Current Status of Transportation Data Analytics and A Pilot Case Study Using Artificial Intelligence (AI)

Task 1 Draft Report:
Review of Current Data Collection and Utilization Practices

Prepared by:

Yuanchang Xie[1]
Danjue Chen[1]
Tingjian Ge[1]
Ali Shirazi[2]

[1] Civil and Environmental Engineering
University of Massachusetts Lowell
Lowell, MA 01854

[2] Civil and Environmental Engineering
University of Maine
Orono, Maine 04469

Prepared for:

New England Transportation Consortium (NETC)

November 2021

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Data is playing an increasingly important role in TSMO applications. Understanding the available data sources and their strengths and weaknesses is important for state DOTs. This report starts with summarizing the traditional and emerging data sources that can be used for TSMO applications, including loop detectors, Bluetooth sniffers, E-ZPass transponders, drone, LiDAR, Automated Traffic Signal Performance Measures (ATSPM) system, connected vehicles, automated vehicles, and data generated by Artificial Intelligence (AI) algorithms. The strengths and limitations of these data sources are also discussed. A detailed analysis of some new data sources is provided at the end of this report, including the National Performance Measurement Research Data Set (NPMRDS), Wejo, Otonomo, and StreetLight. Based on the review of data sources, a list of potential TSMO applications is identified in Sections 2. These potential applications will be further assessed and described in the report for Tasks 2 and 3.

As part of Task 1, the research team interviewed staff from six state DOTs in the New England region and some state DOTs in other regions. The interviews covered TSMO related questions on data needs; emerging data sources; data integration and analysis; data archiving, sharing, security, and privacy; and workforce development. The findings of the interviews are documented in Section 3 of this report. A summary of the findings are presented in the report for Tasks 2 and 3.

1. DATA AND DATA SOURCES

The analysis of data and data sources consists of two sections. The first section focuses on traditional data and data sources, and the second one is for new and emerging data and data sources.

1.1. Traditional Data and Data Sources

1.1.1. Highway Data

Traditional highway data sources mainly include loop detectors, microwave detectors, CCTV traffic cameras, Bluetooth/Wi-Fi MAC address readers. From these sources, occupancy, delay and travel time, spot and segment speeds, volume can be obtained. Many agencies also get data from weather stations and weigh-in-motion stations. Since such data are quite different from traffic flow parameters, they are discussed separately in Sections 1.1.2 and 1.1.3, respectively.

1.1.1.1. Inductive loop detector

Inductive loops are one of the most common data sources for highways. They have been widely adopted by state Departments of Transportation (DOTs) to collect traffic count, speed, length (if dual loop detector), occupancy, etc. on highways. They have also been extensively used at intersections to provide input data to traffic signal controllers.

These loop detectors are less sensitive to the environment (e.g., temperature, lighting, snow, strong wind, vibration) and provide robust traffic measurements. However, since they are installed underneath the pavement, it is difficult to repair them if broken. Another major issue is that such detectors often are used to generate AADT data to meet the Highway Performance Monitoring System (HPMS) reporting requirements. The generated traffic measurements typically are not streamed in real time to Highway Operations Centers (HOC) or Traffic Management Centers (TMC), making them unsuitable for real-time applications such as incident detection and response.

Additionally, these detectors are installed at limited locations on major highways and intersections. Therefore, they can only provide situational awareness for highway segments near those locations. For traffic incidents that happen far away from those locations, they will not be detected in a timely manner, which is critical to emergency response. Even if an incident is detected, it is difficult to accurately estimate its location with a sparse inductive loop detector network. Again, knowing the location of an incident is very important for efficient emergency response operations.

The above issues can be addressed by adding more loop detectors and investing in communication and IT infrastructure. For example, Caltrans maintains a PeMS (Performance Measurement System), which consists of about 40,000 detectors covering freeways across all major metropolitan areas of California, providing both real-time and historical traffic data. However, the cost for doing so can be prohibitive, especially for many New England state DOTs with a significant portion of their highways in rural areas.

1.1.1.2. Microwave sensor

Similar to inductive loop detectors, microwave sensors are installed at limited locations. Also, the collected data sometimes are not streamed in real time to HOC or TMC. In this sense, microwave sensors share the abovementioned limitations of inductive loop detectors. However, compared to inductive loop detectors, microwave sensors are easier and less expensive to install and maintain. They are installed on roadside poles and the installation and maintenance cause no or little impacts on traffic. Some new microwave sensors can each cover more than 20 lanes and over 200 ft of road segments. They can also track individual vehicles, detect lane changes, and measure vehicle length in these segments, while inductive loop detectors can only measure traffic at a single point or over a very short segment (e.g., 20 ft).

1.1.1.3. CCTV camera

All state DOTs operate and maintain a CCTV camera network. For example, RIDOT has about 200 cameras (many of them are around rest stops) and plans to add more. These cameras provide important video feeds for identifying and confirming traffic incidents. However, in most state DOTs such CCTV traffic videos are reviewed manually to confirm traffic incidents (detected/reported using other methods) and provide traffic situational awareness. Such traffic videos typically are not recorded. They are not utilized to automatically detect traffic incidents, although technically it is possible to utilize video image processing algorithms to process live CCTV camera feeds and generate data such as vehicle count, speed, and density for data collection and detecting incidents.

Like inductive loop and microwave detectors, CCTV cameras are deployed at limited locations, although they are getting increasingly popular. One concern with CCTV is privacy, especially for high-definition cameras. Such a problem can be addressed in many ways. One solution is to utilize edge computing devices to process videos in the field without saving them (i.e., only keep and stream the extracted traffic measurements). With the wide deployment of CCTV cameras and adoption of Artificial Intelligence (AI) based video processing algorithms, CCTV cameras may potentially become a major source in the future for traffic data collection and incident detection.

Some toll road authorities are using high-definition CCTV cameras for toll by plate purpose. This application can generate Origin-Destination (OD) and segment travel time data beyond count, speed, and density. Such travel time data allows HOC operators to identify congested segments. However, it cannot provide much useful information related to location (e.g., where a congested segment starts and ends), unless the distance between upstream and downstream cameras is sufficiently short.

Some New England state DOTs have investigated the possibility of turning existing cameras into traffic sensors for traffic data collection and incident detection. A few issues they encountered include the low resolution of existing cameras, Pan-Tilt-Zoom (PTZ) cameras making it difficult to calibrate them, preferring a central video processing solution than adding video processing hardware to individual cameras, etc.

1.1.1.4. Bluetooth data

Bluetooth technology has been widely used in collecting travel time data. It detects the Media Access Control (MAC) addresses of Bluetooth devices on passing by vehicles and matches upstream and downstream MAC addresses to derive travel time. This is like matching upstream and downstream license plate numbers as some toll road authorities are doing (see discussion in Section 1.1.1.3 above) to determine the toll rate for charging users. A main difference is that Bluetooth readers are less expensive and do not require sophisticated data processing algorithms (e.g., AI algorithms for detecting and recognizing license plates).

Portable Bluetooth readers have been developed and can be easily deployed as needed. The collected data can be either stored locally or transmitted to TMC in real time via 4G cellular network. Given that most new cars are equipped with Bluetooth, this data source is becoming increasingly important and reliable. However, there are several major limitations for Bluetooth data. First, the data sample is often biased. It is not uncommon to have several people (i.e., multiple Bluetooth devices) in one vehicle. This can lead to biased travel time measurements. Second, like all previously discussed sensors, the coverage of Bluetooth readers is still limited. A dense network of Bluetooth readers is needed, especially for quickly detecting incidents and accurately estimating their locations. Third, Bluetooth cannot provide information for individual lanes like what inductive loops, microwave, and cameras can do.

Another potential application of Bluetooth sensors is to derive OD and driver route choice data, which are very useful for TSMO. With such information, DOTs can better understand how drivers respond to congestion (e.g., messages displayed on changeable message signs) and make route choice decisions. This may help TMC operators develop effective traffic management and control strategies. However, deriving accurate OD and driver route choice data is not a trivial task, especially given some of the limitations of Bluetooth sensors.

MassDOT has a Bluetooth travel time system called GoTime and seems to be happy with its performance. NHDOT and VTrans have tested Bluetooth travel time systems but are not very satisfied with their performance.

1.1.1.5. Summary

Data from loop detectors often are not streamed to HOC in real time and are mainly used for HPMS reporting purpose. Such data are saved in roadside devices and are manually downloaded. Some New England DOTs (e.g., NHDOT) are thinking about connecting these loop detectors wirelessly to Traffic Management Centers (TMC) so that data can be downloaded in real time and remotely. Most DOTs are moving away from loop detectors due to the high installation and maintenance cost. Installing or repairing loop detectors requires setting up temporary work zones, which is expensive and creates safety risk.

Microwave sensors are widely used by New England DOTs. Data from microwave sensors (e.g., volume, speed, and occupancy by lane, vehicle length, vehicle type) are typically streamed to TMC in real time. Compared to cameras, microwave sensors are more robust and are not affected by lighting conditions. Although thermal cameras can address the lighting issue, they are more expensive and we are not aware of any New England DOTs that are using them.

Some retrofit loop detectors are able to generate inductive vehicle signatures, which can be used to re-identify vehicles at different locations. Similarly, Bluetooth/Wi-Fi sensors and CCTV camera-based vehicle re-identification techniques can match vehicles at different locations, thus generate vehicle OD information. The vehicle signatures generated by loop detectors usually are not very precise and are only good for matching vehicles at nearby locations, while MAC addresses from Bluetooth/Wi-Fi readers and license plate numbers (or vehicle video signatures) generated by CCTV cameras are more accurate. Research is still needed to derive accurate OD information from MAC addresses due to issues such as sampling rate and bias. For the license plate method, it requires a wide deployment of CCTV cameras. Its performance can be affected by lighting and camera angle factors.

1.1.2. Road Weather Information System (RWIS) and Winter Maintenance

State DOTs are using remote weather stations to monitor road surface conditions under different weather. Also, DOTs are interested in integrating data from weather forecast, weather stations, and sensors installed on vehicles (e.g., plow trucks). Some states have plow trucks equipped with AVL and sensors, which provide real-time locations and speeds of plow trucks, material types and application rates, pavement and air temperatures, engine diagnostics, dashcam images, surface friction, and humidity. Information from weather stations, probe vehicles, and weather stations are critical to TSMO under severe weather conditions.

CTDOT installed an Integrated Mobile Observations (IMO) system in maintenance vehicles. Approximately 210 plow trucks (with the eventual goal of all plow trucks) are equipped with forward looking video camera, GPS, and temperature and relative humidity sensors. The collected data is sent to a vendor for CTDOT (called DTN). The data is fed into a local weather forecasting system, and the processed data is then used to inform a Maintenance Decision Support System (MDSS), which recommends which and when roadway segments should be plowed and how much salt should be used in the winter to improve traffic safety.

The MDSS also takes data from a Roadway Weather Information System (RWIS) deployed at 40 fixed locations, which collects pavement friction, wet, dry, icy, salinity surface and sub-surface pavement temperature as well as atmospheric temperature information. A mobile version of this sensor suite is under development (MRWIS).

MassDOT has remote sensors to monitor roadway conditions, real-time locations and speeds of plow trucks, National Weather Service Data, smart work zone data (e.g., locations, durations, configurations). MassDOT’s RWIS is integrated with the DTN’s WeatherSentry platform, which allows MassDOT to view weather conditions across Massachusetts.

Besides NOAA data, NHDOT also collects many data elements (e.g., pavement temperature, visibility, precipitation, water, ice, friction factor) from about 30 weather stations that are mainly on interstate highways and the turnpike. The weather station data are used to guide snow plowing and salting activities. If the friction factor is below standard, a text message will be generated by weather station and sent to maintenance crews.

The plow trucks owned by NHDOT are equipped with the AVL system, which provides real-time information such as speed, location, plow up, plow down, spreading rate, etc. They are planning to add mobile RWIS to these plow trucks. NHDOT also has some maintenance trucks (not plow trucks) equipped with air and pavement temperature sensors. These trucks are driven by highway patrol foremen after snowstorms to determine how road segments should be treated. However, these maintenance trucks do not have the AVL system. They are thinking about integrating the data from weather stations, maintenance trucks, plow trucks, and weather forecast and presenting them in a simple but meaningful format for decision making. The data collected by weather stations and plow trucks' AVL systems can be found in Appendix A.

NHDOT current does not have a Maintenance Decision Support System (MDSS®). A consulting company took all the weather data from NHDOT and tried to develop a system to predict temperature at any points in the road network. The predictions sometimes were accurate, but the accuracy was not stable.

NHDOT, MaineDOT, and VTrans all divide their states into zones (6 for NHDOT, over 100 for Maine, and about 10 for VTrans). They use weather stations, trucks equipped with weather sensors, sensors at highway maintenance sheds, cameras, speed data, and NOAA data to estimate roadway conditions and publish the results on the regional 511 website.

1.1.3. Work Zone

Smart work zone technologies have been widely used by New England state DOTs. These smart work zones utilize sensors such as microwave, camera, Bluetooth to monitor traffic and collect data such as travel time, speed, delay, and queue. MaineDOT uses Linear Referencing System (LRS) to manage their work zone related information in ATMS, although the information is not updated in real time and requires data standardizations. FHWA established the Work Zone Data Exchange (WZDx) program a few years ago, and RIDOT is working on sharing smart work zone data using WZDx.

1.1.4. Weigh-in-Motion (WIM)

State DOTs also collect data from WIM stations, including traffic volumes by vehicle classification and weight, date, time, vehicle length by axle spacing, speed, and axle weight.

1.1.5. Tolling Data

Since toll roads are typically operated by private companies, DOT TSMO divisions often do not have full/direct access to tolling data. They have to request such data through their turnpike authorities. Two types of tolling data can be useful for TSMO purposes: E-ZPass and license plate records. E-ZPass data is similar to Bluetooth data. The main difference is that E-ZPass uses the Dedicated Short-Range Communications (DSRC) technology to read transponders in individual vehicles instead of MAC addresses. Since each vehicle has a unique transponder ID, the travel time derived from E-ZPass data is less biased than Bluetooth data. From E-ZPass records, time-dependent OD can be easily derived.

Some turnpikes also allow vehicles not equipped with E-ZPass to use by tracking their license plate numbers. By matching license plate numbers observed at entrances and exits, accurate OD and travel time information can be obtained. A clear limitation with the E-ZPass and license plate number data is that they are only available for toll roads.

1.1.6. Incident and Crash

All state DOTs have a database for historical incidents and crashes. Such data include incident/crash location, time, duration, etc. Some state DOTs also keep track of highway safety patrol records (e.g., MaineDOT) and 511 phone call records. NHDOT does not have a 511 system anymore but has access to highway safety patrol records and 911 calls related to traffic accidents. Their safety patrol records are in paper format and are entered into a database by TSMO staff. NHDOT uses the incident data information to optimize (based on human intelligence not automated algorithms) safety patrol schedules.

NHDOT's ATMS can take state police inputs to show crash alarms. It also allows traffic operators to define speed thresholds to display roadway segments in different colors based on their speeds. NHDOT finds this to be very useful for detecting incidents. The speed data used in the application come from both TomTom and Wavetronix sensors on the roads. One issue with TomTom (and other similar products) is that DOTs do not have control over the segment length and only average speed data for the entire segment are provided. If an incident happens on a long segment, its impact will take a long time to be reflected in the average segment speed. Therefore, NHDOT uses DOT owned sensors to complement TomTom data for incident detection on long segments.

In Connecticut, some highway patrol vehicles are equipped with data collection devices developed by a company called HAAS. CTDOT takes data from both highway patrol (e.g., HAAS system, 911 calls) and Waze for incident detection. They have access to the incident information entered into the Connecticut State Police (CSP) Computer-Aided Dispatch (CAD) system. CSP dispatchers also notify the CTDOT of incidents on state roadways via dedicated telephone lines. CSP is in the process of revising its CAD time logs to provide additional scene clearance information for improved analysis of incident clearance times. CTDOT found the CSP CAD records to be a timely and reliable data source for incident detection, since almost all drivers have a cell phone. Overall, CTDOT is satisfied with the performance of the CSP CAD and Waze reports for incident detection.

For incidents reported by Waze, CTDOT uses CCTV cameras to further verify them. The incoming Waze reports of the same incident are automatically aggregated so that one incident will not generate many alerts that overwhelm HOC operators. CTDOT can also update Waze incident data to remove false alarms and incidents that have already been cleared. This two-way communications between CTDOT and Waze help provide reliable information to travelers and improve incident response.

1.1.7. Arterial

Loop detectors, traffic cameras, microwave sensors are commonly used at arterial intersections for sensing and data collection. Some state DOTs (e.g., MassDOT) also experimented with drones to collect traffic condition data (e.g., queue length) at intersections. Several state DOTs (e.g., VTrans, MassDOT) have deployed Automated Traffic Signal Performance Measures

(ATSPM) systems, which provide real-time data on traffic detector state (e.g., occupied vs. unoccupied) and health condition, traffic control (e.g., which signal head is currently in green), turning movement counts, queue length, and speed. For intersections not equipped with ATSPM, the above data sometimes are also captured but are often not streamed to TMC. The ATSPM data can have many applications other than monitoring the health conditions of traffic detectors and controllers but have not been fully utilized yet. Since not every state has ATSPM, it is considered as an emerging data source and is further discussed in Section 1.2.3.

1.1.8. Transit

Transit agencies also collect many data that can be used for TSMO purposes, including General Transit Feed Specification (GTFS) data, transit fare collection data (e.g., smart card, Mobile ticket), CCTV camera videos, Automated Passenger Counter (APC) data, and ridership. For example, GTFS data can be used to estimate link travel time on urban arterials. However, the GTFS data is only widely available in major cities with many bus routes like Boston, not on state-maintained highways. Also, such data are owned by transit agencies, and are not directly accessible by TSMO division. For example, in Rhode Island, Rhode Island Public Transit Authority (RIPTA) is an agency separated from RIDOT. Sharing data across agencies is important, but can be difficult, especially for real-time data sharing. This partially explains why none of the six New England state DOTs explicitly utilize transit data for TSMO. For example, NHDOT TSMO does not utilize any transit data.

Delaware Transit Corporation (DTC) supplies fixed route and paratransit services statewide. DTC is an agency under DelDOT. DTC's automated fleet management system is integrated in DelDOT's AI-ITMS. Future AI-ITMS development could include transit system status information.

1.1.9. Parking

Static (e.g., location and # of lots) and dynamic parking data (e.g., parking duration), parking fee data, and Mobile parking app data can also be useful for traffic management and control. For example, CTPS has done license plate surveys at commuter rails stations to derive passenger OD information, which is important for multimodal corridor transportation management.

TSMO division often does not have direct access to parking data. One reason is that many of the parking facilities are owned and operated by private companies. NHDOT provides support to allow third-party Apps to show the availability of parking spaces. However, the data is not utilized for TSMO purpose at this moment. NHDOT is only concerned about parking in a very limited number of areas (e.g., White Mountains).

1.1.10. Assets

Asset condition data are handled by different DOT divisions mostly in GIS format, including pavement, bridges, speed limits, traffic signs and markings, tunnels, ITS equipment, etc. Other than the conditions of bridges and tunnels, state DOTs are also collecting the condition data for ITS assets. NHDOT maintains a detailed database for ITS equipment such as variable message signs, sensors, communication devices, and traffic controllers. More details can be found in Appendix B.

NHDOT is in the process of loading all their ITS assets into a comprehensive GIS database. Previously NHDOT only tracked the locations of ITS devices (e.g., cameras, variable message signs), not the detailed condition information for each asset, for example, the Cabinet for a camera has a modem and a server Rack. They are also working on a work order system to ensure the state of good repair for ITS assets. The ultimate goal is to collect detailed asset condition and configuration data (e.g., modem type, maker) and connect them to asset locations managed by GIS. Also, such a system will be integrated with the work order system (e.g., Assetworks), so that NHDOT can track when and where a device is replaced or repaired. Maintenance is important for ITS assets. To maintain a state of good repair, DOTs need to know how much is needed for the next five or ten years for ITS asset maintenance, which will benefit from a detailed and accurate asset inventory system.

Laser, LIDAR, and camera (mounted on vehicles and drones) have been extensively used by state DOTs to collect asset inventory and condition data. Such sensors have generated an enormous amount of data. More research is needed to reduce such datasets and explore how they may be used for TSMO purposes.

1.2. New and Emerging Data and Data Sources

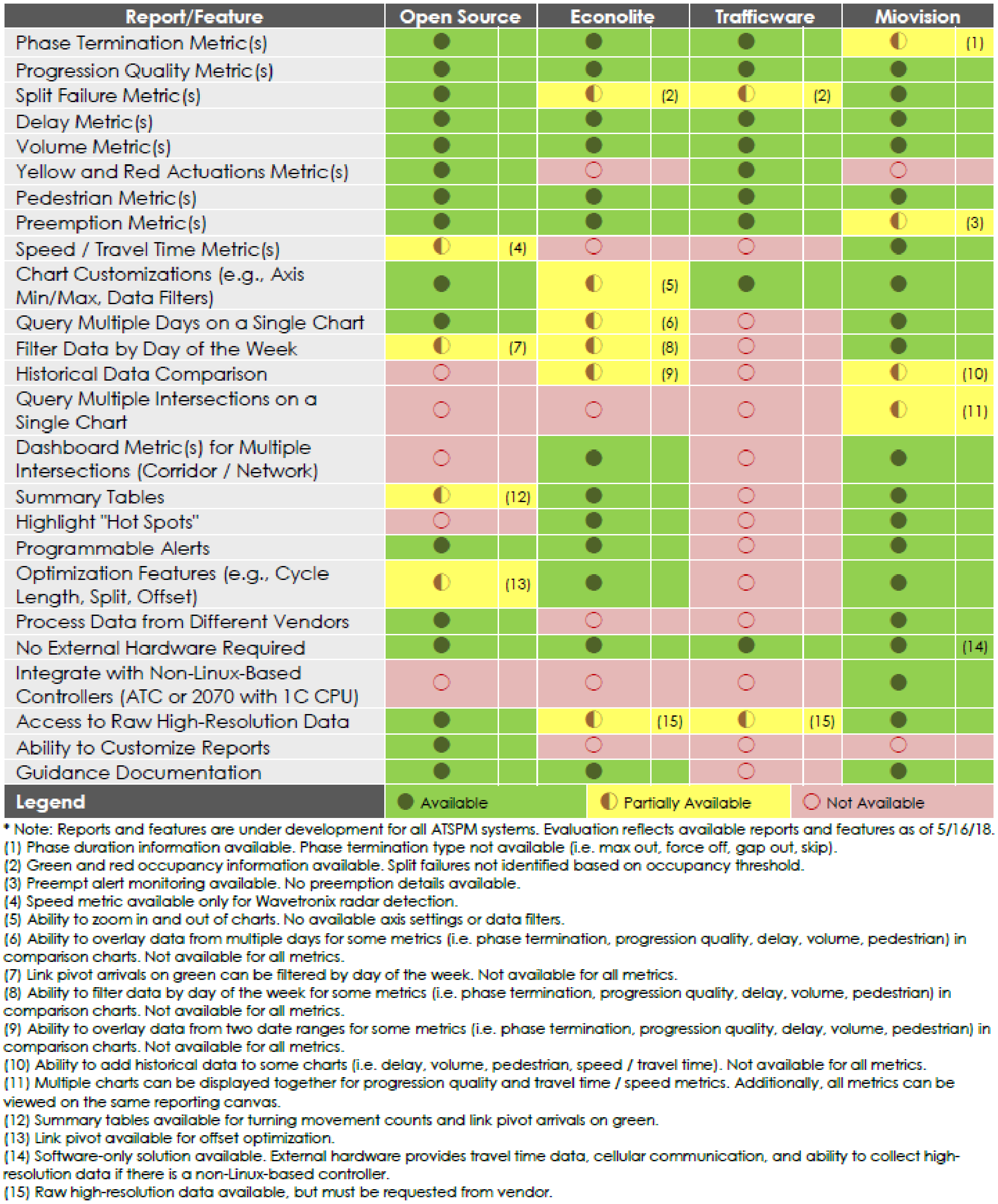
1.2.1. Drone

Drones or Unmanned Aerial Vehicles (UAV) have been used as a popular platform for collecting highway data. RGB and infrared cameras and LIDAR have been mounted on drones for many applications. Those related to TSMO include providing situational awareness at incident/crash scenes and traffic monitoring. Traffic monitoring via drones can overcome the limitations of traditional methods of monitoring due to its simplicity, mobility, and ability to cover large areas. A recent paper presents a good review of research efforts that use drones in relation to online and offline extraction of traffic parameters from video data [*7*]. MassDOT is working on establishing a drone-based emergency response network and has used drone to monitor queue length at signalized intersections. Drone has been extensively used by RIDOT for construction sites monitoring.

1.2.2. LIDAR

LIDAR has also attracted a tremendous amount of attention in the past decade. MassDOT used LIDAR to scan all the state-maintained highways, which led to several hundred Terabyte (TB) data. They extracted useful information such as traffic signs from the LIDAR data.

Drones and LIDAR have generated a huge amount of data. How to extract useful information from such datasets and share and store them now become a major issue. DOTs certainly do not want to discard such datasets and later find that valuable information could have been extracted from them. A good idea is to share such datasets (when possible) with universities, private companies, and the public, allowing them to come up with innovative ideas to analyze and utilize the data.



Report/Feature	Open Source		Econolite		Trafficware		Miovision	
Phase Termination Metric(s)	●		●		●		◐	(1)
Progression Quality Metric(s)	●		●		●		●	
Split Failure Metric(s)	●		◐	(2)	◐	(2)	●	
Delay Metric(s)	●		●		●		●	
Volume Metric(s)	●		●		●		●	
Yellow and Red Actuations Metric(s)	●		○		●		○	
Pedestrian Metric(s)	●		●		●		●	
Preemption Metric(s)	●		●		●		◐	(3)
Speed / Travel Time Metric(s)	◐	(4)	○		○		●	
Chart Customizations (e.g., Axis Min/Max, Data Filters)	●		◐	(5)	●		●	
Query Multiple Days on a Single Chart	●		◐	(6)	○		●	
Filter Data by Day of the Week	◐	(7)	◐	(8)	○		●	
Historical Data Comparison	○		◐	(9)	○		◐	(10)
Query Multiple Intersections on a Single Chart	○		○		○		◐	(11)
Dashboard Metric(s) for Multiple Intersections (Corridor / Network)	○		●		○		●	
Summary Tables	◐	(12)	●		○		●	
Highlight "Hot Spots"	○		●		○		●	
Programmable Alerts	●		●		○		●	
Optimization Features (e.g., Cycle Length, Split, Offset)	◐	(13)	●		○		●	
Process Data from Different Vendors	●		○		○		●	
No External Hardware Required	●		●		●		●	(14)
Integrate with Non-Linux-Based Controllers (ATC or 2070 with 1C CPU)	○		○		○		●	
Access to Raw High-Resolution Data	●		◐	(15)	◐	(15)	●	
Ability to Customize Reports	●		○		○		○	
Guidance Documentation	●		●		○		●	
Legend	● Available		◐ Partially Available		○ Not Available			

* Note: Reports and features are under development for all ATSPM systems. Evaluation reflects available reports and features as of 5/16/18.
(1) Phase duration information available. Phase termination type not available (i.e. max out, force off, gap out, skip).
(2) Green and red occupancy information available. Split failures not identified based on occupancy threshold.
(3) Preempt alert monitoring available. No preemption details available.
(4) Speed metric available only for Wavetronix radar detection.
(5) Ability to zoom in and out of charts. No available axis settings or data filters.
(6) Ability to overlay data from multiple days for some metrics (i.e. phase termination, progression quality, delay, volume, pedestrian) in comparison charts. Not available for all metrics.
(7) Link pivot arrivals on green can be filtered by day of the week. Not available for all metrics.
(8) Ability to filter data by day of the week for some metrics (i.e. phase termination, progression quality, delay, volume, pedestrian) in comparison charts. Not available for all metrics.
(9) Ability to overlay data from two date ranges for some metrics (i.e. phase termination, progression quality, delay, volume, pedestrian) in comparison charts. Not available for all metrics.
(10) Ability to add historical data to some charts (i.e. delay, volume, pedestrian, speed / travel time). Not available for all metrics.
(11) Multiple charts can be displayed together for progression quality and travel time / speed metrics. Additionally, all metrics can be viewed on the same reporting canvas.
(12) Summary tables available for turning movement counts and link pivot arrivals on green.
(13) Link pivot available for offset optimization.
(14) Software-only solution available. External hardware provides travel time data, cellular communication, and ability to collect high-resolution data if there is a non-Linux-based controller.
(15) Raw high-resolution data available, but must be requested from vendor.

Figure 1 TSMO performance measure needs provided by RIDOT [*2*]

1.2.3. ATSPM

Utah was among the first few states to invest in the Automated Traffic Signal Performance Measures (ATSPM) system starting in 2013 [*1*]. Several New England state DOTs now have also deployed the ATSPM, which collects very detailed traffic signal performance measures every 1/10 seconds and stream the data in real time [*2*] to TMC. The data elements collected by well-known ATSPM systems are listed in Figure 1.

NHDOT currently has very few intersections that are connected via fiber to TMC. Over the next few years, many intersections will be connected to TMC either by fiber or wireless network. The ATSPM data can be used for improving traffic safety and operations at intersections. Since this system is relatively new, researchers and practitioners are still trying to find out how to effectively utilize the generated data. A potential challenge for example is that traffic signals are not directly under the TSMO bureau at NHDOT. Traffic signals, pavement marking, and signs are under the bureau of traffic at NHDOT. The NHDOT TSMO bureau does not have traffic signal engineers or technicians on staff. This organizational structure is typical for other DOTs. Overall, how to make the full use of ATSPM data seems to be an interesting and timely topic for DOTs.

1.2.4. Crowdsourced, Probe Vehicle, and Connected Vehicle Data

Smartphones, probe vehicles, and connected vehicles all rely on GPS and have generated a tremendous amount of data that can be used for TSMO purposes. Since these data sources are based on similar technologies, they are thus discussed under the same main category. However, there are some subtle but important differences among these data sources.

Smartphones users contribute their data either actively or passively. One example of active smartphone data contribution is the Waze App. Waze users report roadway conditions such as incidents, debris, and speed traps. Both incidents and debris are critical safety hazards and need to be cleared as soon as possible. Such crowdsourced data are important for DOTs to improve highway safety. Smartphone users in many cases also passively contribute their data. For example, when drivers are using navigation Apps, they often contribute their speed and location information every few seconds. The speed and location information from all drivers collectively can be used to predict travel time, estimate travel speed, detect incidents, and identify safety hazards (e.g., locations with frequent harsh brakes). These Apps share some of the derived data with data contributors, but not all.

Probe vehicles refer to vehicles equipped with GPS and wireless communications technology. Sometimes this is called Automated Vehicle Location (AVL). Many buses and commercial trucks (e.g., owned by UPS and Walmart) are equipped with AVL. With AVL and other onboard sensors, system operators can know in real time where drivers are, how many times a heavy truck backs up, whether turn signals are used when they should, speed violations, etc. Drivers of Transportation Network Companies such as Uber and Lyft need to install an App to connect with customers/passengers. These Uber and Lyft vehicles essentially work as probe vehicles. Companies such as INRIX and TomTom then take probe vehicle data from different sources (The exact sources are not disclosed and they may not include Uber and Lyft), clean them, and sell them to customers such as state DOTs. The INRIX and TomTom data are aggregated. One can only know the average segment speed or travel time, not individual vehicle speeds and locations. Given the original probe vehicle data, theoretically it is possible for INRIX and TomTom to provide disaggregated data to customers. However, this requires customers to have the capability of analyzing very detailed and large datasets.

USDOT has three ongoing connected vehicles pilot studies in Wyoming, New York City, and Tampa. In these studies, thousands of connected vehicles are generating very detailed vehicle trajectory data. However, such data only cover the three pilot sites. On the other hand, many new vehicles on the market are now equipped with GPS and wireless communications capability. Car manufacturers collect vehicle location and speed information, engine and wiper status, etc. and sell them to companies such as Wejo and Otonomo. Both probe vehicle and connected vehicle datasets are based on GPS and wireless communications. The vehicle location and speed data are usually transmitted from vehicles to data center every few seconds. The processed data are then shared with customers in about one minute, which is sufficient for many TSMO applications such as incident detection. Different from INRIX and TomTom, Wejo and Otonomo provide disaggregated data to customers. Although such detailed data can be very useful, analyzing them is difficult. So far, none of the 6 New England state DOTs have used either Wejo or Otonomo data.

In the rest of this section, the above data sources are discussed in more details. In this report, these data sources are grouped into the following three sub-categories:

- ***User Reported Data***: This specifically refers to data contributed actively by travelers using cellphones. Waze is a main source of such data. Some navigation Apps such as Google Maps also allow users to report incidents and speed traps.
- ***Aggregated Probe Data***: This sub-category includes aggregated speed and travel time data such as those provided by INRIX, TomTom, etc. Also, Uber Movement provides zone to zone travel time and road segment travel speed data. Google sells travel time data. All these aggregated datasets are based on GPS coordinates generated by smartphone Apps, AVL, or connected vehicles.
- ***Trajectory Data***: This sub-category is for unprocessed GPS coordinates generated by smartphone Apps, AVL, or connected vehicles. Examples include Wejo and Otonomo.

1.2.4.1. User Reported Data

Almost every driver now has a cellphone. When a crash occurs, it usually will not take much time for the driver(s) involved or passing by drivers to call 911 and report it. Some state DOTs rely a lot on such information for AID. A limitation of driver incident reporting is that non-collision (e.g., road debris) and property-damage-only (PDO) incidents may be under-reported. Also, for drivers calling 911, they sometimes do not know their exact locations on the road (unless they can obtain the location information using a mobile App such as Google Maps).

Most state DOTs have access to Waze data and are using Waze incident reports for AID. Issues with Waze incident reports include: (1) submitting Waze incident report while driving is very dangerous; (2) usually there is a delay between when an incident is spotted and when it is reported, which makes it difficult to identify the exact incident location; and (3) sometimes there are incorrect reports. For example, an incident has already been cleared, but it still shows up in Waze. In a quantitative comparison by Iowa DOT (IDOT) of various sources of incident detection, Waze was ranked the 4th (out of 8) largest contributing sources. While essentially free, Waze incident reports still must be validated by other means, and it captured only 43.2% of

ATMS recorded incidents during the analysis period (although this has most likely increased as the number of users increases).

1.2.4.2. Aggregated Probe Data

Several companies offer aggregate probe data, including INRIX, HERE, TomTom, and Google. A significant advantage of probe data is that state DOTs do not need to invest in any data collection infrastructure, and do not need to worry about the maintenance of such infrastructure either. Although purchasing data can be expensive, clearly many DOTs think it is worthy given the trouble and cost associated with maintaining their own data collection infrastructure. In addition, probe data usually has a much larger coverage than traditional data sources such as inductive loop detectors, microwave detectors, and CCTV cameras. Its actual coverage depends on how many users are contributing their data. Usually, there are more data contributors in urban than suburban areas.

Most probe data vendors provide information aggregated by road segments, such as segment speed and travel time. State DOTs take what these vendors provide and do not know the details of how the data are aggregated. The length of each segment is also decided by vendors. Different vendors often have different standards/ways to divide roads into segments. When DOTs have data from multiple vendors, they often face the challenge to reconcile data aggregated using different segment definitions, which is not a trivial task. In addition, state DOTs lose the opportunity to extract more granular and useful information from the aggregated probe data.

Using Automated Incident Detection (AID) as one example, DOTs may want to have short segments in areas prone to incidents (ideally in all areas if computational power is not a constraint). With short segments, changes in individual vehicles' speeds and travel times can be quickly reflected in the corresponding segment measures. On the other hand, providing aggregated data and hiding the details to some extent is beneficial to DOTs, as they often do not have the human resources to handle the large volume of raw trajectory data and extract critical information out of them.

As discussed previously, the aggregated probe data originally come from detailed vehicle trajectories. Besides the aggregated road segment measures, some data vendors (e.g., INRIX) provide more detailed data at the lane level. They can also generate incident and dangerous slowdown alerts.

Overall, state DOTs are satisfied with probe data quality. CTDOT has validated HERE travel time data. DOT employees had driven some routes to verify the travel time estimated by HERE and found that they meet the DOT data quality standards (FHWA 23CFR511 quality standard for traffic data). CTDOT noted they only display the probe data when it is accurate, which at this point is mostly during daylight hours and some early evening hours on weekdays and weekends. NHDOT was able to detect crashes based on TomTom speed data even before they were notified by state police.

The City of Boston partnered with Waze to identify traffic signals that need improvements. They worked with MBTA to measure impacts of signal timing along the Silver Line [*3*]. Although it was not explicitly mentioned what data from Waze was used, it is very likely the travel speed data (similar to the probe vehicle data) and the more detailed vehicle trajectories (see discussion in the next subsection) were used.

1.2.4.3. Trajectory Data

Smartphones, probe vehicles, and connected vehicles can also generate vehicle trajectories, which are much more detailed than the aggregated probe data described in the previous subsection. Two major trajectory data vendors are Wejo and Otonomo. They provide similar vehicle trajectory datasets, which include data elements such as longitude, latitude, speed, heading, wipers change, seat belt change, autonomous emergency braking, etc. These data elements are collected from some commercial vehicles sold in recent years at short intervals (e.g., every 3 seconds) and are transmitted back to a data warehouse and made available to end users ***within 60 seconds***. Currently, there are over 10 million vehicles contributing trajectory data, and this number is growing as more new vehicles are being sold.

Such vehicle trajectory data can be used to measure traffic operations performance and derive surrogate safety measures such as harsh-braking events. For example, researchers from Purdue University used Wejo data to correlate harsh-braking events with crash occurrences near highway work zones [*4*]. Vehicle trajectories are also very useful for some real-time applications, such as detecting traffic incidents and traffic slowdowns and generating safety hazard alerts (e.g., black ice on road).

The Eastern Transportation Coalition conducted a pilot study to estimate traffic volume using Wejo data in real time [*5*]. Their study utilized data from six states: Alabama, Florida, Georgia, North Carolina, Tennessee, and Virginia. They found that Wejo data covered about 3% of all vehicles on the road and the pilot study received data from each connected vehicle every 3 seconds. Wejo generated over 230B data points in the 3-month pilot study period. Their study concluded that using Wejo data to estimate traffic volume in real time is a viable solution, particularly given that the number of connected vehicles is continuously growing.

Many DOT vehicles are equipped with the AVL system, allowing DOTs to track their vehicles (e.g., plow trucks) in real time. These vehicles can provide valuable trajectory information especially under severe weather conditions. Other public agencies also have AVL in their fleets, such as state police and transit. Integrating all fleet data can generate very useful traffic information benefiting all participating agencies (e.g., first responders always want to have accurate traffic information to find the best routes). In Delaware, all state vehicles have GPS based tracking. As part of their ATCMTD AI-ITMS project, DelDOT is equipping some DOT vehicles to monitor vehicle data port and to integrate the data into their AI-ITMS.

1.2.4.4. Summary

Crowdsourced, probe vehicle, and connected vehicle data are playing a critical role in TSMO. For example, RIDOT is exploring INRIX data for AID. Currently, RIDOT relies on CCTV cameras and reviews the footages manually to detect incidents. RIDOT also has access to the radio of state police, which is another important source (i.e., user reported data) for incident information. Although user reported data via Waze are a little noisy, they are very useful to DOTs due to its coverage.

Overall, state DOTS are satisfied with probe data given that they are maintenance free and cover a very large area. Issues with probe data include low sampling rate in rural areas, reliability (sampling rate for the same segment changes over time), and data conflation.

It is estimated that by 2023 90% of new vehicles in the United States will be shipped with embedded connectivity. The near real-time connected vehicle trajectory data provide a source with many great possibilities for improving the understanding of traffic flows and developing advanced traffic management strategies. State DOTs should carefully monitor the development of connected vehicles and its impacts on traffic data collection and TSMO.

1.2.5. Other Mobile Device Location Data

Since almost every driver now has a cellphone, being able to accurately locate cellphones can help estimate vehicle speeds. Cellphones can be used to generate user reported data (Section 1.2.4.1) and aggregate probe data (Section 1.2.4.2). Besides cellphones, there are many other mobile devices such as tablets and smart watches. It is mentioned in a 2019 news article [*6*] that "Every minute of every day, everywhere on the planet, dozens of companies — largely unregulated, little scrutinized — are logging the movements of tens of millions of people with mobile phones and storing the information in gigantic data files." It is estimated that there is a $12 billion market [*7*] for such data. There is a long list of companies that use mobile device location data for various applications, including AirSage, SkyHook, Cuebiq, and SafeGraph.

There are mainly three types of mobile device location data: cell tower triangulation data, mobile device GPS location data, and mobile carrier data. These data sources are further detailed below.

1.2.5.1. Cell Tower Triangulation

Each cellphone has to be connected to at least one cell tower. The distance between the cellphone and cell tower can be estimated by measuring the strength of wireless signals transmitted between them. Since the cell tower location is fixed and known, the location of the phone can be narrowed down to a circle. If the phone is communicating with two cell towers, its location can be further narrowed down to two points. With three cell towers, theoretically the phone location can be uniquely determined. However, the accuracy of distance estimation based on wireless signal strength is not perfect. Even with three cell towers, a cellphone can usually be located within an area of ¾ square miles.

Mobile device location data obtained via cell tower triangulation usually is not very precise, and cannot be used for calculating speed, travel time, etc. However, it can be used to estimate time-dependent OD data. OD data is important for understanding travel demand. It can be used together with traffic simulation tools to answer questions such as what may happen if a road segment is shut down due to major accidents or construction.

1.2.5.2. GPS Location or Location Based Service (LBS) Data

Smartphone are all equipped with GPS, which provides much more accurate location information than cell tower triangulation. Most mobile device location applications are based on

GPS location data. GPS location data (e.g., obtained via navigation and other LBS Apps) can be used to derive travel time and speed (see Section 1.2.4.2). It also has many other important applications. For example, GPS location data can be used to derive trip generation rates and help business find optimal retail locations. Similar to cell tower triangulation data, GPS location data can be used to derive OD data. In addition, it can potentially be used to drive trip chain, travel mode, and route choice information, which is important for both transportation planning and TSMO.

A major issue with GPS location data is latency. Unlike the connected vehicle trajectory data in Section 1.2.4.3, GPS location data in many cases are not immediately available to end users. An exception is the GPS location data obtained via navigation Apps, which is aggregated and made available to App users in real time. It would be ideal if Google and Apple can share their real-time navigation App data (e.g., trajectory, travel time, incidents) with state DOTs to improve TSMO (e.g., incident detection). Even historical GPS location data can be very useful for DOTs. They can be used to identify and prioritize bottlenecks, safety hazards, etc.

1.2.5.3. Mobile Carrier Data

Through either cell tower triangulation or smartphone GPS, mobile carriers can have the location information of their subscribers. This data source is likely to have a much higher sampling rate of users than other sources such as probe vehicles and generates more accurate measurements of traffic speed and travel time.

The wireless communication solution for future connected vehicles is not clear at this moment. It could be based on DSRC, 5G, or 6G. If 5G or 6G is used as the backbone for connected vehicles, mobile carriers will play a critical role and will have access to a huge amount of vehicle related data, including the trajectory data discussed in Section 1.2.4.3.

Some mobile carriers have also shown great interest in ITS and smart cities. Verizon partnered with some cities (e.g., Boston and Kansas City) to install sensors in the pavement and connect cameras to traffic lights for detecting traffic [*8*] and improving traffic signal operations. AT&T also has an "AT&T Smart Cities Structure Monitoring" program, which adds AT&T LTE-enabled sensors to the existing lighting infrastructure in some U.S. cities (e.g., Atlanta, Dallas) to monitor traffic and parking, and detect gunshots [*9*].

1.2.5.4. StreetLight Data

StreetLight is essentially based on mobile device GPS location data (Section 1.2.5.2). It applies AI algorithms to integrate mobility device location data provided by Cuebiq [*10*], data from DOT permanent traffic counting stations, etc. to estimate AADTs, bike and pedestrian volumes, OD, and so on. It is listed here in a separate subsection because of its popularity. MaineDOT and MassDOT both have StreetLight data. NHDOT is also thinking about purchasing StreetLight dataset, partially because it can provide OD information.

1.2.6. Social Media

Some researchers proposed to use data from social media such as Twitter for TSMO purposes such as AID. They use the Natural Language Processing (NLP) method to extract traffic incident related information from social media feeds. For instance, after identifying an incident-related tweet, words related to “when”, “where”, and “how bad the incident is” will be extracted and analyzed if they exist. A major issue with this data source is that incidents are not guaranteed to be posted in a timely manner and with sufficient details to determine its nature and location information that is accurate enough.

Most state DOTs (e.g., NHDOT and MassDOT) use Twitter and Facebook to share traffic information with the public, not to take information (e.g., major crashes, traffic disruptions due to snowstorms) from those social media platforms.

1.2.7. Autonomous Vehicles

Each autonomous vehicle is equipped with a suite of sensors, which generate a huge amount of data each day. It is not a secret that Tesla collects data from its vehicle owners [*11*] to improve their self-driving algorithms. Even human-driven vehicles are now collecting and sharing data (e.g., Wejo and Otonomo). Some autonomous vehicle companies such as Lyft and Waymo have made part of their collected data (e.g., LIDAR, vehicle trajectory, camera) available to the public.

Autonomous vehicles can sense the surrounding traffic and generate more detailed information than vehicle trajectories. They can detect damaged traffic signs and guardrails, potholes, distracted pedestrians, aggressive drivers, debris on road, etc. However, they are not obligated to share anything with state DOTs. An interesting question is whether it is ethical for car manufacturers to collect data from drivers but do not share them with public agencies (e.g., state DOTs) for the benefits of drivers. The same question can be brought up to tech companies that collect mobile device location information.

1.2.8. Artificial Intelligence (AI)

AI technologies are well known for being data hungry. They often require a tremendous amount of data for model training and validation. On the other hand, AI is also an important tool for generating data. With cameras, drones, LIDAR, etc., transportation agencies have accumulated enormous image, video, and point cloud data that they do not know how to effectively utilize. Well-trained AI models can be used to turn such data into useful information. For example, traffic counts and assets can be derived from videos and LIDAR point cloud, respectively. AI algorithms are also widely used in autonomous driving to process camera, LIDAR, and microwave sensor data.

1.2.9. Summary

The data sources discussed are summarized in Table 1 below. From the discussion above, it can be seen that the landscape of traffic data collection has changed substantially in the past two decades given the advancements in sensors, wireless communications, Internet of Things (IoT) and smart cities, GPS and mobile devices, connected vehicles, and automated driving. Among them, mobile devices and GPS have probably the most profound impacts on traffic data collection. They significantly expand the coverage of traditional sensors (e.g., loop detectors, cameras) and provide a maintenance free approach for transportation agencies to collect detailed data elements such as vehicle trajectory, wipers change, seat belt change, and autonomous emergency braking. Another important front is the wide applications of AI technologies in sensor data processing, generating valuable traffic measurements for data-driven decision making.

Table 1 Traditional and Emerging Data and Data Sources for TSMO

	Traditional Data & Data Source	**New and Emerging Data and Data Source**
Highway	Loop detectors, microwave detectors, traffic cameras, Bluetooth/Wi-Fi MAC address readers, weather stations, weigh-in-motion stations Occupancy, delay and travel time, spot and segment speeds, volume, vehicular OD	Drone, Mobile Lidar Crowdsourced Data (e.g., Waze) Fleet data (DOT vehicles, commercial vehicles) Transportation Network and Logistics Companies (e.g., Uber Movement) Connected Vehicle Pilot Deployment Program TomTom, HERE, WeJo, StreetLight, INRIX, AirSage, SkyHook, Cuebiq, SafeGraph, Google, Apple Mobile Carrier (e.g., AT&T, Verizon) Cell tower triangulation, Cell Phone (or vehicle) GPS Social Media (e.g., Twitter, Facebook) Data from Autonomous Vehicles (e.g., Lyft, Waymo)
Incidents and Crashes	Incident/crash records (e.g., location, time, duration), highway safety patrol records, 511 phone records	
Arterial	Traffic signals, vehicle detectors, cameras, data from Automated Traffic Signal Performance Measures (ATSPM) system, queue length from drone.	
Transit	GTFS, transit fare collection data (e.g., smart card, Mobile ticket), traffic cameras, APC data, ridership, etc.	
Parking	Static (e.g., location and # of lots) and dynamic data (e.g., parking duration), parking fee data, Mobile parking app data	
Assets	*Highway*: conditions of traffic sign, pavement, marking, guardrail, bridges, tunnels, etc. *ITS*: conditions of variable message signs, sensors, communication devices, traffic controllers, etc. GIS maps (e.g., highway geometry), speed limits	
Maintenance & Work Zone	*Maintenance*: real-time locations and speeds of plow trucks, National Weather Service Data *Work Zone*: smart work zone data, location, duration, configuration, etc.	

2. POTENTIAL APPLICATIONS

In the Sources Sought Notice 693JJ3-21-SS-0013 released by the FHWA [*12*] in July 2021, the following applications of AI in TSMO have been identified:

- Predict/detect traffic incidents efficiently and proactively using AI and multisource/multisensor data and generate response plans.
- Predict multimodal delays in real-time using AI.
- Model urban network traffic as completely as possible using AI
- Optimize signal timing plans offline to service all modes of transportation by predicting vehicle and pedestrian arrivals, queues, and delays
- Optimize traffic signals in real-time using AI
- Enhance ramp metering strategy to rapidly adapt to anticipated or predicted conditions
- Use AI techniques to validate and verify datasets
- Use AI techniques to detect work zone location, schematic, and hazards; alert construction crews; and disseminate traveler information
- Detect and predict queues and shockwaves to harmonize speeds for reducing work zone crashes and delays
- Predict road surface conditions before they become dangerous and respond accordingly
- Proactively identify target speeds, lane assignments, ramp metering rates, etc. for improved traffic flow and throughput
- Collaborate across agencies in real-time using DSS and Knowledge Based Expert Systems (KBES)
- Improve situational awareness by fusing data from multiple sources and multiple sensors across the region
- Testing advanced traffic management and connected & automated vehicle technologies

Based on the above ideas, the team further identified the following more specific potential applications of AI for the NETC project panel to consider. We can pick one of these topics for the Phase II pilot study of this project.

- Use AI for data modeling and data-driven decision making
    - Some DOTs (e.g., VTrans, MassDOT) have deployed Automated Traffic Signal Performance Measures (ATSPM) systems. How to make the full use of ATSPM data seems to be an interesting and timely topic for DOTs.
    - DOTs are interested in integrating data from NOAA, weather stations, and sensors installed on vehicles (e.g., plow trucks) for Winter Maintenance Decision Support System. One possible application is to determine the optimal amount of salt to be applied.
    - Queue/slow moving traffic detection using connected vehicle data (e.g., Wejo or Otonomo)
    - Use incident data to optimize safety patrol schedules (currently based on human intelligence not algorithms)
    - Use location-based data to estimate traffic volumes, especially on low-volume roads and intersections.

- Use AI for data processing and reducing
    - Automatically detect incidents and collect traffic data using thermal traffic cameras
    - How to conflate and integrate data from different sources?
    - Use AI to process LIDAR data and automatically extract asset information such as type and location.

3. INTERVIEW QUESTIONS

In addition to reviewing existing and emerging data sources, the team interviewed staff from the six New England state DOTs with a list of questions covering their data needs, data analysis, archiving, sharing, security, and privacy protection practices, etc. The team also interviewed staff from Texas, Oregon, Virginia, and Delaware DOTs and the Eastern Transportation Coalition (previously known as the I-95 Coalition). These questions and the interview results are summarized in the rest of this section.

3.1. Data and Needs

3.1.1. Data Sources

Question: Are we missing any major data elements/sources in Table 1?

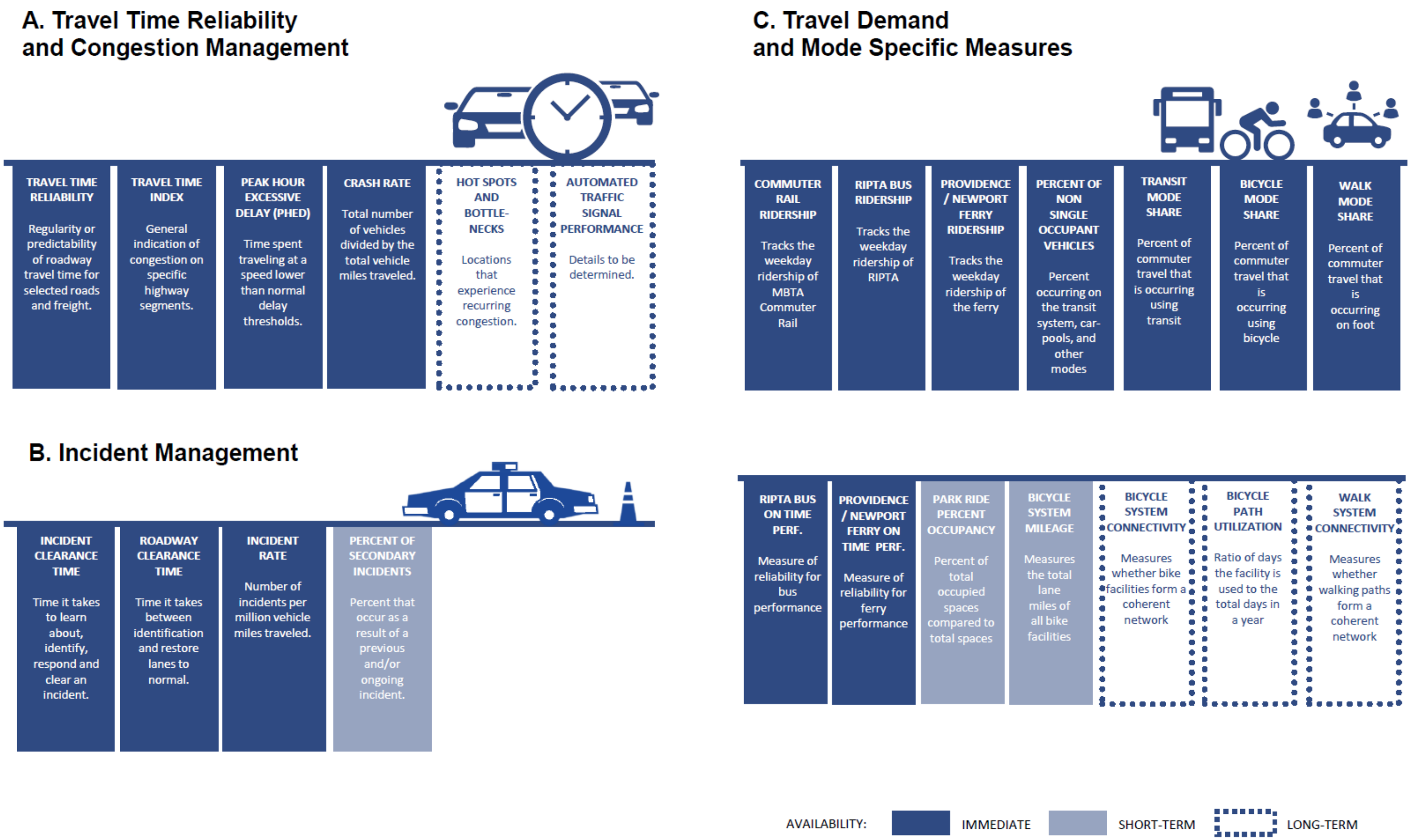
Most DOTs think the data elements in Table 1 are very comprehensive. One comment is that Table 1 can be better organized and made more detailed. For example, to provide detailed information about specific data products, data locations, and temporal limitations (e.g., real time data, archived data).

3.1.2. Data Needs

Question: Any existing and future data needs for TSMO (e.g., estimating OD in addition to segment AADTs)?

The identified future data needs for TSMO include:

- Integration of data from different sensors (e.g., loop detectors, AVL, pavement sensors), at different rates, and in different databases
- Connected vehicle data (e.g., detailed vehicle trajectories)
- Data related to public and private truck parking spaces and availability on major highway corridors
- Better data sharing with travelers, such as broadcasting traffic signal timing information to drivers
- How to determine incident duration, clearance time, and secondary incidents and separate incidents from recurring congestion. A reliable data source for identifying secondary crashes is currently unavailable, and such secondary crashes are sometimes under-reported in police reports (e.g., noted as primary incidents). Recurring congestion makes it challenging in determining when an incident is cleared.



A. Travel Time Reliability and Congestion Management

- **TRAVEL TIME RELIABILITY** Regularity or predictability of roadway travel time for selected roads and freight.
- **TRAVEL TIME INDEX** General indication of congestion on specific highway segments.
- **PEAK HOUR EXCESSIVE DELAY (PHED)** Time spent traveling at a speed lower than normal delay thresholds.
- **CRASH RATE** Total number of vehicles divided by the total vehicle miles traveled.
- **HOT SPOTS AND BOTTLE-NECKS** Locations that experience recurring congestion.
- **AUTOMATED TRAFFIC SIGNAL PERFORMANCE** Details to be determined.

B. Incident Management

- **INCIDENT CLEARANCE TIME** Time it takes to learn about, identify, respond and clear an incident.
- **ROADWAY CLEARANCE TIME** Time it takes between identification and restore lanes to normal.
- **INCIDENT RATE** Number of incidents per million vehicle miles traveled.
- **PERCENT OF SECONDARY INCIDENTS** Percent that occur as a result of a previous and/or ongoing incident.

C. Travel Demand and Mode Specific Measures

- **COMMUTER RAIL RIDERSHIP** Tracks the weekday ridership of MBTA Commuter Rail
- **RIPTA BUS RIDERSHIP** Tracks the weekday ridership of RIPTA
- **PROVIDENCE / NEWPORT FERRY RIDERSHIP** Tracks the weekday ridership of the ferry
- **PERCENT OF NON SINGLE OCCUPANT VEHICLES** Percent occurring on the transit system, car-pools, and other modes
- **TRANSIT MODE SHARE** Percent of commuter travel that is occurring using transit
- **BICYCLE MODE SHARE** Percent of commuter travel that is occurring using bicycle
- **WALK MODE SHARE** Percent of commuter travel that is occurring on foot
- **RIPTA BUS ON TIME PERF.** Measure of reliability for bus performance
- **PROVIDENCE / NEWPORT FERRY ON TIME PERF.** Measure of reliability for ferry performance
- **PARK RIDE PERCENT OCCUPANCY** Percent of total occupied spaces compared to total spaces
- **BICYCLE SYSTEM MILEAGE** Measures the total lane miles of all bike facilities
- **BICYCLE SYSTEM CONNECTIVITY** Measures whether bike facilities form a coherent network
- **BICYCLE PATH UTILIZATION** Ratio of days the facility is used to the total days in a year
- **WALK SYSTEM CONNECTIVITY** Measures whether walking paths form a coherent network

AVAILABILITY: ■ IMMEDIATE ▨ SHORT-TERM ⬚ LONG-TERM

Figure 2 TSMO performance measure needs provided by RIDOT

- Reliable queue length (or unexpected stops/slow-moving traffic on highways)
- Estimate highway and arterial traffic volume/density/capacity from different locations in real time. Existing probe data only covers speed and travel time.
- OD data will substantially increase DOT's ability to predict transportation system use and system response to demand changes. Reliable OD data, together with digital twins, AI, and simulation could substantially improve TSMO by quickly identifying a change, understanding the impacts of the change, evaluating and recommending options, implementing options, and continuously monitoring and adapting to the impacts of the changes.
- DOTs need more detailed and real-time condition information about ITS assets.
- CAVs will generate a lot of data that can be used for TSMO applications. On the other hand, CAVs will need more precise data for making safe, efficient, and eco-friendly driving decisions. In the future, variable message signs probably will not be needed. Instead, DOTs need to provide traveler information in digital formats that can be unambiguously interpreted by CAVs.

Additionally, RIDOT provides a very detailed list of performance measures needed for TSMO, which is presented in Figure 1. Although this list is not specifically for data needs, the listed items will depend on data and will guide researchers and engineers to find appropriate data sources to support the development of such performance measures. One example is the percent of secondary incidents. Such data does not exist in current police reports and is very difficult to obtain. However, it may potentially be obtained from new and emerging data sources such as the connected vehicle data.

In addition to data needs, this study also identifies some needs for data analysis methods, including:

- Data conflation is a major issue faced by many DOTs. For probe data such as INRIX, TomTom, and HERE, data vendors only provide aggregated information, such as segment speed and travel time. State DOTs take what these vendors provide and do not know the details of how the data are aggregated. The length of each segment is also decided by the vendors. Different vendors often have different standards/ways to divide roads into segments. When DOTs have data from multiple vendors, they will face the challenge to reconcile data aggregated using different segment definitions (e.g., LRS used by DOTs to manage pavement conditions, bridges), which is not a trivial task. In addition, state DOTs lose the opportunity to extract more granular and useful information from the aggregated probe vehicle and GPS data. Using automated incident detection (AID) as one example, DOTs may want to have short segments in areas prone to incidents (ideally in all areas if computational power is not a constraint). With short segments, changes in individual vehicles' speeds and travel times can be quickly reflected in the corresponding segment measures. On the other hand, providing aggregated data and hiding the details to some extent is beneficial to DOTs, as they often do not have the human resources needed to handle the large volume of raw trajectory data and extract critical information out of them.
- Data aggregation and mining is important. DOTs need a systematic way of integrating data from different sources (e.g., loop detectors, CCTV cameras, Waze, HERE, INRIX)

and generate useful data for performance measurement, incident detection, traveler information system, etc. Note that these data come in at varying rates, latencies, and accuracies. For example, Waze incident report data can be quite noisy with multiple reports for one incident and the reported locations of the same incident may not match well. Another example is loop detector data can be used to complement and verify INRIX data. However, unlike INRIX data, loop detector data are often not streamed to the HOC in real time and are only for fixed locations instead of segments.

- Innovative data analysis methods and approaches are needed. The existing data analysis methods can be adapted and applied in different ways, depending on what data is available. Using AID as one example, traditional AID methods are based on loop detector data and focus on identifying critical patterns/thresholds in terms of spot speed, volume, and occupancy measured at up- and downstream locations. With INRIX and HERE data, the same statistical or machine learning methods can be used but applied in a different way. The focus now is to compare adjacent segments' current and previous speeds by taking segment length into consideration. If Wejo data (i.e., raw vehicle trajectories in real time) are available, the AID problem will become how to identify changing points in time series (i.e., vehicle trajectories). Therefore, the evolving data sources will call for new methods and innovative applications of existing methods.
- Data sharing and brainstorming: For example, LIDAR data can be used to derive accurate ramp geometry information, which can be combined with drone captured vehicle trajectories to identify safe entrance speed for speed advisory applications. The treasures hidden in various datasets require creative thinking and analysts with both data science background and transportation engineering domain knowledge.
- With real-time data at more granular levels, we need to know how to process and store the data, and how to make sure that we do not overwhelm our communication and computing systems.

3.1.3. Data Quality

Question: Are you satisfied with the existing data quality and reliability? For example, many state DOTs have purchased Waze, INRIX, and StreetLight data. The quality of these data needs to be rigorously checked. Given their accuracies, they can then be used for appropriate applications.

DOTs in general are satisfied with probe data, which seem to be reliable on high-volume roadways (~8-10% penetration rate). For low-volume roadways, traffic data can be unreliable. The reliability of probe /crowdsourced data such as Waze and INRIX depends much on the number of users or data contributors. Therefore, such data for low-volume roads (e.g., rural roads) with few Waze or GPS users can be problematic. Although averaging data over long time periods can partially address this issue, the best way is to increase the number of data contributors.

Probe and crowdsourced data are widely used by DOTs. DOTs are aware of the limitations of such datasets but benefit from their wide coverage and high temporal granularity. DOTs are interested in a systematic assessment of the quality of such data (e.g., sampling rate, accuracy, and confidence level). There are also needs to maintain and modernize traditional sensors, to monitor their conditions and stream data to highway operation center (HOC) in real time.

There is a gap between infrastructure (e.g., bridge, pavement) assets management and ITS/TSMO assets management. Maine, Vermont, and New Hampshire are working on closing this gap, trying to develop an asset inventory for their ITS and TSMO related assets. This effort can be combined with modernizing traditional sensors. ITS assets are different from bridge, tunnels, etc. An ITS device may look normal but can stop working all of a sudden. Therefore, real-time status monitoring is important.

RIDOT always wants to continuously improve the quality of data from different sources, and they agree that it is important to check the quality of data from third-party vendors. It would be ideal that such work can be done at a regional level, instead of by individual DOTs, as the cost of a thorough study can be high. Consultants for RIDOT used StreetLight OD data to identify detour routes for projects. The results seemed to be good. However, RIDOT did not have the resources needed to verify the accuracy of such data. Sometimes, there are no data from traditional sources at all. In this case, it is beneficial to take nontraditional data sources into consideration, which probably is better than nothing.

RIDOT does not have a specific protocol to determine which data source(s) can be used for design and other purposes. At this moment, they approve the use of nontraditional data source(s) on a case-by-case basis.

MassDOT uses INRIX. Similar to other probe vehicle products, the sample size of INRIX varies across different roadways, which is usually 2.5-3% of all vehicles. MassDOT also has StreetLight data, which is primarily used for planning and safety purposes. MassDOT has done some informal studies to compare the traffic counts from StreetLight with field measurements. One concern is that StreetLight traffic counts are calibrated based on permanent counting stations on major highways. The accuracies of StreetLight traffic counts for local roads are not as good as those for major highways. MassDOT also has an initiative to install biking and pedestrian counters on some bike paths.

Like other DOTs, MassDOT has stationary sensors (e.g., loop detectors, microwave sensors) to collect traffic counts, speeds, etc. on major highways. Such data could potentially be used to calibrate the probe data, but there are some issues: (1) these sensors are not connected to the highway operation center (HOC), and traffic operators do not know the status of these sensors in real time. Sometimes MassDOT loses months of data from a sensor before realizing it; and (2) most sensors are for interstate highways, which makes it difficult to validate probe data for other highways. It would be helpful to add new sensors at strategic new locations on local highways for data validation. Also, being able to monitor sensor health status like what the ATSPM system does is important.

VTrans shares the same remark as MassDOT regarding the quality of probe data. VTrans conducted a comparison between Bluetooth data and probe data and found that probe data points with low confidence intervals should be considered with caution.

NHDOT did an internal comparison of TomTom, INRIX, and DOT sensor data. They found TomTom to be more accurate than INRIX, probably because INRIX relies more on freight data than TomTom. NHDOT also found TomTom data to match their own sensor data well.

MaineDOT is validating AADTs from StreetLight, which will be used for purposes other than TSMO.

The Eastern Transportation Coalition (TETC) has conducted many validation studies of the probe data primarily for highways. They plan to expand such efforts to cover arterials and local roads over the next few years.

3.2. Emerging Data Sources

3.2.1. Data Collection Methods

Question: Short- and long-term plans to meet agency's data needs while minimizing the life-cycle cost (e.g., relying on 3rd-party vendors vs. investing in data collection infrastructure) and maximizing the data collection system robustness and reliability (e.g., reliability of crowdsourced data depends on the number of data contributors).

Most state DOTs do not have a conclusion on the direction of future data collection methods. At the current stage, most DOTs are using both methods: DOT data collection infrastructure and third-party data vendors. DOTs will probably continue this practice in the near future.

MassDOT and VTrans have ongoing discussion on this issue but have no conclusions yet. In terms of safety analysis, VTrans is not using any emerging data sources (e.g., AADT from StreetLight) and is more in the wait and see stage.

CTDOT prefers third-party data vendors over DOT's own data collection infrastructure based on life cycle and maintenance costs. Maintaining DOT's own infrastructure often requires setting up short-term work zones. Using data vendors such as HERE does not require infrastructure investment, and it reduces personnel risk during data collection infrastructure maintenance. In Connecticut, Bluetooth is used mainly in smart work zones for travel time data collection. Traffic detection is shifting to systems like Wavetronix and Grid Smart, while in pavement loop detectors are no longer being installed.

RIDOT considers both their own data collection infrastructure and third-party data vendors. They are still maintaining and adding traffic counters and weather stations. The collected data can be compared with data products from third-party vendors, although such data are only for fixed locations. For the future, there is no clear path in terms of which direction to go (e.g., investing in their own data collection infrastructure vs. buying data from third-party vendors). They do recognize the benefits of leveraging data from private vendors, which means no initial investment and the cost and trouble involved in maintenance can be avoided.

Another challenge involved in making this decision is the changes in technologies, which happen constantly and are difficult for public agencies to keep up. Some technologies may become obsolete quickly, and there is risk in investing heavily in such technologies. From this perspective, it makes sense to let private companies bear this risk and buy data products from them. In most cases, private companies can adapt to technology changes more quickly than public agencies. The traditional model of design, build, and maintain may not work very well in the future. It is probably more beneficial to buy data as a service.

The maintenance of ITS and data collection infrastructure is a major challenge to DOTs. If DOTs own those infrastructure, they have to train staff, especially when new technologies are constantly introduced. Sometimes, they may have to hire vendors to maintain the infrastructure. Also, it is important to have consistency in technologies. Otherwise, DOTs may have many different types of devices on the roads, have to keep a wide range of spare parts, and require maintenance staff to be familiar with a variety of technologies. Simply letting vendors to handle operations and maintenance is not a perfect solution either. This leads to the data integration issue. For example, RIDOT is not getting any data from the toll roads. In addition, different vendors have different data structures and formats, which create barriers for integrating data from different sources for in-depth data analysis.

RIDOT is working on expanding their smart lighting infrastructure to remotely monitor LED light status via a wireless network. Smart lighting program itself is for improving lighting infrastructure maintenance practices. However, this can be a good opportunity to add additional sensors/devices for TSMO purpose (e.g., monitoring ramps and major interchanges) at a reduced cost. What RIDOT is doing here is similar to the Verizon [8] and AT&T [9] efforts discussed earlier in this report.

DOTs are constantly facing the question of whether we should repair and install additional microwave sensors or buy INRIX data. A main reason for this is the maintenance cost of existing sensors. If companies like INRIX can also provide volume data, many DOTs probably will lean more towards third-party data vendors to avoid the hassle and costs of maintaining existing sensors. Another issue with INRIX, TomTom, etc. is their reliability, which is heavily affected by the sample size. In Northern New Hampshire, the sample size for TomTom data is lower than Southern NH and the data quality sometimes is not good.

Overall, maintenance is critical in addition to building and investing in data collection infrastructure. Maintaining a state of good repair of ITS infrastructure is very important but has not been given enough attention. DOTs are not only concerned with the price of data products provided by vendors but are also concerned with the life-cycle costs for maintaining their own data collection infrastructure. Some DOTs think future data collection efforts will consider both options.

3.2.2. Emerging Data Sources

Question: Ridesharing companies (e.g., Uber, Lyft), logistics companies (e.g., UPS), Cell Phone data (e.g., from Verizon), and Connected and Automated Vehicles are producing tons of data each day. Any plans to utilize data from emerging sources for TSMO?

Some of these datasets are also included in the probe data such as HERE and INRIX. In this sense, they have been widely utilized by DOTs. Overall, DOTs are aware that there are a lot of emerging data sources beyond the probe data that could be very useful. But DOTs have not yet extensively used those. One reason is that DOTs have not seen convincing examples on how the emerging data can be used for TSMO.

Some mobile carriers such as Verizon have approached DOTs, but no specific applications have been implemented yet. For UPS and FedEx, MassDOT used their data in a SPaT corridor on Route 9. For ATSPM, MassDOT has equipped a lot of signals with ATSPM and is still trying to figure out what to do with the generated data.

VTrans has been approached by a lot of vendors of nontraditional data but does not have a plan to use these emerging sources in the short term. VTrans has done 24/7 traffic counting at some signals using cameras, which yielded accurate results. Compared to probe vehicle data, this method requires equipment installation and is hard to scale up.

UMass Lowell in currently involved in a study funded by Verizon. Cell phone data could provide a much larger sampling rate than current probe vehicle data.

RIDOT conducted an autonomous shuttle pilot study and does have the data from this study. RIDOT is not directly utilizing data from ridesharing companies, logistics companies, or mobile carriers. However, they recognize that data from such private companies (e.g., Tesla) will become more and more important, which may cause some of the existing ITS technologies to be obsolete (e.g., variable message signs). At this moment, they do not have a formal plan regarding how to prepare for such a future.

Autonomous vehicle (AV) companies have collected a huge amount of data, including high-resolution vehicle trajectories of AV and nearby vehicles, and roadway conditions (from camera/Lidar data). There is no clear legislation guidance on what data could and should be shared. AV companies are heavily concerned about data privacy and hence data sharing. On the other hand, public agencies (like DOTs) do not know what data is available, what to request from AV companies, and how to store the massive data securely and analyze them. RIDOT echoed that data sharing is important and mentioned that they had a hard time to get some of the data from their AV pilot. RIDOT has a policy director who deals with data privacy issues.

3.2.3. Data Sharing

Question: Transit agencies (e.g., GTFS data), DOT's maintenance vehicles (e.g., plow trucks), and Automated Traffic Signal Performance Measures (ATSPM) systems can also be used to generate a lot of valuable data. Any plans to coordinate different DOT divisions?

Most DOTs are using or planning to use AVL and their maintenance vehicles for data collection. Several DOTs have deployed the ATSPM system and are interested in knowing more about how such data can be used for improving traffic operations and safety.

VTrans is discussing with transit agencies to utilize their vehicles as mobile sensors, similar to plow trucks. VTrans is at the early stage of utilizing this data source.

It was agreed that data sharing among different DOT divisions would be beneficial. For example, NHDOT TSMO needs to work with traffic division to access traffic signal control data, and work with rail and transit to get the parking App data.

Overall, there are not many data sharing activities among different divisions of DOTs. There could be several main reasons for this. First, there are not many very urgent needs to share data. Second, people are not familiar with what kind of data other divisions own. Third, people do not fully understand the value of the data they have to other divisions.

3.3. Data Integration and Analysis

3.3.1. Artificial Intelligence (AI) and Data Analysis

Question: AI and edge computing technologies are making it possible to extract accurate traffic data using existing traffic cameras (or drone-mounted cameras) and turn cameras into smart sensors, better utilizing the existing data and data collection infrastructure. Is your agency investigating/interested in such technologies?

There are many potential AI and edge computing applications such as traffic camera data processing. Issues related to camera data processing include licensing, access, privacy, and recording, which limit DOTs' ability to test them. Besides cameras, AI has been used for modeling TSMO data, although not done in-house by DOTs. Another important application area of AI and edge computing is with CAV. Edge processing could help with the structuring and application of CAV data, allowing them to be more effectively utilized for downstream analysis.

No applications of AI or machine learning were noted within CTDOT TSMO division. Planning division may be using them. Testing of autonomous vehicles is planned on the CT Fast Track, a dedicated bus corridor, with autonomous buses.

MassDOT is interested in AI and edge computing technologies and applications and is working on pedestrian detection in tunnels using existing cameras. One issue is to bring all cameras to a state of good repair.

VTrans does not have an immediate plan to use AI and edge computing technologies. VTrans is concerned about the large efforts needed to upgrade and expand the existing camera network. VTrans currently streams traffic video back to HOC via Verizon. They find PTZ cameras are more prone to failure and prefer stationary cameras to avoid frequent hardware maintenance needs.

RIDOT is interested in such technologies. They have used drone for construction site inspection, and Miovision for traffic counting and detection. RIDOT also tried Grid Smart, a competitor of Miovision.

Overall, although New England state DOTs have not conducted any in-house case studies using AI, machine learning, and edge computing, they are interested in any innovative technologies that would help improve efficiency and reduce costs.

As part of DelDOT's AI-ITMS project, they are developing and will be testing machine vision capabilities. Their goal is to start replacing in pavement detection technologies with "non-intrusive" detection technology. The total transition will take many years, considering DelDOT is responsible for between 20,000 and 30,000 loop detectors. DelDOT is reviewing detector design requirements to get the full potential of AI. They have developed three project areas that have different transportation demands and will use these areas to develop system requirements and test technologies. A mix of fixed view and PTZ technologies will be used.

3.3.2. In-House Data Analysis or Outsourcing

Question: Should data analysis be done in house or by consultants? How to integrate the data analysis efforts of different DOT divisions?

Most DOTs do data analysis work both in house and by consultants, depending on the type of work, resources available, and their workload.

CTDOT TSMO division relies mostly on consultants for data analysis work. They are using tools developed by UConn and the University of Maryland (e.g., RITIS). Their planning office does some data analysis work in house.

MaineDOT does some of the data analysis work in house. They are facing a critical need of conflating data sources. MaineDOT would like to use as much non-TSMO data as possible in operations analysis. However, data from different sources are organized using different spatial units/referencing systems. Integrating them requires major efforts.

VTrans currently hires consultants to do data analysis. Meanwhile, VTrans is building a data analysis team within their own agency, hoping to do the analysis in house ultimately. It is still unclear to what extent the in-house analysis will cover. For the in-house data analysis, the priority areas probably will be crash data and traffic data (e.g., probe data, ITS data).

Many DOT vendors are storing their data on cloud servers. IT support is important for DOTs to integrate data from different sources (e.g., downloading those datasets to a local server and integrating them). It is anticipated that DOTs will not rely entirely on third-party data, and it would be important to integrate data from DOTs and vendors. Data vendors are profit-driven and are not always motivated to integrate their data with DOT data, and some of the data integration work may have to be done by DOT staff. This issue is related to workforce development, DOT staff training, etc.

For RIDOT, it depends. Sometimes DOTs do not have the resources to do the work. It would be nice to hire consultants to do it. However, it is helpful to have experts in DOTs knowing what consultants are doing and what DOTs are paying for. RIDOT recently has hired data analysts in the performance measure division, suggesting that RIDOT certainly recognizes the needs of in-house data analytics.

At RIDOT, traffic cameras are not synchronized with Waze or INRIX. Theoretically, it would help if these systems were integrated. For example, if Waze reports an incident, the camera feed (if available) covering that incident location should be displayed on the screen and highlighted.

Even if Waze and INRIX may sometimes report false incidents, this is still better than randomly going through different camera feeds. RIDOT is currently working on integrating data from Waze and INRIX to get the most out of them.

NHDOT would like to do the data analysis work in house if resources are available. Otherwise, they may have to outsource the work.

3.3.3. Investing in Data Analytics

Question: Short- and long-term plans for investing in data analytics and workforce development.

Although DOTs may not have specific plans to invest in data analytics and workforce development, they all recognize the importance of such efforts. Several DOTs have recently hired data analysts/scientists or created related positions.

3.4. Data Archiving, Sharing, Security, and Privacy

3.4.1. Data Management Protocols

Question: Any protocols for how data should be reduced and how long should a dataset be archived (in the original form or reduced/processed form)?

Overall, how long a dataset should be kept depends on the nature of the data and agency data retention policies. DOTs all have some kind of policies for data retention, privacy, and security. However, most DOTs struggle with the growing data volumes and how to extract insights out of the massive data.

In terms of data management, historically there was some coordination between divisions in CTDOT, however, it no longer exists and there are no plans on the horizon for it. On the ITS side, there are no data experts in house, and consultants are hired as needed. Data is archived in accordance with state retention requirements set by CTDOT legal department. For example, traffic sign messages displayed during road closure are archived for several years. For data not covered by state retention requirements and not related to DOT liability issues, their storage is decided by different units and offices of CTDOT.

MassDOT is concerned about the huge amount of data they have and recognizes the needs to reduce the data and keep the essential information for long-term storage. MassDOT also struggles with how to analyze the data. Vendors have to comply to relevant MassDOT policies to protect privacy and ensure data security.

VTrans has a similar concern regarding the large amount of data owned by them. Data storage is a major challenge for IT. VTrans historically used DOT owned storage but is moving towards outsourcing data storage (currently transitioning to Amazon web service). It is critical to involve IT in data collection/storage/sharing to make sure that data sharing comply with privacy and security policies.

There has also been discussion on this topic at NHDOT.

3.4.2. Measures for Protecting Privacy and Security

Question: What kind of measures is in place to protect privacy and security?

MassDOT has a position called Director of Data and Policy. NHDOT does not have this position, but they recognize the importance of data privacy and security. NHDOT has IT embedded into the TSMO division to take care of data security issues.

TxDOT has an Information Security Policy Manual, which contains policies for TxDOT's information security functions and creates a dynamic program that protects the confidentiality, integrity, and availability of TxDOT's information resources. All TxDOT employees are required to take annual refresh training course on security topics. ODOT also has a policy to protect data privacy.

3.4.3. Data Sharing

Question: Standards and protocols to guide practices such as: what kind of data should (or should not) be shared and how to share them (e.g., using Amazon Web Services or DOT owned servers)?

Most DOTs use both third-party cloud services and in-house servers for data storage. Most states have their own record retention policies that apply to the collected data.

RIDOT is investigating hosting data in the cloud such as Amazon. Most RIDOT systems are using DOT servers on premise. Some new systems have to be hosted in the cloud, since that is the only option. For example, e-permitting systems can only be hosted in the cloud. RIDOT IT is working on moving GIS infrastructure to the cloud. Policies are being developed now at the state level. There will be a cost shifting from hiring IT staff and purchasing hardware to pay for cloud services.

RIDOT has a record retention policy, which mostly deals with infrastructure related records. There is some information in there related to non-infrastructure regarding how long records or data should be or needs to be retained, and when can it be destroyed. This policy may need to be updated especially for non-asset related data. For new system vendors who host data in the cloud, they have to meet state data security requirements and provide a data retention plan.

NHDOT's ATMS is using amazon web services (AWS). Their asset management system is hosted in the cloud by consultants. Some homegrown systems are hosted on DOT servers.

Most data collected by public agencies are subject to the public information act. A rule of thumb for data collection is not to collect any data with personal identifiable information (PII) or business confidential information. That way, DOTs would not encounter any issues with sharing the collected data. TxDOT is gradually migrating to cloud-based platform for hosting and sharing data.

3.5. Stakeholders and Workforce

3.5.1. Stakeholders

Question: List of stakeholders? We may want to identify stakeholders within DOT to support the investment in data collection and analytics.

Stakeholders identified include:

- Internal: IT, ITS, TMC, safety/traffic, GIS, maintenance, planning, performance measures, communications to the public, legal. Literally, all divisions within DOT should be key stakeholders for data collection and analytics because any decision-making process should be data-driven.
- External: 911, transit, emergency response/first responders, state police, MPO

It is important to get IT involved, especially for system and data integration, and maintenance. NHDOT TSMO has IT embedded and does not see this as an issue. Planning does not need real-time data, but they can certainly benefit from the TSMO data for calculating performance measures.

3.5.2. Organizational Structure

Question: What kind of organizational structure changes are both feasible and necessary to better prepare DOTs for the future? For example, Iowa DOT has an Office of Analytics; Arizona DOT has a Data Analytics section responsible for reporting, maintaining, collecting, analyzing, and visualizing the data on roadways in Arizona; and Florida DOT has a Transportation Data and Analytics (TDA) Office that is FDOT's central clearinghouse and the principal source for highway, traffic, travel time, multimodal, and freight and passenger data information.

There is no one-size-fits-all solution regarding organizational structure, which should be based on the needs and available resources of the specific organization.

MassDOT has been outsourcing some data analytics work but also has a robust data analytics team in house.

VTrans is ultimately going to be responsible for collecting, analyzing, visualizing, reporting, and maintaining data. They currently have one GIS expert analyzing the data.

RIDOT and the state have identified this as a key direction. Data analyst positions have been created within RIDOT. Instead of centralizing these positions, they are being distributed to different divisions of RIDOT. This may cause coordination problems. The GIS (or Transportation Information System) group currently has programmers and GIS professionals to work together and coordinate their efforts. Having a central data analytics division to coordinate efforts is important. Creating a central office will give data analysts a sense of belong to a core group, making it easier for them to exchange ideas and learn from each other.

NHDOT does not have a central data division/office to handle all the data analysis needs. They would like to know what other DOTs are doing. Having an innovative bureau or advanced technology office would be helpful. Such a division can be integrated with the grants division.

ODOT has a Data Section and a Transportation Planning Analysis Unit. VDOT has an Office of Strategic Innovation for accelerating the identification, piloting, and scaling of innovations across the agency, providing divisions with support to purse innovative ideas, and fostering a culture within the agency which values and celebrates innovation.

3.5.3. Workforce Development Needs

Question: Workforce development needs and current strategies?

Workforce development is important. Also, it is important to provide competitive salaries to attract and retain skilled data analysts. DOTs need to hire data analysts with both transportation and data science background. Data scientists without civil/transportation background may not understand the transportation nuances that may cause problems. It would be helpful to train civil students and ask them to take data analytics courses. These new hires can focus on transportation data analytics and commit to it.

Mainstreaming the importance of data and data analytics would be a great strategy. Training existing workforce to acquire needed skills in data would be another strategy.

Many DOTs (e.g., VTrans) are hiring people with data analytics skills. At Oregon DOT (ODOT), a new position named Chief Data Officer has been created.

DelDOT has in-house training. They have ITMS related training developed by consultants and will provide external training. DelDOT deliberately created TMC Technician positions to be career ladder. A critical part of determining promotion requirements is the duties required and the complexity of the system. Programs that include advanced traffic signal systems are more complex.

3.5.4. Additional Thoughts

Question: Additional Thoughts: Any thoughts you have related to data driven TSMO applications.

There is no boundary as far as collaboration in data-driven TSMO applications goes. Collaboration among divisions within DOT is important. Inter-agency collaboration in this area is extremely important as well. It is also desirable to look into the impacts of CAV on TSMO.

4. DETAILED ANALYSIS OF SELECTED DATA SOURCES

Based on our review, three new/emerging data sources are singled out for further discussion in this section due to their popularity and being representative.

4.1. NPMRDS, INRIX, and RITIS

The National Performance Measurement Research Data Set (NPMRDS) consists of archived probe data provided by INRIX. INRIX data are based on GPS readings from fleet vehicles such as delivery vans, long-haul trucks and taxis, users of the INRIX Traffic App, and connected vehicles. All public agencies and their consultants can have free access to it. NPMRDS covers all segments of the National Highway System (NHS). Different from INRIX and other commercial probe data products, the NPMRDS data is only based on the actual speeds of observed probe vehicles, not imputed speeds using historical data [1].

RITIS stands for Regional Integrated Transportation Information System (RITIS). It fuses data from INRIX and local transportation agencies. The information available through the RITIS platform (also INRIX) is aggregated over one minute time intervals and distinct roadway segments. RITIS previously divided roadways into Traffic Message Channels (TMC). Now it uses a new system called eXtreme Definition (XD) to divide roadways. In general, XD segments are shorter than TMC and thus provide more granular information. The lengths of TMC segments vary significantly. Some TMC segments are longer than 30 miles, while other are only 0.2 miles long. XD segments are more uniform in length, and they are mostly shorter than one mile.

For each TMC/XD, the following data are provided by INRIX/NPMRDS:

- One-minute space mean speed,
- Travel time,
- Reference speed,
- Confident Value (C-Value), indicating how well the current reading represents the actual roadway condition based on recent and historical trends, and
- Confidence Score, a discrete variable indicating whether or not the reported value is real-time data. It takes three possible values:
    - 30: the reported value is based on real-time data, and the corresponding segment has adequate GPS readings,
    - 20: the reported value is based on historic averages, and the segment does not have sufficient real-time readings (15-minute granularity), or
    - 10: the reported value is based on the reference speed, and there are no real-time readings.

INRIX provides data aggregated using different time intervals ranging from five minutes to one hour. Figure 3 shows the INRIX data (speed, travel time, confidence score and C-value) for an XD segment on I-90 WB affected by incident. The x-axis is time measured in minutes. Other probe data such as HERE and TomTom are similar to INRIX and are thus not discussed here.

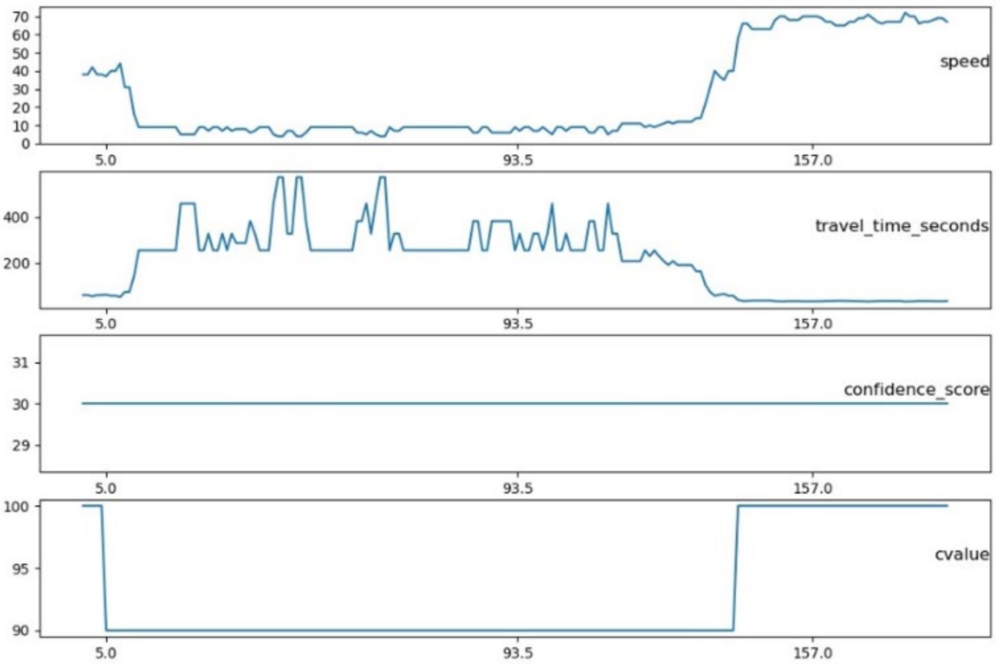


Figure 3 INRIX Data for XD segment 429079803

4.2. Wejo and Otonomo Data

The data provided by Wejo is summarized in Table 2. Different from raw probe data, Wejo also includes information such as exterior temperature, wiper state change, wiper interval, acceleration change type, and autonomous emergency braking type. The temperature and wiper status information can be integrated into DOT's Maintenance Decision Support System (MDSS). The acceleration change type and emergency braking data can be very useful for improving safety. The location, speed, heading, and ignition status data in Table 2 are collected and streamed continuously to a data warehouse. While for other event-driven data such as exterior temperature, wiper state change, and wiper interval, they are collected only when an event (e.g., wiper started) has occurred to reduce the amount of data that needs to be transmitted in real time.

The Otonomo data consists of four components: trip points, trips, road signs, and events. Since we do not have access to Otonomo data, no details regarding each component can be provided. However, the data elements provided by Otonomo should be very similar to those by Wejo.

Table 2 Key Data Elements Provided by Wejo

Data Item	Description
Datapoint ID	Unique identifier for an individual captured datapoint
Journey ID	Unique identifier for an individual vehicle's movements from Journey Start to Journey End
Vehicle Type ID	An identifier for the category of vehicle including body type, fuel type and model year
Date & Time	Timestamp captured for each datapoint (ISO:8601), UTC including Time Zone offset to calculate local time
Latitude & Longitude	The North-South and East-West positioning of the vehicle on the Earth's surface
Speed	The speed in kilometers per hour that the vehicle was traveling when the datapoint was captured
Heading	The direction that the vehicle was heading when the datapoint was captured
Ignition Status	Representation of ignition state when the datapoint was captured
Event type	An identifier for the recorded event
Journey event change type	Ignition on or ignition off
Seatbelt change type	Latched or unlatched
Acceleration change type	Harsh braking or harsh acceleration
Speed threshold event type	Speed above or below threshold
Odometer	Total distance travelled over the life of the vehicle
Fuel consumption	Total fuel used over the life of the vehicle
Fuel Level	Populated when a 'FUEL_LEVEL_CHANGE' event is triggered: Current level of fuel represented as a percentage of total capacity at the time of capture
Exterior Temperature	Exterior temperature recorded by the vehicle at the time datapoint was captured
Seat Occupancy	String representation of the seat occupancy status captured as: OCCUPIED, UNOCCUPIED
Wiper State Change	String representation of the front wiper status captured as: ACTIVATED. DEACTIVATED
Wiper Interval	Number representation of wiper interval at the point the wiper event was captured: 0: default, 1 - 4: interval speed low - high
Autonomous Emergency Braking Type	String representation of the autonomous emergency braking status captured when the autonomous emergency braking is engaged as: ACTIVATED

4.3. StreetLight Data

StreetLight uses GPS data generated by Location Based Service (LBS) Apps installed on smartphones. Such data are associated with individual devices with unique IDs. With the GPS coordinates associated with a device throughout a day, it is possible to find out where the device owner has been, by what mode, using what route, at what speed, and stayed for how long? The same device may have been used for online shopping. Based on the online shopping activities, additional socio-demographic information of the device owner can be estimated. The raw LBS App data are quite noisy, as LBS Apps generate GPS coordinates at different frequencies. Some companies acquired such data and further cleaned them (e.g., removing privacy related information) and are selling them.

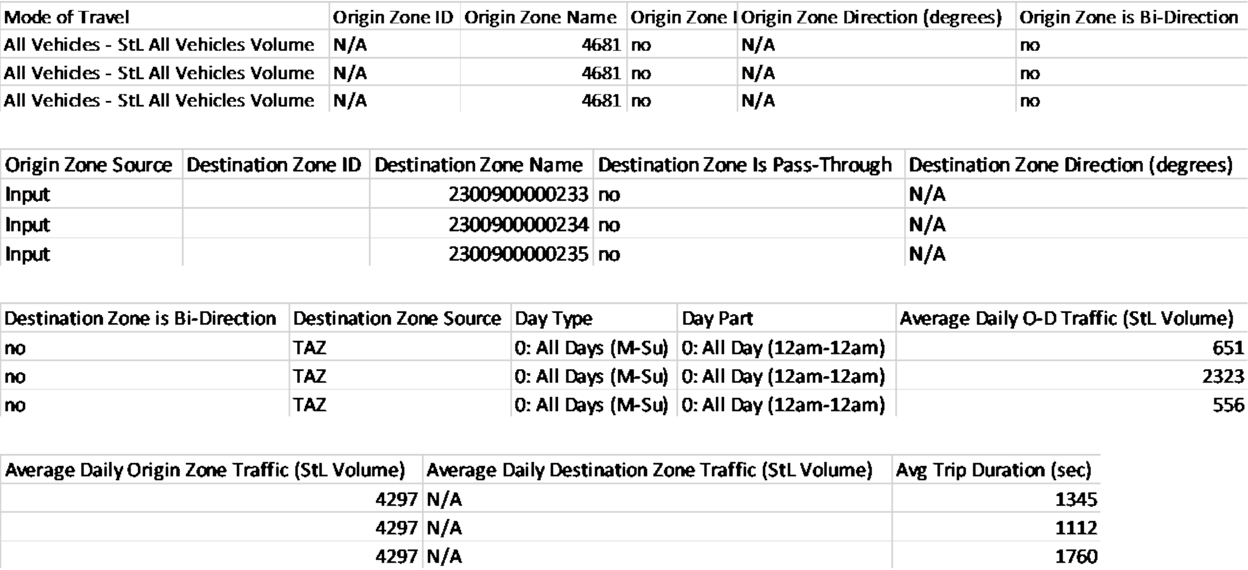
The LBS App data can be used to derive speed, travel time, travel mode, and route choice data for individual devices. By integrating LBS data with loop detector counts, land use, employment, and other transportation planning datasets, traffic volume and OD information for different modes (e.g., vehicle, pedestrian, bike, bus, and rail) can also be estimated.

StreetLight provides five types of analysis options, which are modular analysis, exploratory analysis, segment analysis, turning movement counts, and AADT, with the names providing some explanation as to what they do. All these analyses start with a user defined map of areas of interest, called Zones in StreetLight. Zones can represent either areas, roadways, railways, or pathways and can be created within StreetLight or in a GIS program and uploaded as a ESRI Shapefile.

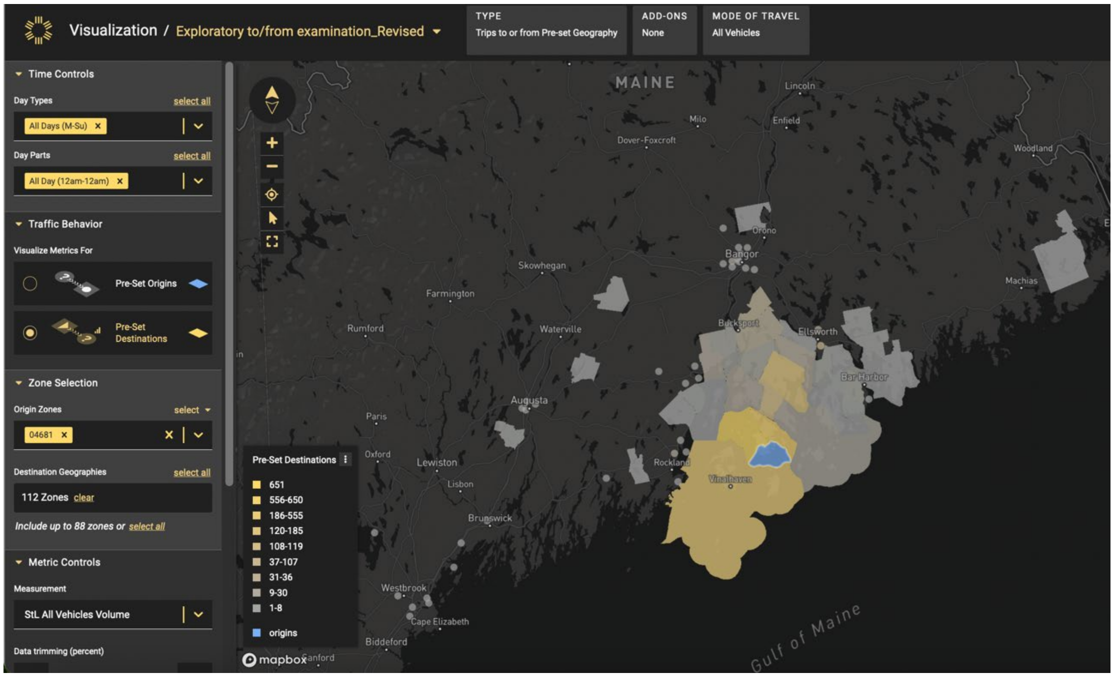
- *Modular Analysis* provides estimated trip counts between zones (i.e., OD data) by travel mode, the route choice results for each OD pair, average trip speeds and lengths, and traveler demographics. Among them, OD, route choice, trip speeds and lengths can be provided using intervals as short as 15 minutes.
- *Exploratory Analysis* is similar to *Modular Analysis*. It provides trip counts on the most frequently used routes into, out of, and between zones. Average trip speeds and lengths are also available. The smallest reporting interval in this case is 1 hour.
- *Segment Analysis* provides traffic counts, average speeds, speed distributions, and speed percentiles for individual roadway segments. Data are available for cars, medium trucks, and heavy trucks separately in intervals down to 15 minutes.
- *AADT Analysis* generates estimated AADTs for roadway segments of interest.
- *Turning Movement Counts (TMC) Analysis* provides TMC data by mode for three- and four-leg intersections in intervals down to 15 minutes.

The StreetLight analysis results can be viewed in its internal visualization tool or downloaded as .csv files. Table 3 shows a sample .csv file downloaded from StreetLight describing trips from/to a zone. The spreadsheet has been broken into four rows to better fit the page. Figure 4 is an example of the StreetLight visualization tool showing how trips from an origin (the blue polygon in the figure) are distributed among different destination zones (other polygons of different colors).

Table 3 Default StreetLight Output Showing Trips to/from a Zone



Mode of Travel	Origin Zone ID	Origin Zone Name	Origin Zone I	Origin Zone Direction (degrees)	Origin Zone is Bi-Direction	Origin Zone Source	Destination Zone ID	Destination Zone Name	Destination Zone Is Pass-Through	Destination Zone Direction (degrees)	Destination Zone is Bi-Direction	Destination Zone Source	Day Type	Day Part	Average Daily O-D Traffic (StL Volume)	Average Daily Origin Zone Traffic (StL Volume)	Average Daily Destination Zone Traffic (StL Volume)	Avg Trip Duration (sec)
All Vehicles - StL All Vehicles Volume	N/A	4681	no	N/A	no	Input		2300900000233	no	N/A	no	TAZ	0: All Days (M-Su)	0: All Day (12am-12am)	651	4297	N/A	1345
All Vehicles - StL All Vehicles Volume	N/A	4681	no	N/A	no	Input		2300900000234	no	N/A	no	TAZ	0: All Days (M-Su)	0: All Day (12am-12am)	2323	4297	N/A	1112
All Vehicles - StL All Vehicles Volume	N/A	4681	no	N/A	no	Input		2300900000235	no	N/A	no	TAZ	0: All Days (M-Su)	0: All Day (12am-12am)	556	4297	N/A	1760



Visualization / Exploratory to/from examination_Revised

TYPE: Trips to or from Pre-set Geography

ADD-ONS: None

MODE OF TRAVEL: All Vehicles

Time Controls

Day Types: All Days (M-Su)

Day Parts: All Day (12am-12am)

Traffic Behavior

Visualize Metrics For: ○ Pre-Set Origins ● Pre-Set Destinations

Zone Selection

Origin Zones: 04681

Destination Geographies: 112 Zones clear

Include up to 88 zones or select all

Metric Controls

Measurement: StL All Vehicles Volume

Data trimming (percent)

Figure 4 StreetLight Visualization Tool

5. REFERENCES

1 The Eastern Transportation Coalition. (2019). Optimal Traffic Monitoring in a New Data Age.

2 Tanaka, A., Schroeder, B., Trask, L., & Chase, T. (2019). NCDOT Guide on Automated Traffic Signal Performance Measures. North Carolina Department of Transportation.

3 Enwemeka, Z. (2015). Boston Tapping Waze Data to Help Improve Traffic Signals, Ease Congestion. Available online at https://www.wbur.org/news/2015/02/13/waze-boston-data-partnership, accessed on 10/26/2021.

4 Desai, J., Li, H., Mathew, J. K., Cheng, Y. T., Habib, A., & Bullock, D. M. (2021). Correlating hard-braking activity with crash occurrences on interstate construction projects in Indiana. *Journal of Big Data Analytics in Transportation*, *3*(1), 27-41.

5 The Eastern Transportation Coalition. (2020). Hurricane Proof of Concept Results States' Experience with Real-Time Connected Vehicle Data Coalition.

6 Thompson, S.A. & Warzel, C. (2019). Twelve Million Phones, One Dataset, Zero Privacy. Available online at https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html, accessed on 10/26/2021.

7 Keegan, J. & Ng, A. (2021). There's a Multibillion-Dollar Market for Your Phone's Location Data. Available online at https://themarkup.org/privacy/2021/09/30/theres-a-multibillion-dollar-market-for-your-phones-location-data, accessed on 10/26/2021.

8 Verizon. (2017). How the Internet of Things can create congestion-free cities - Verizon's Intelligent Traffic Management solution is already changing the way cities plan their streets. Available online at https://www.verizon.com/about/news/how-internet-things-can-create-congestion-free-cities, accessed on 10/26/2021.

9 AT&T. (2018). AT&T Expands Smart Cities Offerings with New Structure Monitoring Solution for U.S. Railways and Roadways. Available online at https://about.att.com/story/structure_monitoring_solution_for_railways_and_roadways.html, accessed on 10/26/2021.

10 Sawers, P. (2020). StreetLight Data raises $15 million to bring big data analytics to city transport Available online at https://venturebeat.com/2020/08/06/streetlight-data-raises-15-million-to-bring-big-data-analytics-to-city-transport/, accessed on 10/26/2021.

11 Lambert, F. (2020). Tesla is collecting insane amount of data from its Full Self-Driving test fleet. Available online at https://electrek.co/2020/10/24/tesla-collecting-insane-amount-data-full-self-driving-test-fleet/, accessed on 10/26/2021.

12 FHWA (2021). Sources Sought Notice 693JJ3-21-SS-0013 AI for ITS. Available online at https://govtribe.com/file/government-file/693jj3-21-ss-0013-ai-for-its-dot-pdf, accessed on 10/26/2021.

6. APPENDIX A – NHDOT TSMO ASSET MANAGEMENT SYSTEM OVERVIEW

7. APPENDIX B – NHDOT TSMO ASSET MASTER DIAGRAM