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

Tasks 2 and 3 Draft Report:

Task 2 - Assessment of Data Needs, Emerging Data Sources, and Data Processing and Analytics
Task 3 - Recommendations

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)

April 2022

TABLE OF CONTENTS

1. ASSESSMENT OF DATA NEEDS, EMERGING DATA SOURCES, AND DATA PROCESSING AND ANALYTICS

The section provides recommendations on data needs, emerging data sources, data processing and analytics, and others to state DOTs in the New England region.

1.1. Recommendations on Data Needs

ID	Data Needs	Recommendations
1	• Incident detection • Traveler Information Systems (TIS) • Travel time estimation	The existing probe data (e.g., TomTom, INRIX) in general provides a good coverage of highways. The penetration rates of emerging connected vehicle data (e.g., Wejo, Otonomo) are continuously growing. DOTs should not invest in additional roadside sensors such as Radar and camera for incident detection, TIS, and travel time estimation purposes, unless it is for areas that are poorly covered by the above data sources, or these data sources are unreasonably expensive.
2	• Vehicle trajectories	Safety is an important aspect of TSMO. Safety analysis has been done reactively and based primarily on historical crash data. It is interesting to use vehicle trajectory data to proactively evaluate safety risk in the future. Vehicle trajectories from connected vehicles (e.g., Wejo, Otonomo) cover a large area but only a small sample of all vehicles. Roadside sensors (e.g., high-resolution Radar, camera, LiDAR) cover a short road segment but can capture all passing vehicles. Both data sources are important for proactive safety risk analysis. DOTs are encouraged to investigate both data sources (i.e., connected vehicles and roadside sensors). When investing in new roadside sensors, DOTs are encouraged to consider sensors that can generate vehicle trajectories during both day and night.
3	• Passenger and freight OD	Data from mobile device GPS (e.g., location-based service data) and various vehicle ReID technologies make it possible to derive traffic OD for a large geographic area. This may potentially be done for passenger vehicles and heavy trucks separately. Such OD information is not only important for planning purposes, but also will substantially increase DOTs' ability to understand driver behavior and predict transportation system use and response to disruptions. TSMO and planning divisions are encouraged to work together on deriving and evaluating OD information using LBS and vehicle ReID data.

ID	Data Needs	Recommendations
4	• Traffic volume and capacity	Existing probe data only covers speed and travel time. Estimating traffic volume and capacity (e.g., under different weather conditions) can be very interesting. Such information can be used together with OD to predict when congestion (not caused by incidents) may occur and the corresponding queue growing and dissipating processes. Although some data vendors claim that they can provide traffic volume data such as segment AADTs and intersection turning movement counts, the accuracy of such data needs to be thoroughly evaluated, especially for rural areas where there are not many permanent traffic monitoring stations to provide calibration data. Existing traffic monitoring stations are mainly on major highways to satisfy the HMPS requirement. DOTs should expand the station network using roadside sensors. Such sensors may also be used to provide vehicle trajectory data for safety analysis, vehicle OD, and detailed vehicle classification data (see below).
5	• Detailed vehicle classification and ReID data	AI technologies make it possible to detect, track, and classify vehicles reliably from RGB camera, thermal camera, Radar, LiDAR, and traditional loop detectors. For example, retrofitted loop detectors and camera + AI technologies can differentiate among vehicles such as flatbed, dry goods semitrailer, tankers, refrigerated trucks, and recreational vehicles. DOTs are encouraged to consider such technologies. DOTs are not encouraged to install new loop detectors due to their high installation and maintenance costs. However, retrofitting existing loop detectors can extend their service life and generate more useful information.
6	• EZ-pass and Bluetooth data	For areas without good probe data coverage, DOTs are encouraged to consider installing Bluetooth sniffers/readers to collect travel time data. DOTs can also install sensors to read EZ-pass transponders. For example, New York City has been using EZ-pass transponder data to track vehicles and measure travel time.
7	• Corridor freight data	Parking information along major corridors such as I-95 is important for truck drivers. DOTs may use camera + AI + edge computing + 4G technologies to collect and share such information.

ID	Data Needs	Recommendations
8	• ITS asset condition data	Detailed and real-time condition information about ITS assets is critical. This is especially true for traffic controllers (e.g., ATSPM) and ITS assets that provide real-time traffic data. Tracking such data is important for ensuring system safety (e.g., a malfunctioning traffic signal can cause accidents) and developing preventative maintenance plans. It is strongly recommended that DOTs invest in this area. Some of the data does not need to be transmitted to the Traffic Management Center (TMC) in real time. For instance, the detector condition data may be reported every hour instead of minute to the TMC.

1.2. Recommendations on Emerging Data Sources

ID	Emerging Data Sources	Recommendations
9	• Connected vehicles and travelers	It may take many years for automated vehicles to occupy the streets. However, connected vehicles are very close to us now. Many auto makers have already been collecting data using their new vehicle models. These datasets are packed and sold by companies such as Wejo and Otonomo. They include vehicle trajectories as well as event logs such as wiper speed and activation/deactivation. Travelers nowadays depend heavily on mobile devices and various Apps, even knowing that their privacy is at risk. These mobile devices and Apps are contributing critical data (e.g., StreetLight) for understanding traveler behavior under different traffic conditions. Useful information can be derived from such data sources, including OD, route and mode choice, driver behavior, and safety issues associated with highway geometric designs. DOTs should explore the potential applications of such datasets and their impacts on traffic operations and safety. DOTs should also work with legislators to push technology companies such as Google to make such datasets available to public agencies. Such datasets are collected from the public and probably should be made available for free or at a reduced price to public agencies for the benefits of whoever contribute the data.

ID	Emerging Data Sources	Recommendations
10	• Sensors powered by AI and edge computing: thermal and RGB cameras, loop detectors, LiDAR, Radar, EZ-pass transponder	Advanced sensors powered by AI and edge computing technologies will be another important data source. Thermal and RGB cameras can detect, track, and classify vehicles, pedestrians, and bicycles. They can detect lane changing activities, vehicles stopped in the emergency lane, bus lane violations, reidentify vehicles at different locations, etc. High-resolution LiDAR and radar can generate more accurate vehicle speed and location information than cameras and cover larger areas. Vehicle signatures from retrofitted loop detectors can be used to classify and reidentify vehicles. New York City has been using EZ-pass transponder data to estimate travel time. DOTs are encouraged to explore the potential of traditional and new sensors mounted on portable platforms. These portable platforms can be moved to different locations to (1) collect trajectory data for safety studies, and (2) collect speed and travel time data to complement the probe and connected vehicle data in rural areas.
11	• Automated vehicle data	Car manufacturers such as Tesla are collecting a huge amount of data (e.g., videos, vehicle control parameters) from vehicle owners. The data covers driver behavior and the surrounding environment. For example, Tesla uses such data to calculate safety scores for drivers. Such data can also be used to detect road debris, pavement cracks, pavement marking conditions, damaged traffic signs, problematic highway geometric designs, etc. There are already commercial products based on probe (e.g., INRIX) and connected vehicles (e.g., Wejo, Otonomo) data. It is anticipated that there will be commercial datasets available in the future that are collected by semi- or fully automated vehicles. DOTs should take this potential data source into consideration when making future data and data collection infrastructure decisions.

ID	Emerging Data Sources	Recommendations
12	• Relying on data vendor vs. investing in data collection infrastructure	Some DOTs are reluctant to invest in new roadside traffic sensors such as inductive loop, Radar and camera due to installation and maintenance costs. They are more willing to simply purchase probe data. DOTs should conduct studies to compare the life-cycle costs of relying on data vendors and their own data collection infrastructure. In the future, DOTs can invest in mobile/portable data collection units (similar to portable variable message signs) for areas that are not well covered by probe data. These portable data collection units can also be used to collect trajectory data for safety studies. Also, DOTs should invest in retrofitting existing traffic cameras and loop detectors using AI and edge computing technologies to expand the capacities of these traditional sensors. DOTs may work together and develop data and communication interface standards for vendors. In this way, DOTs can easily switch from one vendor to another to obtain the same data elements. This flexibility and independence may potentially increase the competition among vendors and reduce the sensor maintenance and replacement costs.
13	• Data quality validation	DOTs should continuously monitor the quality of probe and connected vehicle data, particularly for rural areas where the penetration rates might be low.

1.3. Recommendations on Data Processing and Analytics

ID	Data Processing and Analytics	Recommendations
14	• Data integration and conflation	It would be interesting to integrate crash history, pavement condition, and probe vehicle data to find connections among them. However, these datasets are organized using different referencing systems. Crash data is often based on x and y coordinates; pavement condition data is typically stored using linear referencing systems; while probe data is organized by segments (e.g., INRIX previously used TMC and is now using XD segments). Data conflation is a major issue faced by many DOTs and should be given enough attention.

ID	Data Processing and Analytics	Recommendations
15	• More detailed incident data analysis	With probe data such as TomTom and INRIX, DOTs can derive more detailed incident information, including duration, queue length, clearance time, and effects on secondary incidents. Such information can be correlated with incident characteristics such as # of lanes closed, # of vehicles involved, and injury and casualty to establish models to predict future incident impacts. In addition, probe data can be used to separate recurring congestion from incidents and for queue detection and warning. The recurring congestion information in conjunction with OD and travel mode choice (e.g., from StreetLight) data can be used to develop comprehensive transportation network improvement solutions. DOTs are encouraged to explore this area and conduct more detailed analysis of probe data.
16	• Connected vehicle data analysis	USDOT has funded three connected vehicle pilot projects. These vehicles have generated a huge amount of exciting data. In the meantime, many auto makers have already been collecting data using their new cars. These datasets are packed and sold by companies such as Wejo and Otonomo. These datasets are not aggregated by segments (like what TomTom and INRIX do) and contain more details. DOTs are encouraged to investigate such datasets and explore their applications beyond incident detection and travel time estimation. They can potentially be utilized to estimate crash risk and identify safety issues due to inappropriate highway geometric designs.
17	• Effective utilization of existing data	Existing datasets are not effectively utilized or explored. For example, StreetLight data is mainly used for planning purpose. It can provide useful OD and mode/route choice information for developing contingency traffic management plans for special events, major construction projects, and accidents. Data from loop detectors are often not streamed to highway operations center in real time. Traffic cameras are only used for incident verification and traffic videos are reviewed manually. Waze data is not seamlessly integrated with INRIX or TomTom data for incident detection/verification. DOTs are encouraged to explore methods to integrate such data sources and automate the process of integrating them.

ID	Data Processing and Analytics	Recommendations
18	• ATSPM data analysis	Several New England State DOTs have implemented or are planning to implement the Automated Traffic Signal Performance Measure (ATSPM) system. ATSPM allows DOTs to detect traffic signal related hardware and control plan issues in real time and remotely from the Traffic Management Center (TMC), identify potential causes, and quickly dispatch staff as needed. It helps to minimize the impacts of traffic signal control malfunction and improve traffic safety at signalized intersections. ATSPM systems generate high-resolution (e.g., every 1 second) detector and signal controller data (e.g., detector on/off, green light on). How to effectively utilize such data beyond calculating Signal Performance Measure (SPM) is a very interesting question, which has not been adequately investigated. ATSPM is getting increasingly popular. DOTs are encouraged to explore such datasets for both traffic operations and safety applications.
19	• Innovative data analysis methods	Emerging data sources such as probe vehicle, connected vehicle, and ATSPM require innovative data analysis methods. For example, previous incident detection methods based on loop detectors are not applicable to probe vehicle data. DOTs should investigate innovative analysis methods to get the most out of these new data sources.
20	• Data sharing and brainstorming	DOTs are encouraged to share data with the public when applicable. This may help to generate new application ideas. For example, MBTA makes real-time GTFS data public, based on which many mobile Apps have been developed without costing MBTA anything. With the shared data, DOTs may hold data analytics competition among college and high school students to identify interesting ideas and attract students into the transportation data analytics area.
21	• AI + Edge computing for data analysis and reduction	Most DOTs struggle with the growing data volumes and how to extract insights out of the massive data. With real-time data at more granular levels, DOTs need to investigate how to best process and store the data, and how not to overwhelm communication and computing systems. For example, DOTs are encouraged to explore AI and edge computing technologies to speed up the processing of images and videos. This will significantly reduce the amount of data that needs to be transferred and stored. DOTs are encouraged to work with universities on this topic.

ID	Data Processing and Analytics	Recommendations
22	• Road Weather Information System	Although all six New England state DOTs have invested a lot in stationary and mobile weather stations, more still needs to be done to analyze the collected data. For example, such data can be used to estimate the optimal amount of deicing materials to be applied.

1.4. Other Recommendations

ID	Others	Recommendations
23	• Collaboration among DOTs	State DOTs in the New England region face many similar issues that are unique to this region (e.g., winter maintenance). It is strongly recommended that leaders from their TSMO divisions get together regularly to share best practices, experience, and issues encountered. For procurement decisions (e.g., which probe vehicle dataset to purchase), working together will give New England state DOTs more bargaining power.
24	• Organizational changes	Since we are increasingly relying on data to make decisions, DOTs should have a central office to handle data related issues. Instead of hosting data scientists/analysts in different DOT divisions, having a central office is beneficial for workforce training, recruiting, and retaining. Employees in this data office can easily help and learn from each other, which is helpful for data modeling. The data office will be similar to the IT department. Every DOT division can have some IT experts. However, it makes more sense to have a central IT department. Almost every DOT division depends on data and needs to collect, analyze, and store data. Having an office of data analytics will allow things to be done more efficiently and professionally (in terms of data safety, retention, sharing, etc.). With a holistic view of all the DOT data assets and how they are being utilized, it would be easier to develop data sharing, retention, privacy, and security policies. This central office can discuss the data retention needs and sharing policies with individual DOT divisions.

ID	Others	Recommendations
25	• Data storage and sharing among different DOT divisions	Most DOTs use both third-party cloud services and in-house servers for data storage. Most states have their own formal and informal record retention policies that apply to the collected data. DOTs are recommended to move their data to the cloud when applicable, which will make it easy to share data and help to ensure data safety, security, privacy, and integrity. More work needs to be done to promote and facilitate data sharing among different divisions of DOTs and different agencies (e.g., Transit vs. Highway; Turnpike vs. TSMO). Having a central Data Office may help to facilitate data sharing.
26	• Workforce	Many DOTs are creating data scientist/analyst positions and they are encouraged to continue doing this as needed. Although DOTs can always outsource the data analytics work to private companies, it is important for DOTs to understand what is being done by private companies.
27	• Personalized TIS with more dynamic and precise traffic information	A major part of TSMO is TIS. In the future, personalized data sharing with travelers would be important (e.g., sharing traffic signal timing data with connected vehicles, Alexa type of system instead of 511 phone system, recommender system that provides personalized traffic information based on a traveler's location and trip history). Google maps to some extents are doing this. With detailed and comprehensive (e.g., Transit, work zone) information, DOTs should explore what roles public agencies can play in future TIS. For example, can DOTs develop an App to share information not readily available on Google maps (e.g., scheduled work zones) with travelers in this region? Such an App can also collect travelers' mobile device GPS information (when it is within the boundary of state highways) for estimating travel time and detecting incidents. Such information will not be used for any commercial purpose unlike Google maps. Connected and Automated Vehicles (CAV) will generate a lot of data that can be used for TSMO applications. On the other hand, CAV will need precise traffic data for making safe, efficient, and eco-friendly driving decisions. In the future, variable message signs most likely will be phased out. Instead, DOTs need to provide traffic information in digital formats that can be easily and precisely interpreted by CAV.

ID	Others	Recommendations
		The traffic information will be much more detailed than what is displayed on a variable message sign today and can include information such as which lane is closed, taper length, distance to lane closure point, average left-turn phase duration, average queue length, etc.
28	• Drone as a data collection platform	Drones have been widely used by many DOTs for infrastructure inspection and providing situational awareness. DOTs are encouraged to investigate the potential of AI + drones (e.g., drone-in-a-box solution) for post-disaster roadway condition assessment.

2. RECOMMENDED PROJECT IDEAS FOR PHASE II

We proposed three topics in the original proposal: (1) providing an objective evaluation of the accuracies of emerging/non-traditional datasets such as Waze, TomTom, and INRIX using DOT data; (2) applying Artificial Intelligence (AI) to integrate the best emerging dataset and other datasets to predict traffic; and (3) applying the best emerging dataset for safety modeling.

In this report, we include three new ideas: (4) Evaluating the detection and tracking capabilities of advanced sensors such as LiDAR, Radar, and thermal cameras; (5) Evaluating the potential of connected vehicle data for applications such as queue detection/warning, incident detection, modeling driver behavior on horizontal curves; and (6) Weather station data modeling.

2.1. Topic 1 – Probe data validation

We plan to utilize existing DOT detectors (e.g., loop detector, Bluetooth, GoTime) to evaluate TomTom or INRIX data, to find out how reliable they are under different traffic and weather conditions. A focus is to find out how reliable such probe data is in rural areas where the penetration rates could be low. We believe an objective and thorough evaluation of third-party data vendors' products is important and will help state DOTs make more informed decisions. After validating the data, we will pick the best one and explore its possible applications in traffic operations. Specifically, we plan to identify signalized intersections for retiming using artificial intelligence methods. It is recommended by the Institute of Transportation Engineers (ITE) that traffic signal timing plans should be updated at least every 3 years. Due to resource constraint, many transportation agencies exceed this recommended interval. The proposed research will provide a handy tool to help traffic engineers prioritize intersections (especially those not equipped with ATSPM) and retime the most congested ones first.

2.2. Topic 2 – Travel time prediction

Existing traveler information system (TIS) are mostly based on measured (a few minutes ago) information, not predicted data. Being able to accurately predict future travel time (e.g., in the next 30 minutes) or the impacts of an incident (beyond just detecting an incident) by fusing data from different sources (e.g., INRIX, upstream detectors) can be very useful for TSMO. The results can be used for route guidance, estimating incident impacts (e.g., total queue length, recovery time), etc. We will consider the latest Graph Neural Network models for this topic.

2.3. Topic 3 – Safety modeling using Probe data

Traditional crash count modeling relies on AADT. Although each state DOT does maintain an AADT database to meet the HPMS requirement, the AADTs for many road segments (particularly minor roads) are often estimated and are unreliable. Another issue with AADT is that it does not reflect the variation in traffic volumes throughout a day and over different seasons.

This research will consider non-traditional datasets such as INRIX for modeling crash counts. It will consider speed limit, spatial and temporal speed variations throughout a day, time duration of each speed pattern, etc. to predict crash counts. Co-PI Ge has developed a novel approach combining graph neural embedding with probabilistic graphical models to give interpretable/explainable predictions of critical events such as crashes based on streaming spatial-temporal data. If this new modeling approach can accurately predict crash counts, it will help to address a major problem faced by traffic safety engineers by eliminating the need to rely heavily on AADT data. Since the spatial and temporal variations of variables are considered, this new modeling approach is anticipated to generate more accurate crash count predictions, which will help state DOTs accurately identify high-risk locations and dangerous traffic conditions and make better safety improvement and investment decisions (e.g., warn drivers of risky traffic conditions using variable message signs).

2.4. Topic 4 – Vehicle detection and tracking using AI + Edge computing powered sensors

Many DOTs have been approached by vendors with ideas to use RGB cameras for detecting and tracking vehicles. This research will investigate the limitations of existing technologies and explore the potential of new sensors such as LiDAR, high-resolution Radar, and thermal cameras (see a sample in Figure 1). These new sensors are not affected by light conditions and work well during both day and night. We also plan to invite vendors to field test their solutions (at vendors' costs) under the same conditions so that DOTs can know the pros and cons of different products clearly.

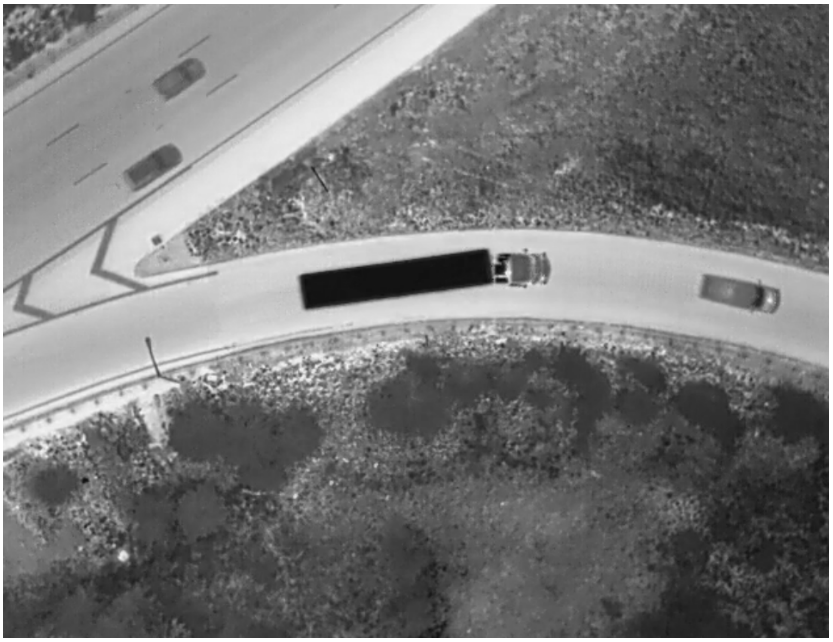

Figure 1. A Top-down View Captured by Thermal Camera at 400 ft in Andover MA

2.5. Topic 5 – Evaluating the potential of connected vehicles

This topic aims to explore the potential of connected vehicles data. We plan to use the USDOT Connected Vehicle pilot projects data from Wyoming or Tampa. We plan to study how effective such data may be used to detect queues and incidents and analyze driver behavior on horizontal curves. For example, we may find out how drivers decelerate when approaching a horizontal curve. Also, it will allow us to quantify the impacts of speed limit or advisory signs, time of day, vehicle type, etc. on driver behavior. The findings can help DOTs better set up traffic signs and design curves.

Although the USDOT Connected Vehicle pilot data is not for the New England region, the developed methods can be used in our region once we have local connected vehicle data available. Depending on the cost, we may also consider Wejo and Otonomo data. So that we can study some local issues.

2.6. Topic 6 – Weather data modeling

Many DOTs have stationary and mobile weather stations. The data from stationary and mobile weather stations can be correlated with deicing material application data to identify the optimal amount deicing materials that should be applied. The findings here can have significant environmental and safety benefits. Too much deicing materials will increase the negative environmental impacts, while inadequate deicing materials will increase safety risk.