the scope of the Replica offering, and then seek to procure a similar data sample from StreetLight.

4.3 PROCUREMENT OPTIONS, CONSIDERATIONS, AND STRATEGIES

In the near-term future, the TPB could also take other approaches in parallel or consider the following:

- **Request for Information (RFI)** | The TPB should consider issuing a Request for Information (RFI) to the vendors of the preferred Big Data products to gain further insights on product-specific strengths, limitations, pricing models, and data sharing opportunities. An RFI could be used to obtain detailed pricing quotes including accounting for collaboration with other agencies; formal data-sharing agreements; and the most up-to-date analytics or data available, including custom analytics that may not be available through a traditional subscription.
- **Collaborative versus Independent Procurement** | If partner agencies in the Metropolitan Washington region are able and willing to collaboratively invest in a particular data product, this process of product procurement would involve coordination between agencies to collect an agreed-upon data subscription or set of historical data.
- **Direct versus Open Procurement** | A direct procurement would offer a more expedited schedule and a generally simpler procurement process; an open procurement may offer additional competition resulting in lowered prices (not guaranteed). An open procurement would expend additional COG/TPB resources completing the request for proposals (RFP) process, although it may be a necessity depending on cost of data procurement or pooled funding requirements with partner agencies.
- **Obtaining TNC Data** | Another related near-term action the TPB could take is obtaining the existing TNC datasets from the DC government. To do this, the TPB would need to develop an agreement with the DC government regarding the purpose of the data access. In procuring this dataset, the TPB would be able to fully explore the available information and evaluate the limitations in using this data to address the SWG's identified research area needs. In addition, TPB could leverage the end-user platform, such as SharedStreets, to visualize TNC datasets.

4.4 LONG-TERM RECOMMENDATIONS

The research team also identified a number of long-term recommendations for the TPB to consider as the landscape of Big Data continues to evolve.

- **Identifying Promising Research Leading to Practice**. Promising research is currently being sponsored through FHWA and other entities, such as universities and research institutes, to reduce the unknowns in using Big Data and introduce transparency to the analytic process. These projects are resulting in open-source code and documentation that sheds light on details of Big Data processing procedures and introduces new opportunities for processing raw data in the future. While these projects are still at a highly granular level and may not be ready for the TPB's application at the time of this report, they hold promise and could be evaluated further as research evolves into practical application. The TPB should partner with universities in the region to continue carrying out research on Big Data.
- **Staying Proactive**. As the landscape of Big Data products continues to evolve at a rapid pace, the TPB should take a proactive approach to monitoring the development of new products, innovative technology in data acquisition, and new product validation studies. This could be done by conducting an annual or biannual survey of peer agencies and reporting updates of this Big Data product evaluation, specifically geared towards identifying how the industry is embracing new Big Data products (e.g., new product vendors within existing data categories as well as data from completely new data streams, such as CAVs, micromobility, TNC, and gig travel). This continuous update also would enable

the TPB to stay up to date on the latest experiences with common Big Data products and recent validation information from research and practice.

- **Regional Advocate for Big Data and the Importance of Maintaining Infrastructure**. Finally, the TPB can be a regional leader advocating for the continued investment in transportation infrastructure that supports the use of Big Data. Currently, Big Data validation relies on the baseline data provided by transportation infrastructure; therefore, the continued maintenance of this infrastructure is imperative. The TPB can actively support continued investment in regional traffic count stations, sensors for monitoring bicycle and pedestrian activities, multimodal traffic signal infrastructure and communications, transit fare collection and counter technology, advanced transit systems (real-time performance, scheduling, crowding), and further ITS infrastructure. The TPB can also advocate for data standards for various Big Data sources (for example, the Mobility Data Standard for micromobility data). Finally, the TPB can continue to be an advocate for the behind-the-scenes "infrastructure" for storing and archiving data, such as bringing Big Data products or outputs into the Regional Transportation Data Clearinghouse or another regional solution facilitating open data access.

APPENDIX A

PEER AGENCY SURVEY RESPONSES

APPENDIX B

EVALUATION DOCUMENTATION

APPENDIX C

LOCAL MEETING SUMMARIES

APPENDIX D

COMPREHENSIVE PROCUREMENT OPTIONS