

NCHRP Research Report 1071: Application of Big Data Approaches for Traffic Incident Management

Appendices

NCHRP Project 03-138

The National Cooperative Highway Research Program (NCHRP) is sponsored by the individual state departments of transportation of the American Association of State Highway and Transportation Officials. NCHRP is administered by the Transportation Research Board (TRB), part of the National Academies of Sciences, Engineering, and Medicine, under a cooperative agreement with the Federal Highway Administration (FHWA). Any opinions and conclusions expressed or implied in resulting research products are those of the individuals and organizations who performed the research and are not necessarily those of TRB; the National Academies of Sciences, Engineering, and Medicine; the FHWA; or NCHRP sponsors.

CONTENTS

Appendix A: Traffic Crash Data Detailed Assessment Outcomes	1
Appendix B: ATMS/Integrated ATMS-CAD Data Detailed Assessment Outcomes	4
Appendix C: CAD Data Detailed Assessment Outcomes	6
Appendix D: SSP Data Detailed Assessment Outcomes	9
Appendix E: Free Navigation App Data Detailed Assessment Outcomes	15
Appendix F: DOT ITS Fixed Sensor Data Detailed Assessment Outcomes	21
Appendix G: Vehicle Probe Data Detailed Assessment Outcomes	23
Appendix H: Roadway Inventory Data Detailed Assessment Outcomes	31
Appendix I: LRS Data Detailed Assessment Outcomes	33

Appendix J: Third-Party Road Network API Detailed Assessment Outcomes	35
Appendix K: SharedStreets Referencing System/OpenStreetMap Detailed Assessment Outcomes	36
Appendix L: ARNOLD Detailed Assessment Outcomes	38
Appendix M: MADIS Data Detailed Assessment Outcomes	41
Appendix N: Road Weather/Weather Data Environment (WxDE) Data Detailed Assessment Outcomes	45
Appendix O: Third-Party Weather API Data Detailed Assessment Outcomes	50
Appendix P: Third-Party CV Data Detailed Assessment Outcomes	53

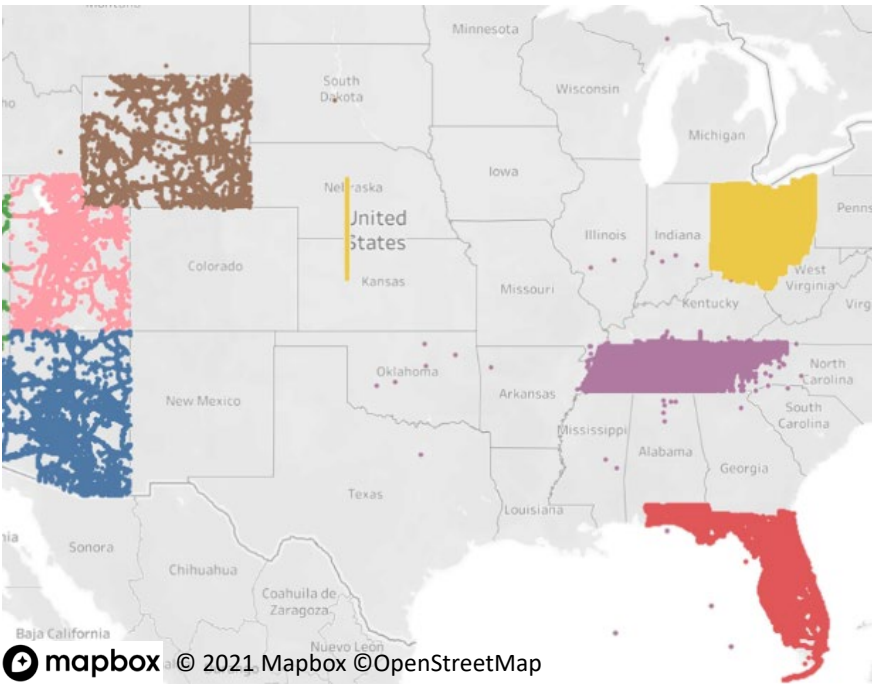
APPENDIX A

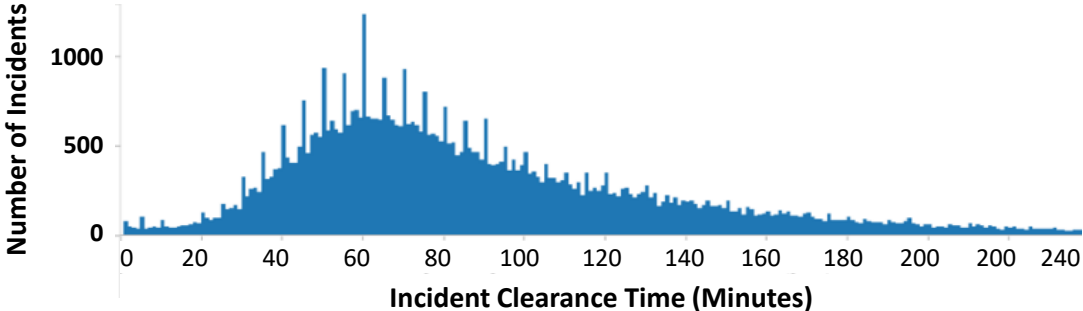
Traffic Crash Data Detailed Assessment Outcomes

Table 25 details the overall assessment of crash data based on data from nine states.

Table 1. Traffic Crash Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>Crash reports are required in all states if the crash involves an injury or a fatality. In cases of property damage only, state limits vary on what constitutes a reportable crash (e.g., over \$1500 in property damage). Therefore, state crash report datasets include most crashes; however, minor crashes may not be reported. There are many data elements and attributes (>1000 on most crash report forms), and depending on the state and data element, all may not be mandatory. Therefore, some data elements/attributes may be available, and others may not, which can make it difficult to properly analyze and combine crash data across states. For NCHRP Project 03-138 and the selected use cases, data on secondary crashes and roadway/incident clearance times were specifically needed. While roadway clearance time (RCT) and secondary crashes are both data elements in the 5th edition of the MMUCC (2017), many states do not have them on the crash report. Even if they do, in some cases they are not mandatory (e.g., Tennessee), which would result in incomplete data for the use cases.</p>
Timeliness	<p>For some states, crash data can be made available within a few days, while at the other extreme, some states do not make crash data available one or more years after the year of interest. Some datasets can also be augmented later if they are initially missing information because of pending legal action, and the timeline for when this missing information will be added is unknown.</p>
Consistency	<p>The MMUCC is a voluntary data collection guideline developed cooperatively by the National Highway Traffic Safety Administration (NHTSA) and the Governors Highway Safety Association (GHSA). The original guideline was developed in 1998, and the MMUCC is now in its 5th edition (published in 2017). The MMUCC 5th Edition includes 115 data elements. The goal of the MMUCC is to drive consistency in crash data collection across the states; however, because the MMUCC is voluntary, states can exercise control over what data elements are included on their crash forms and how these data are collected. And while some states strive to remain aligned with new editions of the MMUCC as a “standard,” others use it as more of a guideline. As such, data are inconsistent across states and sometimes even within a state (e.g., different jurisdictions may use different versions of the crash report and/or may include supplemental data elements like the TIM performance measures).</p> <p>In addition, despite instruction manuals and training, how data elements on the crash report are interpreted can vary widely from one officer to another. This has been shown by examining the use of various data elements associated with responder struck-by crashes, from implicit data elements (e.g., “working in trafficway – incident response”) to more explicit data elements (e.g., “was a responder hit”). In most cases, despite the data element used, a review of the crash report narrative is required to determine if the crash involved a responder being struck by a vehicle. This variability limits the possibility of automation and incorporation of these data into big data analysis.</p> <p>Additional issues to be aware of:</p> <ul style="list-style-type: none"> • Commas within cells, free text fields, proper nouns, etc.

Assessment Criteria	Assessment
	<ul style="list-style-type: none"> • Within dataset discrepancies (human entered data, changing formats from year to year, varying levels of completeness between agencies, varying levels of training between officers, subjective responses) • Between dataset discrepancies (different fields, different datatypes between similar fields, different available responses, different collection processes)
Conformity	<p>Within a single state’s dataset, most data attributes conform to a given format, but across states these formats vary. This makes it difficult and time-consuming to convert data from multiple states into a common format for analysis.</p> <p>Crash data files are often stored in a CP1252 encoding and need to be re-encoded to UTF-8 for easier processing. Files are often in CSV or pipe-delimited formats and sometimes have formatting errors, such as extra quotation marks. These formatting issues must be corrected before various software tools are able to ingest and process these data. It helps to convert these data to a consistent, strongly typed schema so that it is easier to analyze the data across state lines, but in doing so some data points are left ignored and other data points are unavailable in some states. This reduces the usefulness of many data points when doing cross-state analyses. As crash data for a single state can include hundreds of thousands or millions of records, it is best to convert these data into formats (such as parquet) deployed to a cloud environment for easier querying and analysis. It is important to use strong typing for consistent querying capabilities across different states’ datasets.</p>
Accuracy	<p>Crash data are inherently limited in accuracy because they are manually recorded by humans. Manual data entry of any kind is prone to error.</p> <p>Figure 45 shows that there are crashes located outside the state lines for a few of the states, particularly for Ohio, Tennessee, and Florida. This is due to errors in the location coordinates for the crashes in the databases. What is not shown here are the crashes that were missing location data/coordinates (e.g., crash data from Colorado did not include location coordinates).</p>  <p style="text-align: center;">Figure 1. Lat/long Errors Seen in Crash Data</p>

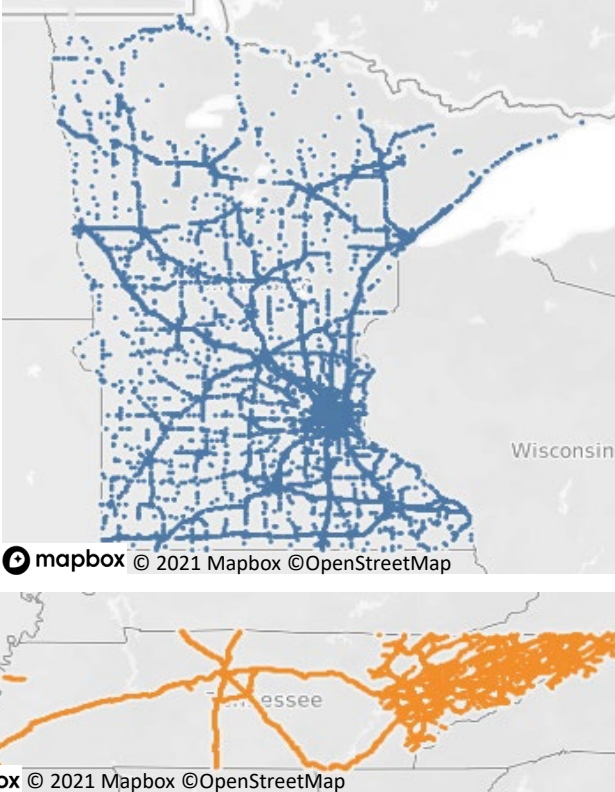
Assessment Criteria	Assessment
	<p data-bbox="337 289 1463 667">Given the nature of crashes, law enforcement officers have many things to tend to (e.g., keeping victims safe, calling in necessary resources, setting up temporary traffic control) in addition to completing crash reports. Additionally, they do not always have all the information needed to complete the form. For example, the time of the crash may be unknown, and the officer may have to estimate or rely on those involved in the crash. This leads to rounding artifacts and makes it so that certain attributes are often left unrecorded. Two areas that were identified as particularly problematic were: 1) confusing “no value” with “0”, and b) rounding. For example, if responders report a zero for the value of RCT where the roadway did not need to be cleared, and these zeros are not removed when calculating the average RCT, this will artificially decrease the average RCT. Furthermore, it has been observed that when humans enter in times, it appears as though they often round to the nearest 5-10 minutes, sometimes the nearest 30 min, creating artificial spikes in the data as seen in Figure 46.</p>  <p data-bbox="418 1041 1393 1073">Figure 2. Example of Probability Distribution Curve of ICTs from One State</p>
Integrability	<p data-bbox="337 1104 1451 1266">Crash data have good integrity within states, but these datasets can be difficult to analyze because of the variety of formats in which the data are stored across states. The attributes available in a particular state’s crash reports may vary year-over-year as well, which can complicate the data ingestion and unification process. It is important to store metadata about the data points that are available for each state and analysis period.</p>

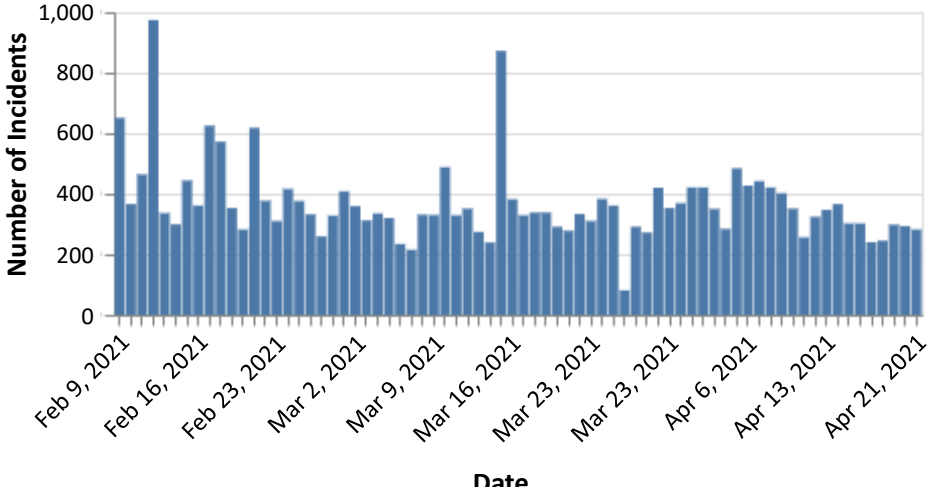
APPENDIX B

ATMS/Integrated ATMS-CAD Data Detailed Assessment Outcomes

Table 26 details the overall quality assessment of ATMS data based on data from these states.

Table 2. ATMS/Integrated ATMS-CAD Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>ATMS data typically have a high rate of completeness for the areas of coverage; however, some TMCs focus on managing traffic and incidents in urban areas. As such, incidents that occur outside of these urban areas may not be in the ATMS data. Some of these incidents may never be reported to the transportation agency, as they are handled by local or state police. Figure 47 shows the locations of ATMS incidents in the Minnesota and Tennessee datasets. While incidents in IRIS appear evenly distributed throughout the state of Minnesota, incidents in Tennessee’s Locate IM system are not evenly distributed and instead are numerous in the eastern part of Tennessee yet are only located along the interstate highways in the rest of the state.</p> <div style="text-align: center;">  <p>The figure consists of two maps. The top map shows the state of Minnesota with a network of blue dots representing ATMS incident locations. The dots are most concentrated in the eastern part of the state, particularly around the Minneapolis-St. Paul area, and are also distributed along major interstate highways. The bottom map shows the state of Tennessee with orange dots representing ATMS incident locations. Similar to Minnesota, the dots are heavily concentrated in the eastern part of the state and along major interstate highways, with fewer dots in the western and central regions.</p> </div> <p>Figure 3. Location of ATMS Incidents in Minnesota and Tennessee</p> <p>When ATMS systems are integrated with law enforcement CAD systems or public safety answering point (PSAP) systems, this can expand coverage beyond the normal urban coverage area of the TMC. A good example is in Minnesota. Figure 48 shows the number of records available each day (between February 9 and April 21, 2021) in MnDOT’s IRIS system, which is integrated with the Minnesota State Patrol CAD system.</p>

Assessment Criteria	Assessment
	 <p style="text-align: center;">Figure 4. Count of MnDOT Daily ATMS-CAD Records (Feb 9 to Apr 21, 2021)</p>
Timeliness	<p>ATMS data are often very timely, with new data becoming available as soon as they are captured by the ITS devices or entered in the ATMS by TMC operator/field staff. However, most agencies do not have real-time feeds set up to share their ATMS data and can only provide historical data in batches on a request-by-request basis. This makes it difficult to use the data in real-time analyses. The team was able to get a real-time data feed from IRIS. Utah DOT provided a data feed through XML web services (SOAP standard).</p>
Consistency	<p>There are many different vendors of ATMS software. Therefore, while there are examples of ATMS software provided by the same vendor that would provide some consistency between the states using these systems, overall, there is little consistency between data from states that use different systems. Even among states that use the same software, customizations in software can lead to inconsistencies in the data. One example is that some systems allow for free text entry into some fields, which can result in inconsistent spelling and multiple values that represent the same detail. One common misspelling occurs when dates are entered, where some operators may enter year-day-month instead of year-month-day.</p>
Conformity	<p>ATMS data vary widely from agency to agency and sometimes even TMC to TMC within an agency depending on the vendor of the software, so there is little conformity between systems. One item of note on conformity of ATMS data is how operators interact with the ATMS interface. If operators do not use the system as intended, entering all data (e.g., lane by lane closures and openings), less information is available about the incident and the corresponding TIM activities.</p>
Accuracy	<p>As with all human-entered data, accuracy can be a concern.</p>
Integrability	<p>ATMS data are often represented as a table or as a CSV file, where one row represents a collection of attributes about one incident. As such, integrating these data with other datasets should be straightforward for most cases. A broader challenge is associating data by latitude/longitude and date/time. Integrating data from one ATMS with another ATMS can be difficult, as there is no accepted unified standard for column names and data types, and</p>

Assessment Criteria	Assessment
	acceptable cell values may differ. When this occurs, some standardization process needs to be followed before full integration can occur.

APPENDIX C

CAD Data Detailed Assessment Outcomes

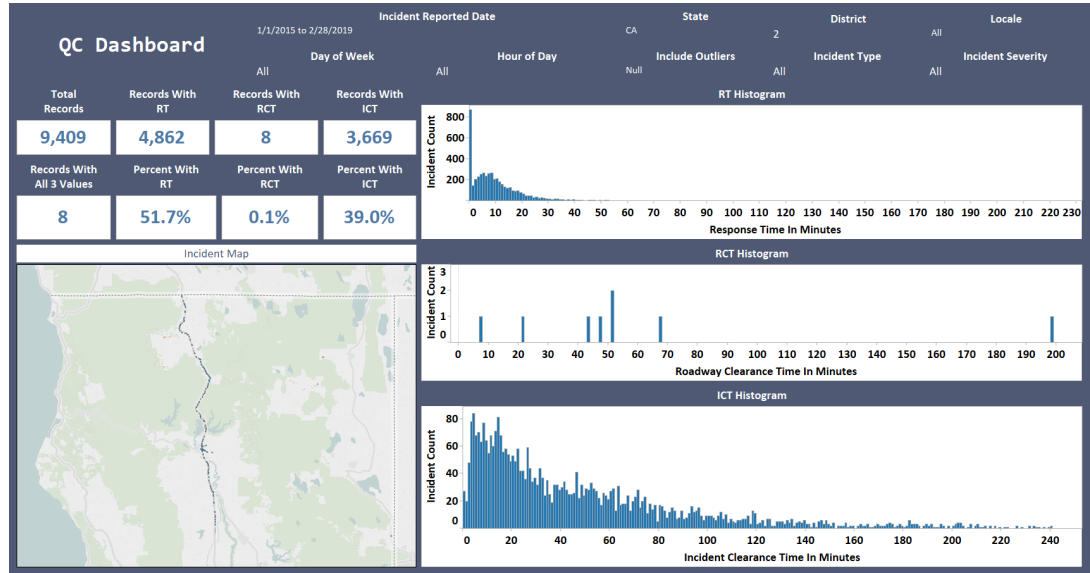
Table 27 details the results of the assessment of the CHP CAD data.

Table 3. CAD Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>While CAD systems keep a complete log of all agency responses, the complete dataset is rarely shared as some data are considered sensitive (e.g., PII, criminal activities) and cannot legally be shared. However, the events recorded in an individual CAD system represent only part of all traffic-related incidents.</p> <p>On the other hand, the completeness of CAD event status can be an issue when viewed from a TIM lens. Police, fire, and EMS may place different emphasis on the events that occur during an incident. Recording the TIM timestamps is an example. Following is a list of record completeness established from the publicly available CHP CAD data.</p> <p>Location data completeness:</p> <ul style="list-style-type: none"> • No missing latitude and longitude points • Less than 0.1 percent missing location description • Less than 0.1 percent missing roadway name <p>Event time data completeness:</p> <ul style="list-style-type: none"> • 19 percent missing valid incident date. • 19 percent missing time incident was reported. • 32.4 percent missing time incident was verified. • 61.3 percent missing time responder was dispatched. • 57.5 percent missing time responder arrival on scene • 99.6 percent missing time roadway was cleared. • 99.6 percent missing time lane(s) opened. • 69.6 percent missing time incident was cleared/last responder departed. • 16.4 percent missing number of responders • 51.8 percent missing injury type/code <p>In the analysis of the CHP CAD, the team also noticed that while some events were missing explicit information on timestamps, a quick natural language analysis of the text for the event status updates and the times when they were posted could be used to infer some of them. This is not ideal for real-time processing yet would allow more value to be extracted from CAD data.</p>
Timeliness	CAD systems are often based on real-time systems capable of delivering updates to users in seconds and automatically recording event status updates in near real-time.
Consistency	The data recorded by CAD systems are not always consistent. Even though most systems are capable of real-time, automated data collection and voice transcription, there are still areas where

Assessment Criteria	Assessment
----------------------------	-------------------

communication coverage is limited (e.g., rural areas), and responders are not able to report event status updates until they have better coverage. The CHP CAD data collected in Caltrans District 2, which is the most rural in the state, presents this kind of shortcomings.



Source: FHWA. Unpublished. Developed as part of FHWA Every Day Counts Round 4 (EDC-4) Using Data to Improve Traffic Incident Management Innovation.

Figure 5. Quality Analysis of TIM Performance Measures – CHP CAD data in Caltrans District 2

It can be noted in Figure 49 that the CHP CAD data for District 2 contain fewer incidents and that most of them are recorded along main roads.

Conformity	<p>CAD systems are often equipped with effective processes capable of validating and formatting location, time, and message content using natural language processing (NLP) to detect keywords and 10-codes in the event status updates. These data are typically standardized (e.g., Integrated Justice Information Systems Institute, or IJIS; Public Safety Technology standards, or IPSTSC; the National Fire Incident Reporting System, or NFIRS; the National Information Exchange Model, or NIEM; and the Global Justice XML Data Model, or GJXDM). These different standards, while overlapping, are inconsistent and are often customized to the culture and habits of agency personnel for maximum effectiveness. CAD data are manually collected, which can lead to natural variations from event status updates depending on the operators and responders, time of the day, experience, etc.</p> <p>As observed in the CHP CAD data, while the most common status updates are reported as 10-codes or 11-codes, less common event updates are reported using police lingo that varies from responder to responder. Before effectively parsing the text associated with event status updates, it is essential to learn/understand the lingo; using 10- and 11-codes is not sufficient.</p> <p>Locations expressed in CAD systems are typically uniformly expressed with recommended federal coordinate referencing systems, such as NAD 83 in the case of the CHP CAD data.</p>
-------------------	--

Assessment Criteria	Assessment
Accuracy	<p>CAD data are fairly accurate in terms of time and location. Time is captured automatically at the time of a call, and the near real-time processing by CAD systems allows the data to be processed and stored in less than a few seconds. CAD location data are often captured automatically using automated vehicle location (AVL) systems onboard responder vehicles, which are also fairly accurate. However, some incidents may have inaccurate times/locations. Timestamps can be inaccurate due to human error (e.g., officer forgetting to announce departure from the scene/being recorded several minutes or hours later). Locations can be affected by the environment (e.g., canyon effect in city centers with tall buildings can lead to imprecise GPS coordinates).</p>
Integrability	<p>CAD data are most often shared using data formats such as XML and JSON. XML, which while an aging format, is still dominant in most relational database systems including CAD systems. For such systems, exporting data in XML format is easy, reliable, and secure; however, most modern systems are built on more recent technologies, such as NoSQL, and have little to no support for XML data. This leads to the need for additional coding to integrate data with these modern systems. In the case of the CHP CAD XML data feed, the data are published in XML format and use a different XML standard than the strict XML document standard, most likely proprietary. To be parsed by common XML tools and loaded into more easily management formats like JSON that are usable by modern data analysis tools, the CHP CAD XML data require additional text processing to correct and make them adhere to the strict XML standard.</p> <p>CAD systems often use and publish data according to recommended federal coordinate referencing systems, such as the North American Datum of 1983 (NAD 83, a geocentric datum and geographic coordinate system based on the 1980 Geodetic Reference System ellipsoid (GRS80)), to geolocate event status updates. However, this is not ideal for integration with commercial vendor datasets, which use worldwide coordinate referencing systems such as WGS 84 (the standard for GPS) or Web Mercator (the de facto standard for web maps and online services). To integrate CAD data with these vendor datasets it requires the CAD data to be reprojected to fit the referencing systems of the other datasets prior to being integrated. This can be costly, especially in real-time systems. As the CHP CAD data were expressed using the NAD 83 coordinate referencing system, the team had to reproject the data using the WGS 84 so that the data could be merged with other datasets.</p>

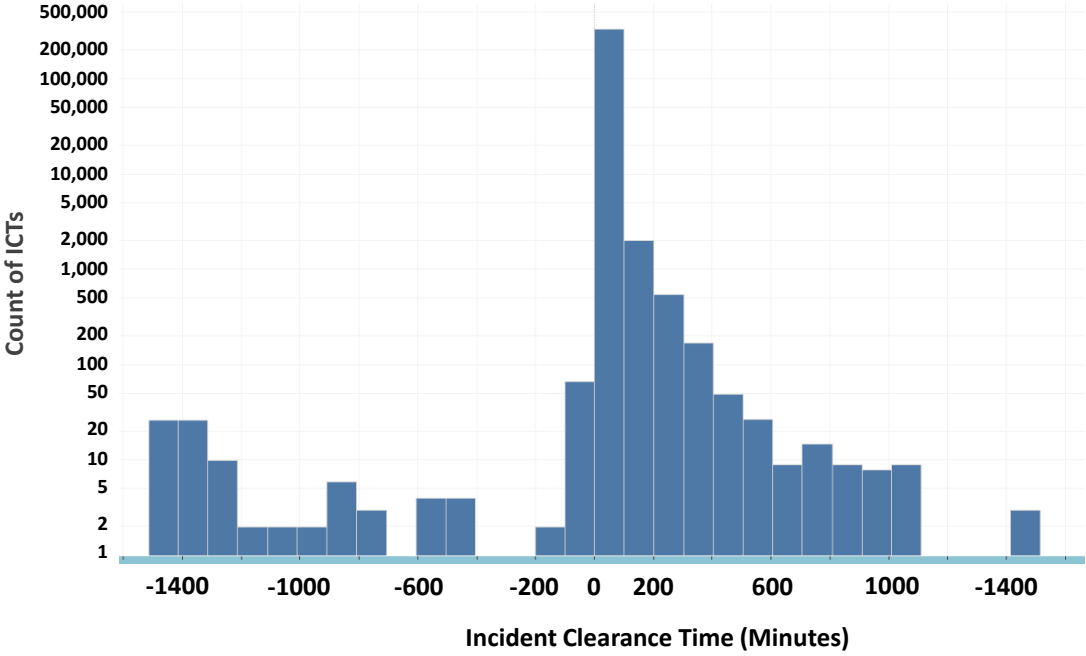
APPENDIX D

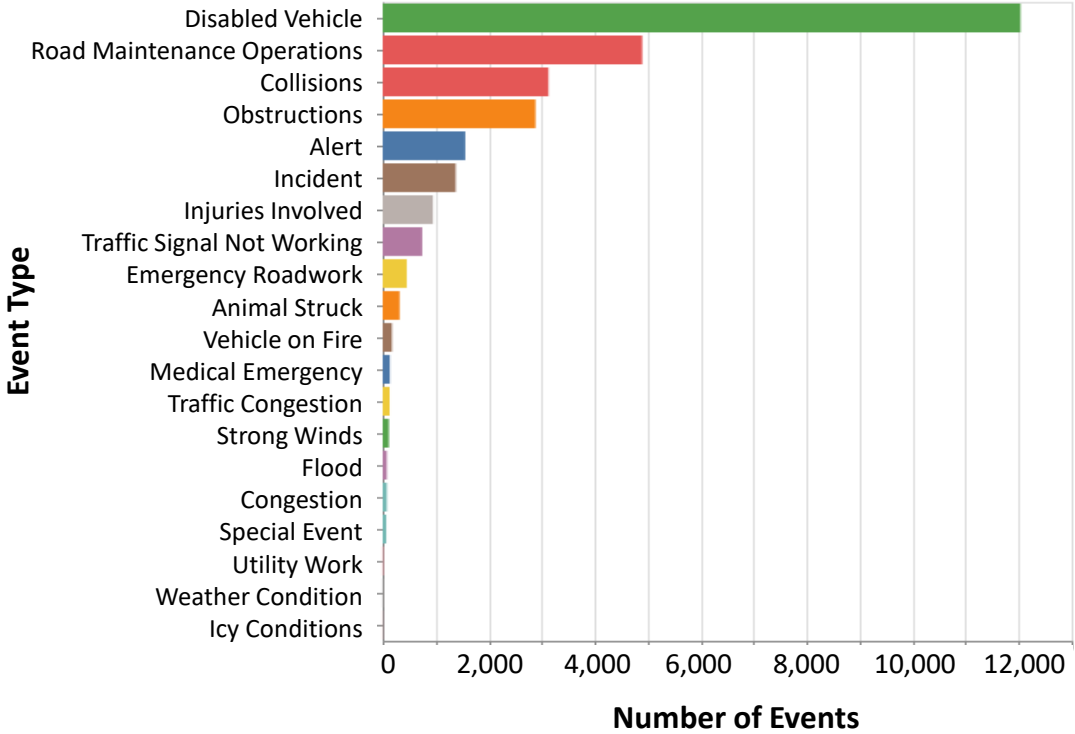
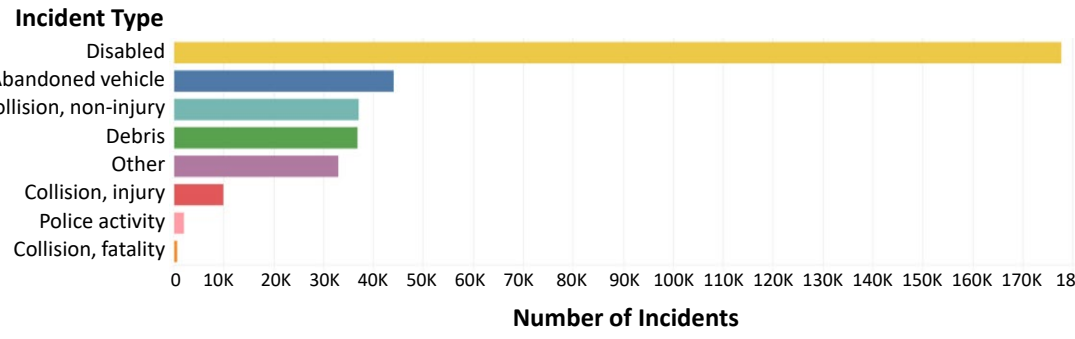
SSP Data Detailed Assessment Outcomes

Table 28 details the results of the assessment of the SSP data obtained by the team.

Table 4. SSP Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>The completeness of SSP data varies from agency to agency. To assess completeness, the team used the WITS dataset, which covers 5 years and sufficiently identified patterns of missing data. The WITS dataset contains 342,976 records composed of 26 variables, including latitude, longitude, time of arrival at scene, and time of departure. In this dataset about 13.3 percent of the variables are missing. Below is a list of the missing variables and their associated percentages:</p> <ul style="list-style-type: none"> • PrimaryIRDriverResponseIdentifier – 342,975 (> 99.9 percent) missing values. • ArrivalatSceneTime – 7,514 (2.2 percent) missing values. • AllLanesClearTime – 249,151 (72.6 percent) missing values. • SecondaryLaneClosure – 342,217 (99.8 percent) missing values • TotalNumberVehiclesInvolved – 50,228 (14.6 percent) missing values. • IRResponseComment – 70,675 (20.6 percent) missing values. • LandmarkDescription – 116,368 (33.9 percent) missing values. <p>For variables such as AllLanesClearTime and SecondaryLaneClosure, missing values are expected because they express information that is not relevant to every incident (although the rate of missing values is still high, especially for AllLanesClearTime). The other variables (ArrivalatSceneTime, PrimaryIRDriverResponseIdentifier, TotalNumberVehiclesInvolved, IRResponseComment, and LandmarkDescription) should not be missing. PrimaryIRDriverResponseIdentifier is missing for most records, which means that no information is available about which responder (driver) responded to the incident. The ArrivalatSceneTime is missing for only a small percentage of the records, which is acceptable. The TotalNumberVehiclesInvolved, IRResponseComment, and LandmarkDescription are missing for a non-negligible number of records. While the LandmarkDescription variable, which contains manually entered addresses or cross-street locations, can be easily completed using the latitude and longitude, the other two cannot, which impacts the ability to classify almost a third of the dataset in more detail.</p>
Timeliness	<p>The data reviewed were historical. The timestamps provided in the data are somewhat consistent with the other timestamps in the same record. They are within the same year, but some records show timestamps several weeks apart within the same record. The team visualized the frequency</p>

Assessment Criteria	Assessment
	<p>distribution of incident clearance time (ICT) show this anomaly (Figure 50). Negative ICTs on the left side of the distribution indicate erroneously collected timestamps for some records.</p>  <p>Figure 6. Frequency Distribution Plot of ICT from WITS (2015 and 2020)</p>
Consistency	<p>Agencies collect data on a variety of incident types, and these definitions vary from program to program. Washington’s incident response report includes fatal, injury, and non-injury collisions; blocking and non-blocking disabled vehicles; abandoned vehicles; and debris blocking traffic. Incidents/assists found in the CHART data are shown in Figure 51 and are categorized differently than Washington’s, which are shown in Figure 52. Most of the SSP data also contain information like RCT and ICT to track performance of the patrols.</p>

Assessment Criteria	Assessment																																																												
	 <p>Figure 7. Counts per Incident Type from the Maryland CHART SSP Dataset</p> <table border="1"> <thead> <tr> <th>Event Type</th> <th>Number of Events</th> </tr> </thead> <tbody> <tr><td>Disabled Vehicle</td><td>12,000</td></tr> <tr><td>Road Maintenance Operations</td><td>5,000</td></tr> <tr><td>Collisions</td><td>3,500</td></tr> <tr><td>Obstructions</td><td>3,000</td></tr> <tr><td>Alert</td><td>1,800</td></tr> <tr><td>Incident</td><td>1,500</td></tr> <tr><td>Injuries Involved</td><td>1,000</td></tr> <tr><td>Traffic Signal Not Working</td><td>800</td></tr> <tr><td>Emergency Roadwork</td><td>500</td></tr> <tr><td>Animal Struck</td><td>400</td></tr> <tr><td>Vehicle on Fire</td><td>300</td></tr> <tr><td>Medical Emergency</td><td>200</td></tr> <tr><td>Traffic Congestion</td><td>150</td></tr> <tr><td>Strong Winds</td><td>100</td></tr> <tr><td>Flood</td><td>50</td></tr> <tr><td>Congestion</td><td>50</td></tr> <tr><td>Special Event</td><td>50</td></tr> <tr><td>Utility Work</td><td>50</td></tr> <tr><td>Weather Condition</td><td>50</td></tr> <tr><td>Icy Conditions</td><td>50</td></tr> </tbody> </table>  <p>Figure 8. Count of WITS Incident Response by Type between 2015 and 2020</p> <table border="1"> <thead> <tr> <th>Incident Type</th> <th>Number of Incidents</th> </tr> </thead> <tbody> <tr><td>Disabled</td><td>175,000</td></tr> <tr><td>Abandoned vehicle</td><td>45,000</td></tr> <tr><td>Collision, non-injury</td><td>38,000</td></tr> <tr><td>Debris</td><td>38,000</td></tr> <tr><td>Other</td><td>35,000</td></tr> <tr><td>Collision, injury</td><td>12,000</td></tr> <tr><td>Police activity</td><td>5,000</td></tr> <tr><td>Collision, fatality</td><td>2,000</td></tr> </tbody> </table>	Event Type	Number of Events	Disabled Vehicle	12,000	Road Maintenance Operations	5,000	Collisions	3,500	Obstructions	3,000	Alert	1,800	Incident	1,500	Injuries Involved	1,000	Traffic Signal Not Working	800	Emergency Roadwork	500	Animal Struck	400	Vehicle on Fire	300	Medical Emergency	200	Traffic Congestion	150	Strong Winds	100	Flood	50	Congestion	50	Special Event	50	Utility Work	50	Weather Condition	50	Icy Conditions	50	Incident Type	Number of Incidents	Disabled	175,000	Abandoned vehicle	45,000	Collision, non-injury	38,000	Debris	38,000	Other	35,000	Collision, injury	12,000	Police activity	5,000	Collision, fatality	2,000
Event Type	Number of Events																																																												
Disabled Vehicle	12,000																																																												
Road Maintenance Operations	5,000																																																												
Collisions	3,500																																																												
Obstructions	3,000																																																												
Alert	1,800																																																												
Incident	1,500																																																												
Injuries Involved	1,000																																																												
Traffic Signal Not Working	800																																																												
Emergency Roadwork	500																																																												
Animal Struck	400																																																												
Vehicle on Fire	300																																																												
Medical Emergency	200																																																												
Traffic Congestion	150																																																												
Strong Winds	100																																																												
Flood	50																																																												
Congestion	50																																																												
Special Event	50																																																												
Utility Work	50																																																												
Weather Condition	50																																																												
Icy Conditions	50																																																												
Incident Type	Number of Incidents																																																												
Disabled	175,000																																																												
Abandoned vehicle	45,000																																																												
Collision, non-injury	38,000																																																												
Debris	38,000																																																												
Other	35,000																																																												
Collision, injury	12,000																																																												
Police activity	5,000																																																												
Collision, fatality	2,000																																																												

Assessment Criteria	Assessment																																							
	<p>There can be many inconsistencies in what data are recorded for each incident, and these data may be inconsistent with crash reports or other datasets. For example, UDOT's SSP data differ from Maryland CHART's SSP data in structure and content. UDOT's data are split across multiple files and contain various free-text descriptions of incidents, whereas CHART data have standardized incident types. The WITS dataset has a mix of standardized data types, including categorical data types for fields such as incident response type and primary lane closed, as well as free text for fields such as landmark description and incident response comments. The latter allows many variations of the same information, which creates unnecessary inconsistency in the dataset. Figure 93 shows an example of manually entered values in the WITS IR response comment field, leading to inconsistent spelling for the incident information.</p>	<table border="1"> <thead> <tr> <th colspan="2">IR Response Comment</th> </tr> </thead> <tbody> <tr> <td>Null</td> <td>70,711</td> </tr> <tr> <td>-</td> <td>1</td> </tr> <tr> <td>-</td> <td>2</td> </tr> <tr> <td>- Assist tow</td> <td>2</td> </tr> <tr> <td>- Assist tow</td> <td>1</td> </tr> <tr> <td>- Assit Tow</td> <td>1</td> </tr> <tr> <td>- Called off</td> <td>1</td> </tr> <tr> <td>- Calling for a tow</td> <td>1</td> </tr> <tr> <td>- Driver is lost</td> <td>1</td> </tr> <tr> <td>- Engine computer issue</td> <td>1</td> </tr> <tr> <td>- Flat tire no spare..</td> <td>1</td> </tr> <tr> <td>- Front Reg AKT4433..</td> <td>1</td> </tr> <tr> <td>- Has AAA tow assist com..</td> <td>1</td> </tr> <tr> <td>- Has Assist coming</td> <td>1</td> </tr> <tr> <td>- Has tow Assist coming</td> <td>1</td> </tr> <tr> <td>- Has Tow Assist coming</td> <td>1</td> </tr> <tr> <td>- Has Tow coming</td> <td>1</td> </tr> <tr> <td>- Has tow coming</td> <td>1</td> </tr> </tbody> </table> <p>Figure 9. Example WITS IR Response Comments</p>	IR Response Comment		Null	70,711	-	1	-	2	- Assist tow	2	- Assist tow	1	- Assit Tow	1	- Called off	1	- Calling for a tow	1	- Driver is lost	1	- Engine computer issue	1	- Flat tire no spare..	1	- Front Reg AKT4433..	1	- Has AAA tow assist com..	1	- Has Assist coming	1	- Has tow Assist coming	1	- Has Tow Assist coming	1	- Has Tow coming	1	- Has tow coming	1
IR Response Comment																																								
Null	70,711																																							
-	1																																							
-	2																																							
- Assist tow	2																																							
- Assist tow	1																																							
- Assit Tow	1																																							
- Called off	1																																							
- Calling for a tow	1																																							
- Driver is lost	1																																							
- Engine computer issue	1																																							
- Flat tire no spare..	1																																							
- Front Reg AKT4433..	1																																							
- Has AAA tow assist com..	1																																							
- Has Assist coming	1																																							
- Has tow Assist coming	1																																							
- Has Tow Assist coming	1																																							
- Has Tow coming	1																																							
- Has tow coming	1																																							
Conformity	<p>SSP data should conform to best practices for data typing and formatting, such as standard timestamps and the WGS84 referential system with nine decimal precision; however, there are cases where data do not conform to industry best practices. For example, none of the SSP data reviewed has a specified time zone in their schema despite using timestamp standards correctly. This is often considered unnecessary due to assumptions that the data will only be used within the state. Also, some data, such as addresses and even timestamps, are entered by hand without any guidance or checks at the interface or database level (e.g., WITS landmark data field, Utah incident timestamps). This leads to numerous inconsistencies and inaccuracy in the data, rendering the dataset much more difficult to analyze. In some cases, SSP data are exported in a format that is unreadable by any software. For example, UDOT's incidents table was exported as a tab-delimited CSV but is missing quotes around fields containing tabs, which made it impossible to parse, short of going through it line by line and guessing the end of each field for each record and removing tab or adding quotes.</p>																																							

Assessment Criteria	Assessment																																												
Accuracy	<p data-bbox="391 302 683 873"> Island cest on ramp l/s Island Crest J/N Capital J/S 80TH J/S 185TH JAMES C/D James St (Mainline) JAMES ST ENTRANCE .. just east of &\$ crossover just north of Forest ST Lake City Way Machias Road maker marker MERCER mile marker MP-197 N OF 4 th N OF 15th NW N of 85th </p> <p data-bbox="391 884 651 982"> Figure 10. WITS Landmark Description Sample </p> <div data-bbox="383 1003 1430 1608"> <table border="1" data-bbox="383 1003 1430 1608"> <caption>Data for Figure 11: WITS Total Number of Vehicles Involved Distribution</caption> <thead> <tr> <th>Number of Vehicles Involved per Incident</th> <th>Count of Total Number of Vehicles Involved</th> </tr> </thead> <tbody> <tr><td>0</td><td>1</td></tr> <tr><td>50</td><td>400,000</td></tr> <tr><td>100</td><td>60,000</td></tr> <tr><td>150</td><td>15,000</td></tr> <tr><td>200</td><td>12,000</td></tr> <tr><td>250</td><td>4,000</td></tr> <tr><td>300</td><td>5,000</td></tr> <tr><td>350</td><td>6,000</td></tr> <tr><td>400</td><td>5,000</td></tr> <tr><td>450</td><td>6,000</td></tr> <tr><td>500</td><td>4,000</td></tr> <tr><td>550</td><td>2,000</td></tr> <tr><td>600</td><td>3,000</td></tr> <tr><td>650</td><td>5,000</td></tr> <tr><td>700</td><td>3,000</td></tr> <tr><td>750</td><td>2,000</td></tr> <tr><td>800</td><td>2,000</td></tr> <tr><td>850</td><td>4,000</td></tr> <tr><td>900</td><td>3,000</td></tr> <tr><td>950</td><td>3,000</td></tr> <tr><td>1000</td><td>3,000</td></tr> </tbody> </table> </div> <p data-bbox="500 1633 1336 1667"> Figure 11. WITS Total Number of Vehicles Involved Distribution </p> <p data-bbox="708 289 1463 961"> The SSP data are generally accurate in terms of time and geolocation, but they lose accuracy when manual or non-standardized data entries are considered. The Maryland CHART dataset has latitude/longitude with six decimal places, which is accurate to ± 11 centimeters. UDOT’s location data are encoded in an unknown coordinate referencing system. WITS latitude and longitude are accurate and standardized even accurately showing WITS responses in Oregon. They are also all matched to a linear referencing measure and a unique state route ID. Accuracy drops when reviewing manually entered data fields, such as “Landmark Description” where no standardized way to express address and location is enforced. Figure 54 shows a sample of the “Landmark Description” field and the variety of ways location is entered. Some coded (not free text) fields in the WITS dataset also have accuracy issues, such as the total number of vehicles involved in the incident, which is not bounded on entry and results in an excessively large number of vehicles involved entered in several hundred incident records. Figure 55 shows the distribution of number of vehicles involved in incidents, ranging from 50 to 1000 vehicles. </p>	Number of Vehicles Involved per Incident	Count of Total Number of Vehicles Involved	0	1	50	400,000	100	60,000	150	15,000	200	12,000	250	4,000	300	5,000	350	6,000	400	5,000	450	6,000	500	4,000	550	2,000	600	3,000	650	5,000	700	3,000	750	2,000	800	2,000	850	4,000	900	3,000	950	3,000	1000	3,000
Number of Vehicles Involved per Incident	Count of Total Number of Vehicles Involved																																												
0	1																																												
50	400,000																																												
100	60,000																																												
150	15,000																																												
200	12,000																																												
250	4,000																																												
300	5,000																																												
350	6,000																																												
400	5,000																																												
450	6,000																																												
500	4,000																																												
550	2,000																																												
600	3,000																																												
650	5,000																																												
700	3,000																																												
750	2,000																																												
800	2,000																																												
850	4,000																																												
900	3,000																																												
950	3,000																																												
1000	3,000																																												
Integrity	<p data-bbox="370 1696 1463 1900"> SSP data are typically shared as Excel or CSV files, which are good formats to parse and ingest, but they lack metadata to increase data clarity (e.g., the referential system used by the latitude and longitude in the table). This means that additional documentation needs to be acquired to correctly parse and ingest the data. Timestamps are often missing time zone information, which is not an issue in most states but causes problems with states like Florida that have more than one time zone and Arizona, which does not participate in Daylight Savings Time. In these cases, some additional data </p>																																												

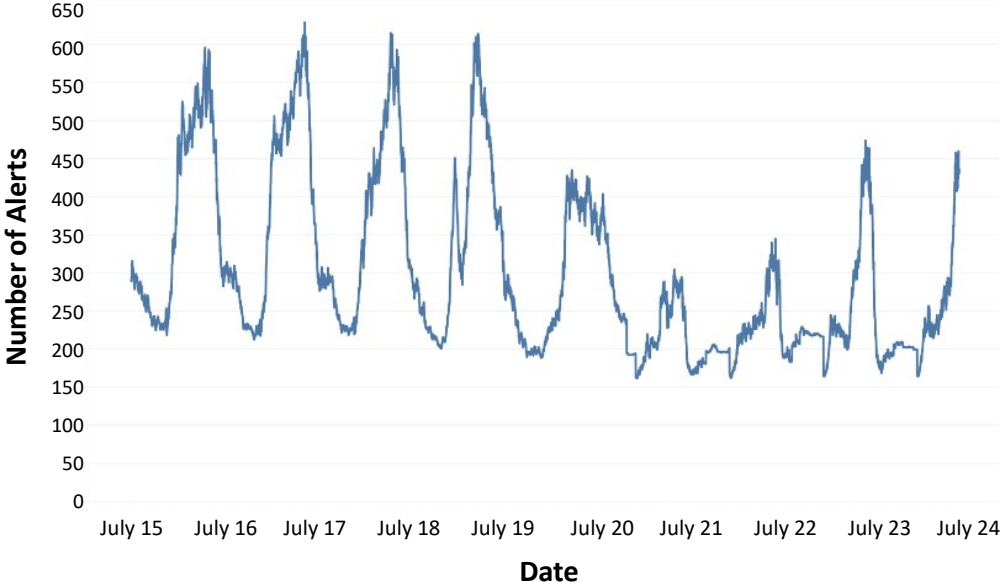
Assessment Criteria	Assessment
	<p>processing is required to identify the correct timestamp time zone. Regarding the categorization or classification of incident response, while there is some overlap across states, each state creates its own taxonomy, often loosely based on Traffic Management Data Dictionary (TMDD) but not entirely compliant. This makes it difficult to integrate data from multiple states as a mapping of categories needs to be established prior to merging the datasets. This can be difficult since the incidents in SSP datasets may not have identifying information that enables easy association with other datasets (such as crash reports) since the SSP data originate from entirely different systems.</p>

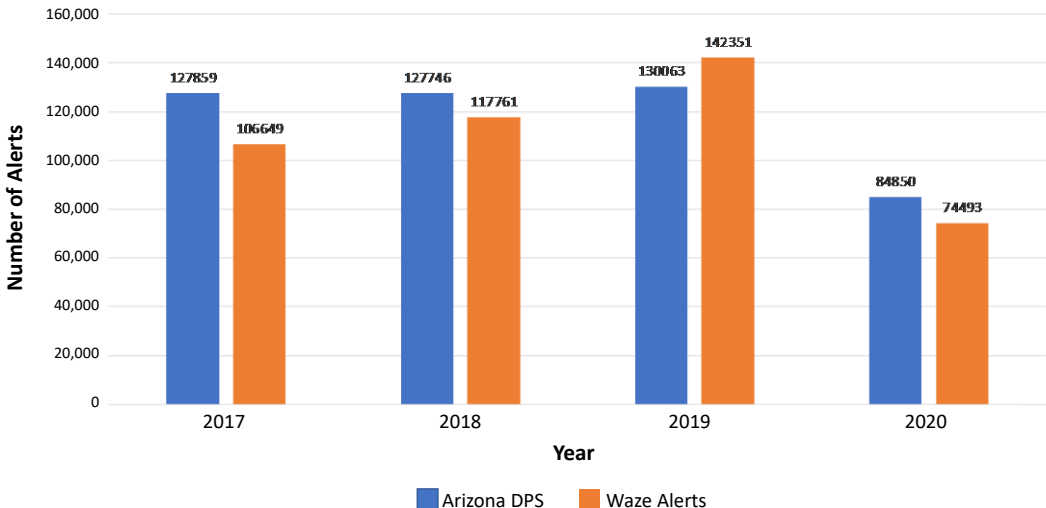
APPENDIX E

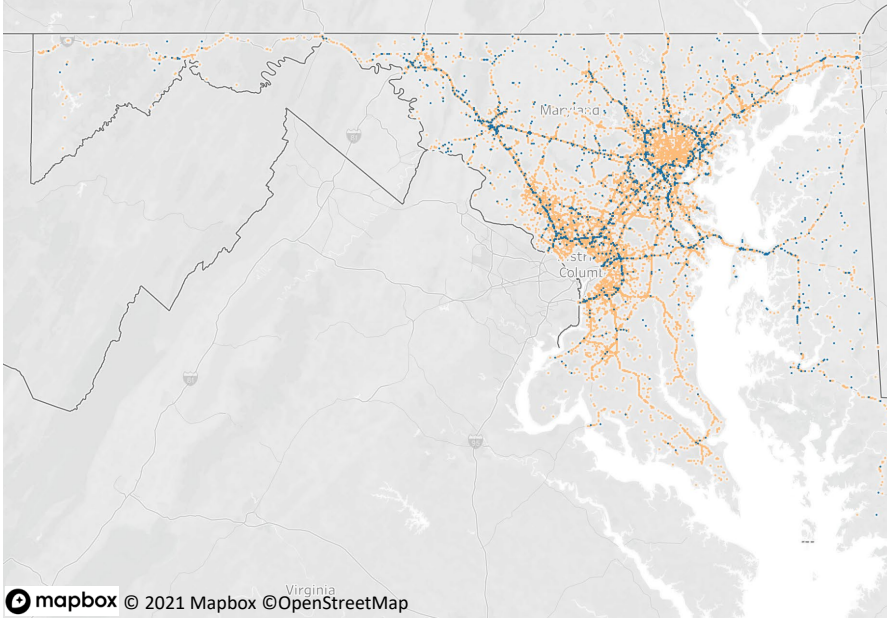
Free Navigation App Data Detailed Assessment Outcomes

Table 29 details the results of the assessment of the free navigation app data obtained by the team.

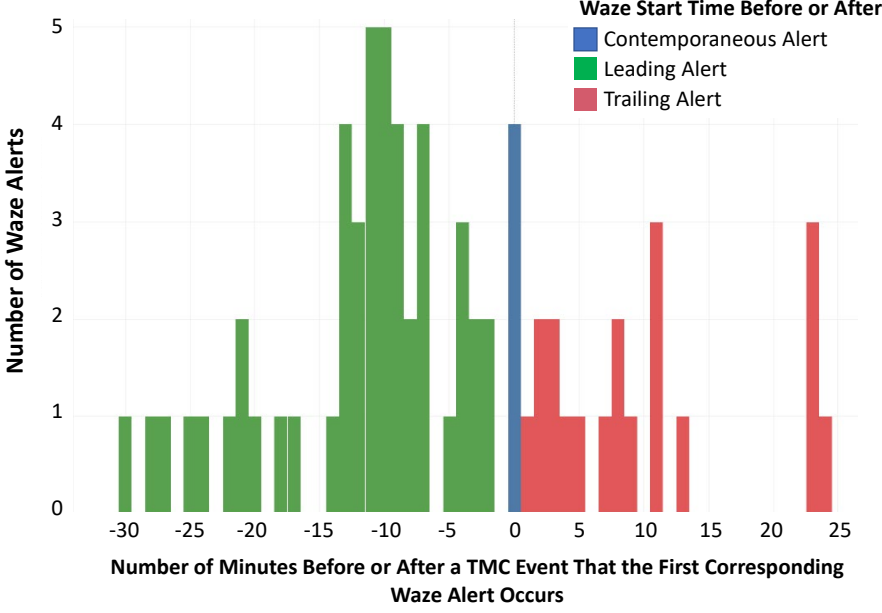
Table 5. Free Navigation App Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>Since the navigation app data assessed are crowdsourced from the public, there may be discrepancies in the number of users reporting in different regions, which results in various levels of completeness. The navigation app provider does not provide any metrics on how many users actively report on a specific roadway segment. Overall, the navigation app data captures more traffic-related events than traditional TMCs.</p> <p>The real-time navigation app data collected in Minnesota, Massachusetts, Maryland, Utah, and historical data archives did not show any apparent gaps. Figure 56 shows the number of alerts published by the navigation app provider in the entire state of Massachusetts between July 15 and July 24, 2020. The number of alerts peaked toward the evening commute and dropped overnight. The alerts also dropped starting July 20, 2020, following the Massachusetts COVID stay-at-home order, and the number of alerts published afterward was significantly less but still showed a peak during the evening commute.</p>  <p>Figure 12. Navigation App Alerts Per Hour in Massachusetts (July 15-24, 2020)</p> <p>Figure 57 shows a histogram comparing the number of crashes reported by the Arizona Department of Public Safety (AZDPS) and the number of “accidents” reported by navigation app provider by year for the years 2017 and 2020 across the state of Arizona. The number of navigation app alerts is within about 10 percent of the AZDPS crash count, which seems acceptable when considering the crowdsourced nature of the navigation app data.</p> <p>At the data fields level, navigation app alerts were not always complete; some alerts were missing some data fields. For example, city names may be missing when an incident occurs on a highway</p>

Assessment Criteria	Assessment															
	<p>between towns or when alerts were submitted using the navigation app touchless option (hand waving), which does not collect alert subtype. Following are the percentages of missing data found in the US DOT national navigation app data archive dataset:</p> <ul style="list-style-type: none"> • Street name missing in 5.4 percent of all records. • Alert subtype missing in 5.3 percent of all records. • City name missing in 49.4 percent of all records. • Number of thumbs up missing in 25.9 percent of all records • Road type missing in 5.6 percent of all records. <p>Location data fields, such as latitude, longitude, and heading were always present and usually within expected bounds. Latitude and longitude were sometimes inaccurate due to the use of a geohash and the location approximation it creates by the navigation app. Alert timestamp, alert update timestamp, alert reliability, and alert confidence were also always present in the data.</p>  <table border="1" data-bbox="389 724 1429 1228"> <thead> <tr> <th>Year</th> <th>Arizona DPS</th> <th>Waze Alerts</th> </tr> </thead> <tbody> <tr> <td>2017</td> <td>127,859</td> <td>106,649</td> </tr> <tr> <td>2018</td> <td>127,746</td> <td>117,761</td> </tr> <tr> <td>2019</td> <td>130,063</td> <td>142,351</td> </tr> <tr> <td>2020</td> <td>84,850</td> <td>74,493</td> </tr> </tbody> </table> <p>Figure 13. Comparison between the number of crashes recorded by the AZ DPS and the number of accidents reported by the third-party in 2017, 2018, 2019 and 2020</p> <p>Figure 58 shows the locations of navigation app reports (orange) and incidents recorded by Maryland CHART (blue) in 2020. It is easy to observe that navigation app users provide insights on many more incidents than CHART with existing methods of incident detection.</p>	Year	Arizona DPS	Waze Alerts	2017	127,859	106,649	2018	127,746	117,761	2019	130,063	142,351	2020	84,850	74,493
Year	Arizona DPS	Waze Alerts														
2017	127,859	106,649														
2018	127,746	117,761														
2019	130,063	142,351														
2020	84,850	74,493														

Assessment Criteria	Assessment
	 <p data-bbox="367 926 1446 999">Figure 14. Map displaying incidents recorded by navigation app (orange) and MD CHART (blue) in 2020</p>
Timeliness	<p data-bbox="344 1024 1463 1373">The navigation app data are provided in real-time through a GeoRSS¹ data feed, which is updated every minute. This refresh rate is acceptable for ingesting navigation app data into traditional data systems, but it is on the slow side when considering modern data systems, which are capable of ingesting and processing navigation app data coming in real-time through a web socket and then processing updates as soon as they are published. Despite this limitation, the navigation app data are timely; however, 99 percent of the timestamps for all navigation app alert updates are the same as the time the event started. This makes it impossible to know when the update was published. The only way to capture the timestamps of these updates is to do so each time the navigation app GeoRSS data feed changes. This time can be used as a proxy for the alert update time, but it is limited in precision by the GeoRSS data feed refresh rate.</p> <p data-bbox="344 1388 1463 1627">Figure 59 shows a histogram representing the count of navigation app alerts binned by minute that occurred 30 minutes before (green) and 30 minutes after (red) and within half a mile of a TMC recorded crash (blue) on US-50 in Maryland in March 2020. Figure 59 shows that using the recreated update timestamps, navigation app alert updates occurred as soon as 30 min before the TMC had knowledge of the crash, and the rate of alert updates started drastically increasing 15 min before the TMC knew of the crash. The team found equivalent results through a similar comparison with crashes in Massachusetts.</p>

¹ <https://www.ogc.org/standards/georss>

Assessment Criteria	Assessment
	 <p data-bbox="373 934 1437 997">Figure 15. Count of Alerts Occurring 30 Minutes Before and After a TMC Incident Record Time on US 50 in Maryland</p>
Consistency	<p data-bbox="344 1018 1461 1186">The navigation app data are simple and consistent, even with the errors (e.g., missing update timestamps) across all types of alerts. The data are consistent across the U.S. over several years with a slight increase in 2019 followed by a drop in 2020, like what can be observed in Figure 56 for Massachusetts. There are little unexpected variations (less populated areas, nighttime, holidays, etc.) in data quality or reporting frequency nationwide.</p>
Conformity	<p data-bbox="344 1207 1453 1312">The navigation app provider follows its own specification rather than publishing data using some other specification. The specification is simple and lacks the details found in other specifications. Roadway name and type are arbitrary and so are confidence and reliability indices.</p> <p data-bbox="344 1323 1437 1459">The navigation app provider has also adopted international standards for its location and timestamp data that are widely recognized and heavily used in modern data systems. The alert locations are expressed using decimal latitude and longitude in the WGS84 referential system, and timestamps are expressed using UNIX epoch time in milliseconds.</p>
Accuracy	<p data-bbox="344 1480 1461 1753">The accuracy of the navigation app data is another matter. The data are crowdsourced and human entered, which inherently creates errors (e.g., creating the wrong type of event, pointing the phone in the opposite direction of traffic when recording an alert associating it with the opposite traffic direction, recording a premature event, such as a disabled vehicle on shoulder alert when a vehicle is stopped to pick up a phone call). The navigation app alert locations are also inaccurate by design due to the distance traveled by the user after observing the event before reporting it. Thus, the alert locations are typically within a bounded distance from when they are reported often within a mile after the actual location of the event.</p> <p data-bbox="344 1764 1461 1858">This means that navigation app alerts should be carefully reviewed and assessed before being considered for action. This is also made difficult as update timestamps are inaccurate 99 percent of the time.</p>

Assessment Criteria	Assessment																												
	<p>The navigation app provider includes three separate measures to assess the trustworthiness of data: a reliability score, a confidence score, and the number of “thumbs-up” reports received from other users. The exact algorithms that the provider uses to calculate reliability and confidence are not publicly available, but the number of “thumbs-up” reports received is a clear and intuitive measurement that performs well to identify reliable reports.</p> <p>The distributions of the reliability and confidence scores (shown in Figure 60) indicate that accurate event reporting (levels 8 to 10) is scarce, as most navigation app alerts have a consistent confidence score of 0 and a consistent reliability score of 6. It can also be noted that some event confidence scores have a value of -1, which most likely indicates failure of the algorithm to calculate a confidence score for these events.</p> <div data-bbox="370 667 1409 1045"> <table border="1"> <caption>Estimated Data for Figure 16</caption> <thead> <tr> <th>Score</th> <th>Frequency (1e8)</th> </tr> </thead> <tbody> <tr><td>-1</td><td>0.02</td></tr> <tr><td>0</td><td>1.05</td></tr> <tr><td>1</td><td>0.10</td></tr> <tr><td>2</td><td>0.05</td></tr> <tr><td>3</td><td>0.02</td></tr> <tr><td>4</td><td>0.01</td></tr> <tr><td>5</td><td>0.05</td></tr> <tr><td>5</td><td>0.95</td></tr> <tr><td>6</td><td>8.5</td></tr> <tr><td>7</td><td>1.0</td></tr> <tr><td>8</td><td>0.5</td></tr> <tr><td>9</td><td>0.2</td></tr> <tr><td>10</td><td>1.2</td></tr> </tbody> </table> </div> <p>Figure 16. 2012-2017 Nationwide Navigation App Event Confidence Score (Left) and Reliability Score (Right) Distribution</p> <p>There are a limited number of navigation app alerts that contain erroneous timestamps, which are scattered between year 0001 and year 3713, but their frequency is so small that it can be considered as noise.</p>	Score	Frequency (1e8)	-1	0.02	0	1.05	1	0.10	2	0.05	3	0.02	4	0.01	5	0.05	5	0.95	6	8.5	7	1.0	8	0.5	9	0.2	10	1.2
Score	Frequency (1e8)																												
-1	0.02																												
0	1.05																												
1	0.10																												
2	0.05																												
3	0.02																												
4	0.01																												
5	0.05																												
5	0.95																												
6	8.5																												
7	1.0																												
8	0.5																												
9	0.2																												
10	1.2																												
Integrability	<p>Real-time navigation app event data are provided through a GeoRSS data feed, in either JSON or XML. Historical navigation app data are provided through access to a cloud database that contains the same data as the real-time data feed. By using JSON and XML, the real-time navigation app data can easily be integrated with traditional and modern data systems. The GeoRSS publishing process is limited but robust enough to support integration with systems designed to be near real-time. Furthermore, the Open Geospatial Consortium (OGC) retired the GeoRSS standard in September 2020, meaning it is becoming a legacy standard.</p> <p>On the contrary, the historical navigation app datasets, available through a cloud platform, are not easily integrated without adopting the platform and migrating part of the integration processes and data to that platform. While it is convenient to access such large datasets through data querying services, it represents a type of vendor lock by not allowing the data to be accessed through other platforms.</p> <p>The location data and bearing data provided in navigation app event messages also make it easy to associate events with agency road networks, even in real-time.</p> <p>Finally, the imprecise nature of the navigation app time and location data make it challenging to integrate it with other incident datasets, such as TMC or crash report data, as fuzzy matches resulting in</p>																												

Assessment Criteria	Assessment
	multiple matching candidates may be possible, and some rules or algorithms will need to be devised to select the correct matching event.

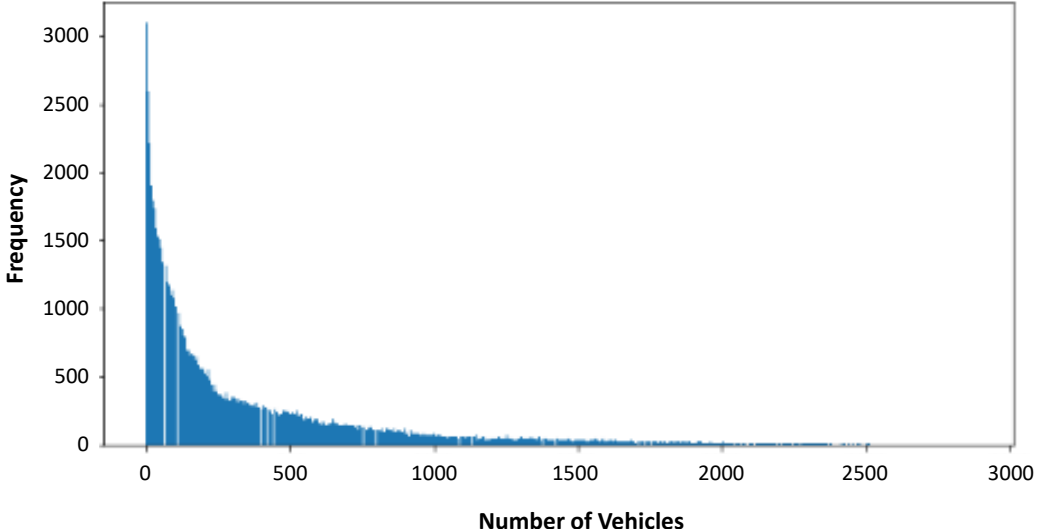
APPENDIX F

DOT ITS Fixed Sensor Data Detailed Assessment Outcomes

Table 30 details the results of the assessment of the DOT ITS fixed sensor data obtained by the team.

Table 6. DOT ITS Fixed Sensor Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>In the case of the data from California’s PeMS, there are raw data from March 2001 to February 2021, with some missing months for certain districts. Sensors fail from time to time for several reasons. There are stations with missing data on some days, as can be observed in Figure 61 for eight stations on January 1, 2021.</p> <p style="text-align: center;">Stations with Fewer than 750 Rows of Raw Data (January 1, 2021)</p>  <p style="text-align: center;">Figure 17. Missing data from shortly after noon to shortly after 7 p.m. from several sensors in the California PeMS data on January 1, 2021.</p> <p>Ohio’s speed sensor data are comprised of readings from only 158 sensor locations across the state. Of these locations, some are missing data, as displayed in Figure 62.</p>  <p style="text-align: center;">Figure 18. Missing data from Ohio's speed sensor data.</p>
Timeliness	<p>The Caltrans station data, provided by PeMS, are available in raw form for the previous day, which is timely enough for certain applications but not for applications that require real-time or near real-time data. The same thing can be assumed for other states implementing PeMS, such as Utah. The latency of Florida and Ohio’s data is unknown; the team did not have direct access to the systems to assess the timeliness with which the data are refreshed.</p>
Consistency	<p>For Caltrans, data are consistent across districts and times. Data are available in 5-minute, hourly, and daily roll-ups. Whether data are observed or estimated is indicated explicitly.</p>

Assessment Criteria	Assessment
	Ohio’s data are inconsistent across districts in terms of quantity of data, as shown above in Figure 62, but are otherwise consistent.
Conformity	<p>The data come in different formats depending on the type of dataset, but they conform to a given schema across years and districts for each dataset. For example, the California PeMS data have multiple averaged datasets (station 5-minute, station hour, station day) that all have the same data structure. PeMS also has station AADT and station raw datasets (among others) that each have their own schema and data definition sets. Sensor data from Ohio has a different schema and format. More specifically, Ohio’s data track the number of vehicles per speed range (e.g., 0-40 MPH, 40-45 MPH, 45-50 MPH, etc.), whereas the PeMS data track speed averages for the given time interval.</p>
Accuracy	<p>In the PeMS data, outliers have either been culled or imputed already, or the sensors are relatively accurate. No extreme outliers were found in the samples. Ohio’s data schema does not allow for outliers in speed, as the speed bins are predefined, and while there are records with high numbers of vehicles, the distribution does not indicate any extreme outliers (Figure 63).</p>  <p style="text-align: center;">Figure 19. Histogram of number of vehicles per 5-minute period in Ohio’s speed sensor data.</p>
Integrability	<p>The California sensor data are only available through the PeMS portal, and while terabytes of data are available, downloading and processing the data is difficult. Files must be batch-downloaded and moved to a data lake or other “big data” environment to be analyzed. Ohio’s sensor data must be requested from ODOT. Once obtained, the data can be integrated with other datasets via the latitude and longitude coordinates; however, as indicated above under “completeness,” the distance of the sensors from crashes/incidents will reduce the usefulness of the data.</p>

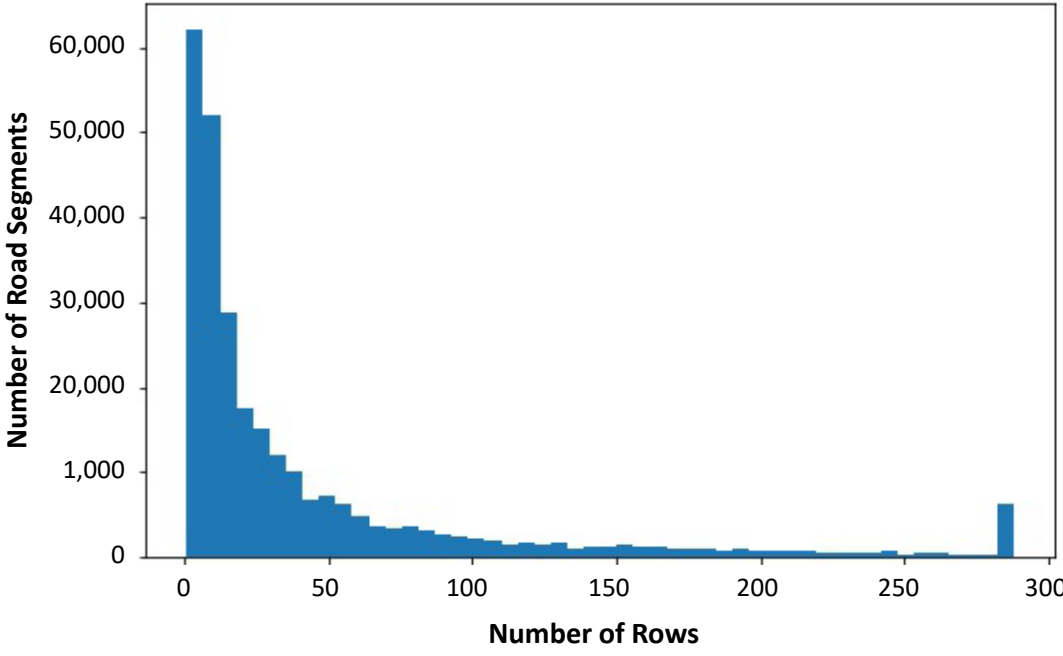
APPENDIX G

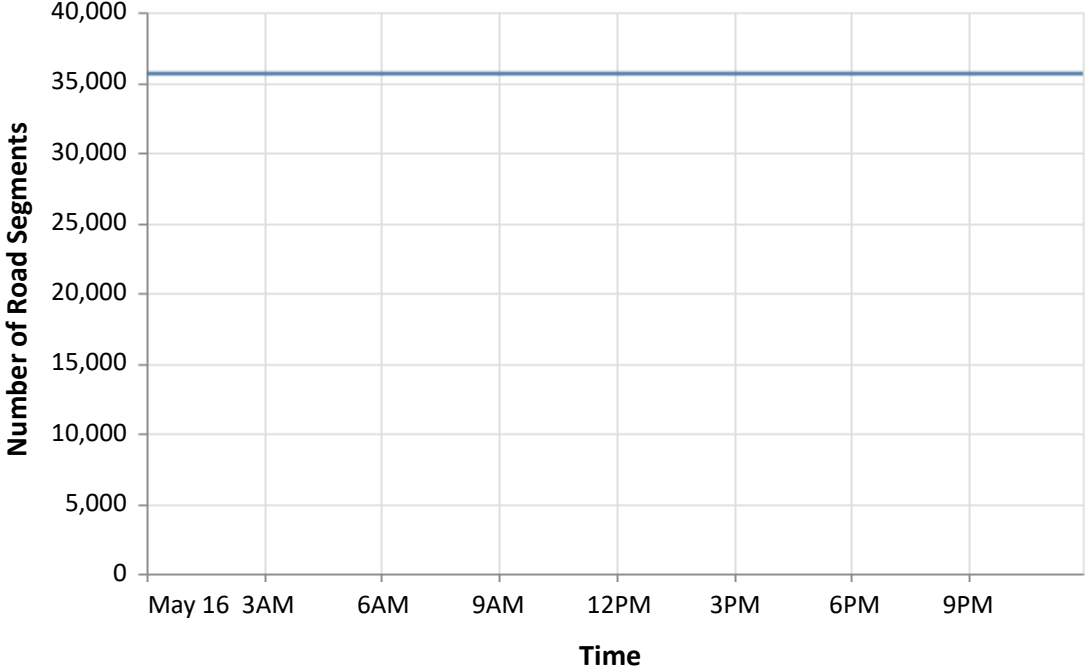
Vehicle Probe Data Detailed Assessment Outcomes

Table 31 details the results of the assessment of the probe vehicle data obtained by the team.

Table 7. Vehicle Probe (NPMRDS and a Third-Party Tool) Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>While the TMC road network used by NPMRDS covers the nation rather well; when querying NPMRDS, the number of road segments returned varies depending on the location or state, as well as the time interval for which the data are requested. This difference can be drastic; for some states, the data returned can have a complete dataset covering the states, and for other states, the data returned show very sparse road segments with probe data covering an exceedingly small amount of the state road network. The team looked at all TMC road segments with probe data returned from the NPMRDS for May 16, 2021, for the entire state of Ohio. The Ohio data are complete, with 35,626 road segments (out of a total of 35,698 TMC road segments in Ohio) containing probe data for every 5 min interval of that day. Similarly, the team looked at all TMC road segments with probe data returned by NPMRDS on the same data for the entire state of Minnesota. The Minnesota data are sparse, with only 27 road segments (out of a total of 9,970 TMC road segments in Minnesota) containing probe data for every 5 min interval. This means that NPMRDS data for the same day are available for 99.8 percent of the Ohio road network but only 0.27 percent of the Minnesota road network.</p> <p>The third-party uses the same data providers as NPMRDS, and the team noticed similar issues with the completeness of the data. Completeness depends on the density of vehicle probe data on the road network at any given time. Figure 64 shows a road segment count distribution in the third-party data for the state of Utah on May 14, 2021. There are many road segments with fewer than 25 rows, which is less than 4 hours of data in that day, in the dataset. Figure 64 represents 24 hours of data (May 14, 2021) downloaded via a third-party tool for all road segments in Utah (a total of 12,085,127 records over 276,916 total road segments). Figure 64 illustrates the number of road segments (y-axis) binned by the number of rows of data/records (x-axis) for every five-minute interval. In other words, for each five-minute interval, available vehicle probe readings are added as a row in the data. Therefore, for the 24-hour period of interest, there is a maximum of 288 five-minute intervals. The bar on the far-right of the chart illustrates that there are only 6225 road segments (2 percent) that have the full (or close to full) 288 rows of data, while the far-left of the chart shows that there are</p>

Assessment Criteria	Assessment
	<p>many road segments with data/records for very few of the five-minute intervals. This may be due to many rural roads that have few or no vehicles over many five-minute intervals.</p>  <p>Figure 20. Histogram of Row Counts (Probe Vehicle Records) per Road Segments</p>
Timeliness	<p>The NPMRDS data can be downloaded manually via RITIS’s data downloader tool at 5-minute intervals (the lowest temporal resolution available). This process is not real-time or near real-time. It does take some time for it to process and return the data, from several minutes to days depending on the geographical area and/or time range selected when requesting the data. This means that the NPMRDS data are not readily available for big data analysis interactively and that real-time analysis is not possible. The latter is expected as NPMRDS was not designed to support real-time data feeds.</p> <p>The third-party data are typically available within 10 minutes of recording. Data are available using the API or the webpage by time of day, day of week, and time range. The tool is not meant to provide discrete data, aggregated or trended data. The latency of the data is typically a month; so just like NPMRDS, the third-party tool is not meant to support interactive data analysis over massive quantities of data or real-time analysis (i.e., big data analysis).</p> <p>Because the process for both data sources is manual and requires multiple requests of the system to obtain the necessary data, it prevents any automated processing on the data in a timely manner, which is required for big data analysis.</p>
Consistency	<p>NPMRDS data are consistent when considering metrics over time. Figure 65 shows the number of road segments for Ohio with data returned every 5 minutes for 24 hours on May 16, 2021. In Ohio, the NPMRDS data are consistent over 24 hours. A similar graph was drawn for Minnesota (not shown here) where, even though a small portion of the Minnesota network (27 road segments) is returned, the returned data are consistent over 24 hours.</p> <p>On the contrary, data from the third-party tool are not as consistent. Figure 66 shows a similar count of the number of links (road segments) in Utah with data every 5 minutes for 24 hours on May 14, 2021. This time the count is a lot less consistent, showing a significant drop of road segment counts at</p>

Assessment Criteria	Assessment
	<p data-bbox="358 296 1466 359">night, an improvement during the day, and a large spike during the evening, which is not very intuitive or explainable.</p>  <p data-bbox="383 1073 1442 1136">Figure 21. Count of Road Segments with Vehicle Probe Data on May 16, 2021, in Ohio (from NPMRDS)</p>

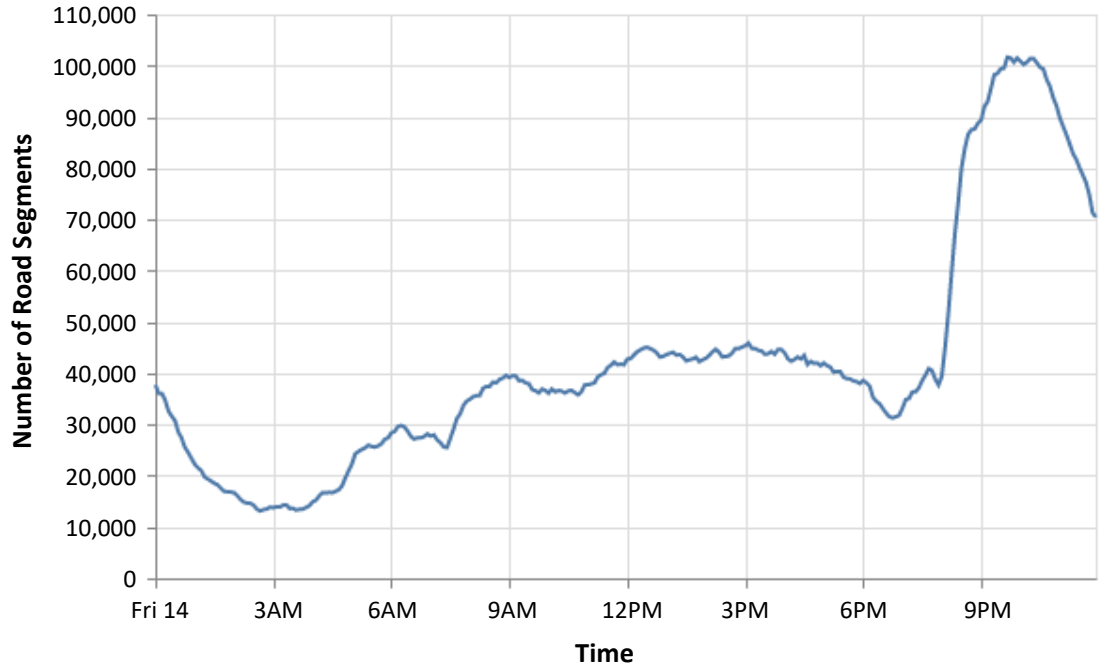
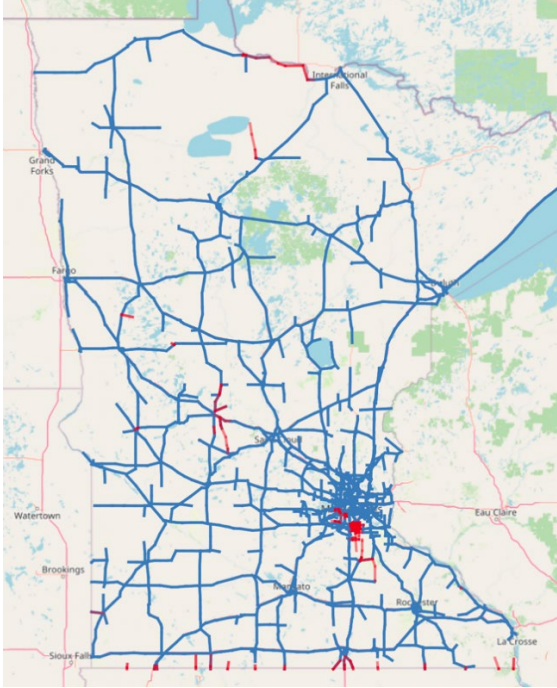
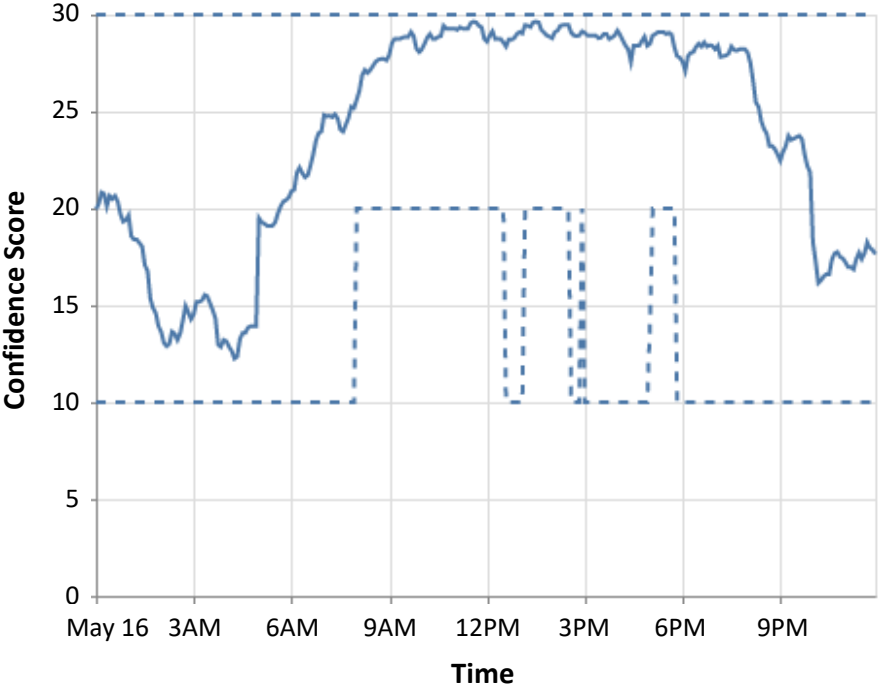


Figure 22. Count of Road Segments with Vehicle Probe Data on May 14, 2021, in Utah

While NPMRDS is consistent over time, it is not geographically consistent. This is due to the drastic discrepancies between states (detailed in the completeness section) and because of the two versions of the TMC network that NPMRDS uses. There are discrepancies between the TMC network NPMRDS used for data pre-2017 and the TMC network NPMRDS uses post 2017. Figure 67 shows an overlay of the pre (blue) and post 2017 (red) NPMRDS TMC networks in Minnesota. Road segments in red, not covered by blue ones, are new road segments that were not part of the TMC network prior to 2017. There are also road segments present in the pre-2017 TMC network that do not exist in the post-2017 NPMRDS TMC network, but they cannot be seen on the map.

Assessment Criteria	Assessment
	 <p data-bbox="475 982 1438 1039">© OpenStreetMap – Basemap used under the Creative Commons Attribution-ShareAlike 2.0 License (CC BY-SA 2.0), https://creativecommons.org/licenses/by-sa/2.0/legalcode (no changes made).</p> <p data-bbox="428 1052 1398 1083">Figure 23. Differences between NPMRDS TMC network pre and post 2017</p>
Conformity	<p data-bbox="358 1110 1463 1497">Both NPMRDS and data provided by a third-party have their own specifications. Their specifications conform to expected data types, and both use TMC (Traffic Message Channel) codes or Link IDs to identify road segments and associated TMC identification files containing metadata for each segment, as well as coordinates for each beginning and end point of a segment. The location data are simple, using plain text representation of latitudes and longitudes of road segments in multiple columns. The data do not use common modern geographical and geometry representation formats such as GeoJSON (Geometry JavaScript Object Notation) or WKT (Well-Known Text). Times are in local time without time zone information, and speeds are in miles per hour (mph). Time zone information is included. Both specifications also include some sort of quality metrics, such as the confidence score values and the C-Values, to provide more insights into the actual metric values, which aligns with data management good practices.</p>
Accuracy	<p data-bbox="358 1518 1463 1898">For both NPMRDS and the third-party, the data returned appear to be accurate without any obvious anomalies. However, many of the metrics returned are not the direct result of the aggregation of probe data; many values are imputed (i.e., missing values are estimated and inserted into the dataset) because there are not enough actual data for some periods of time or locations. This is because probe data vary widely based on the location, time of the day, week, and years as it depends on the number of vehicles traveling on a road segment at the time. Therefore, the team reviewed accuracy in both datasets by examining the quality metrics available with each record. Figure 68 and Figure 69 show plots of NPMRDS confidence scores for all road segments in Ohio and Minnesota, respectively, on May 6, 2016 (the dotted lines represent the minimum and maximum confidence scores found in the datasets throughout the day; the maximum all day for both states is 30 – shown at the top of the chart). For both states, the confidence scores, and therefore the accuracy of the data, drop</p>

Assessment Criteria	Assessment
	<p>significantly overnight, as fewer vehicles are on the roadways, and more imputed data are used. It is interesting to note that in Minnesota, even at some points in the afternoon, the minimum confidence score is as low as it is overnight. This means that even though the average confidence score is high, some road segments in Minnesota are not traveled enough during the afternoon, and the data need to be imputed and are therefore less accurate. Interestingly, Ohio shows a consistent minimum confidence score of 10 during the entire day, revealing that some road segments never had enough data to generate good calculations across the entire day.</p>  <p>Figure 24. NPMRDS Confidence May 6, 2021, Across Minnesota</p>

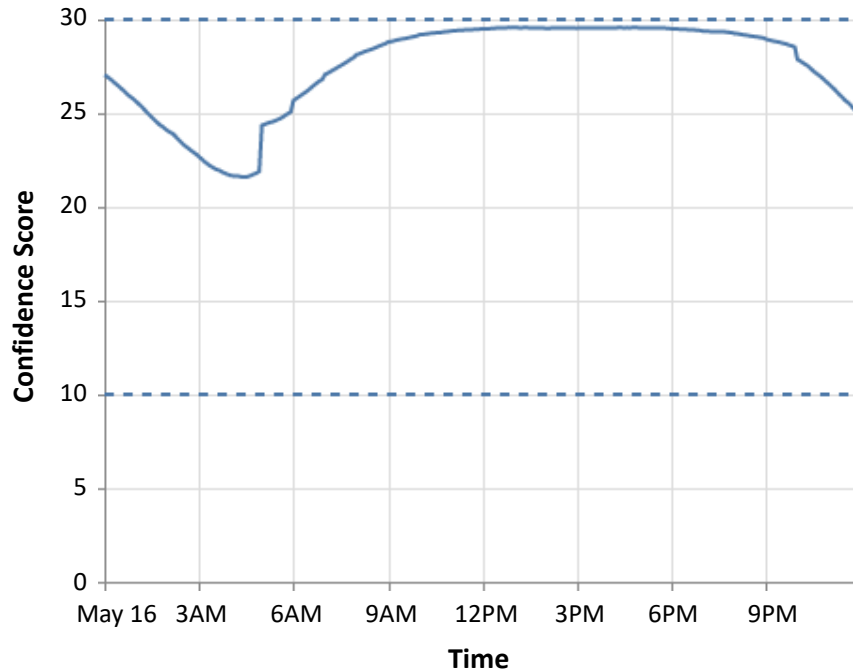
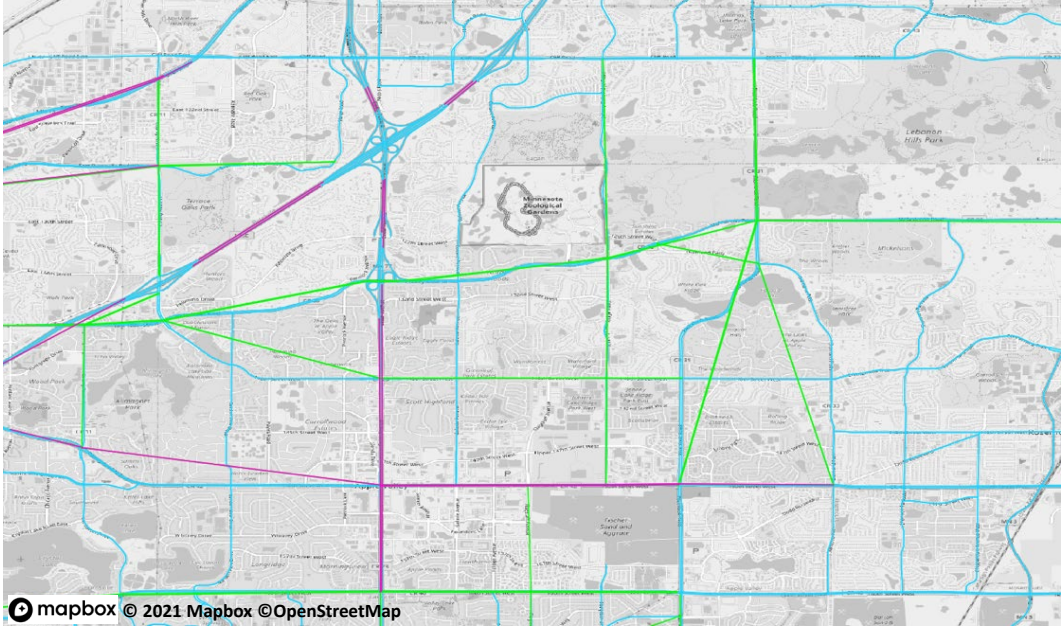


Figure 25. NPMRDS Confidence Score on May 6, 2021, Across Ohio

Another aspect of accuracy in both NPMRDS and the third-party data is the geographical accuracy of the road segments. Both use similar versions of the TMC network. Figure 70 shows a small section of Minneapolis. Green lines are the pre-2017 TMC road segments, purple lines are the post-2017 TMC road segments, and blue lines are the ARNOLD road segments. It is interesting to note how coarse the TMC network is in comparison to the ARNOLD network. Some TMC segments simplify the road network to the point of excluding two or more intersections and crossroads. This makes the precision of the speed provided for a segment coarse and difficult to extrapolate what the speed was near an actual point, an accident or construction zone, and along the actual road, especially if intersections before and after this point are considered in the data. The ability to snap a crash location from a crash report to a TMC road segment could also be challenging. The TMC road segments are simplified to straight lines between two intersections and can deviate several blocks away from the actual road. This makes it challenging to connect some crashes to the TMC segments. More sophisticated snapping methods than a simple range search need to be employed, and some crashes may still require manual intervention to be connected to TMC road segments.

Assessment Criteria	Assessment
	 <p data-bbox="513 936 1312 968">Figure 26. NPMRDS TMC Network Resolution in Minneapolis</p>
Integrability	<p data-bbox="360 997 1445 1060">Data are accessible, but only through a tedious manual process of downloading a series of individual files. The data can be tied to other datasets through the TMC codes.</p> <p data-bbox="360 1073 1458 1346">While NPMRDS and the third-party make data accessible by using modern formats and standards and providing APIs and web interfaces, they only do so through a tedious and restrictive manual process from the point of view of big data analysis and provide data only “drop by drop.” Both systems are optimized to show data as maps, aggregates, and trends on small portions of the network and time (e.g., route, month of the year). While both tools have the capacity to provide data that could be integrated at scale or in real-time, both are designed to limit the ability of users to perform analysis on data at a large scale. This is likely due to restrictions imposed on tool designers from the data source providers.</p>

APPENDIX H

Roadway Inventory Data Detailed Assessment Outcomes

Table 32 details the results of the quality assessment of the roadway inventory data from these states.

Table 8. Roadway Inventory Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
<p>Completeness</p>	<p>Roadway inventory data are often incomplete and suffer a lack of conscientious data management. Roadway inventories are often missing road segments, as well as metadata for road segments. For example, in Tennessee’s roadway inventory, the “Pavement Roughness” metadata field is populated for only 0.06 percent of the records. The rest is listed as “No Data.” The dataset does not cover local roads. Figure 71 shows the “Local Roads” in Colorado. While local roads are included in the roadway inventory, the data are not available for many of them. Blue segments are included in the roadway inventory, and gray segments are on the map only, indicating that they are missing from the dataset.</p> <div data-bbox="639 842 1211 1339" data-label="Image"> </div> <p>© OpenStreetMap – Basemap used under the Creative Commons Attribution-ShareAlike 2.0 License (CC BY-SA 2.0), https://creativecommons.org/licenses/by-sa/2.0/legalcode (no changes made).</p> <p>Figure 27. Roadway Inventory—Local Roads in Colorado</p> <p>In Massachusetts, the roadway inventory dataset is complete, covering all roads in the state even bicycling and walking pathways; however, metadata are lacking, with 25 percent of all metadata not populated (e.g., 35 percent of roadway numbers, 55 percent of road segment IDs, 10 percent of street names, 89 percent of road segment length and 58 percent of speed limits).</p>
<p>Timeliness</p>	<p>Roadway inventory data are not frequently updated. It may take several months to several years to update the geometry and metadata of roadway inventories. For example, Tennessee’s “Pavement Roughness” inventory as received in November of 2020, was last updated in June 2016. These delays reduce data value, as this stale road inventory data become difficult, if not impossible, to integrate with other commercial or public data that include more up-to-date location information.</p>

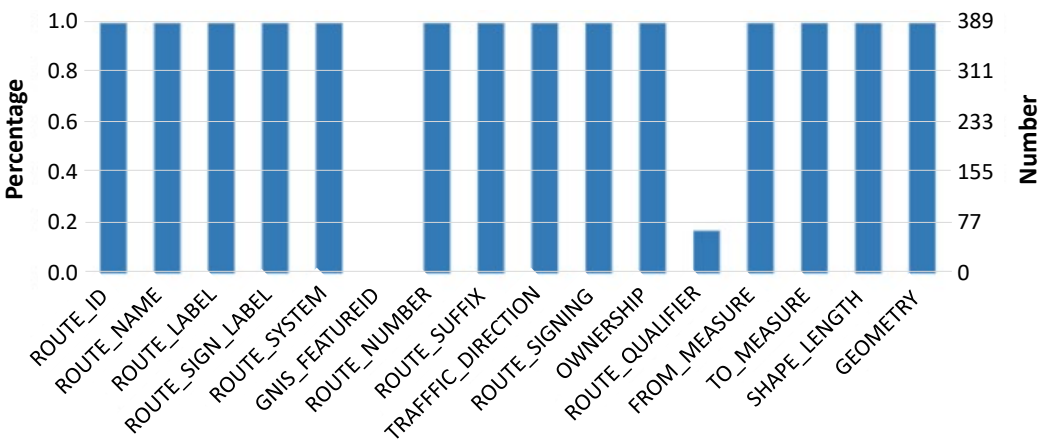
Assessment Criteria	Assessment
Consistency	<p>Inconsistencies can be found between data updates. Inconsistencies may be found because new formats are used to add updated data to the roadway inventory without modifying prior data. Inconsistencies can also be observed between highway and arterial metadata and local roads metadata where data come from different agencies using different standards, nomenclature, and precision requirements.</p>
Conformity	<p>Data conform to best practices for data typing and formatting, but it is important to review data for odd date formats or geometries that are not in the most common Coordinate Referencing System (CRS). Often conversions are needed so that the data can be used with latitude/longitude coordinates in other datasets.</p> <p>The Model Inventory of Roadway Elements (MIRE) is a recommended listing of roadway inventory and traffic elements critical to safety management. MIRE is intended as a guideline to help transportation agencies improve their roadway and traffic data inventories. It provides a basis for what can be considered a good/robust data inventory and helps agencies move towards the use of performance measures to assess data quality. The MIRE listing contains 202 data elements divided among three broad categories: roadway segments, roadway alignment, and roadway junctions. The composition of MIRE was purposefully designed to link with supplemental databases including roadside fixed objects, signs, speed, automated enforcement devices, land use elements related to safety, bridge descriptors, and railroad grade-crossing descriptors. Yet its adoption is only partial in most roadway inventories and often limited to the most traveled roadways, such as highways and arterials. Conformity drops significantly when considering smaller and less traveled roadways.</p>
Accuracy	<p>Geometries represented within road inventories may be off by tens of meters in some places but are generally accurate. The accuracy of the asset data can also be affected as agencies may not have the resources to update asset records as soon as an asset is upgraded or replaced, resulting in stale asset data several weeks or months after asset work has been performed. In the Massachusetts roadway inventory, for example, several versions of the same roadway segments coexisted without any reliable way to identify which was the most current. This drastically lowers the accuracy of the dataset.</p>
Integrability	<p>With a conversion of the geometries to a standard WGS84 CRS, roadway inventory data can be matched with other datasets; however, as is true with all geographical datasets, it may be difficult to match road segments within a roadway inventory to other maps or other geographic datasets. An intermediary process may be required to “snap” road segments to connect multiple data sources. As has been observed with accuracy, the presence of multiple versions of the same road segment without any clear indication of the currency will make integration with other datasets more challenging than it needs to be. Missing data, such as heading, roadway segment ID, or roadway name, can affect the integration with other datasets and only allow partial matching. A significant lack of metadata, such as speed limit, lane count, and presence of shoulder, can also drastically reduce the usability of roadway inventory dataset and lead data analysts to ignore them and use other datasets that are more current and more complete.</p>

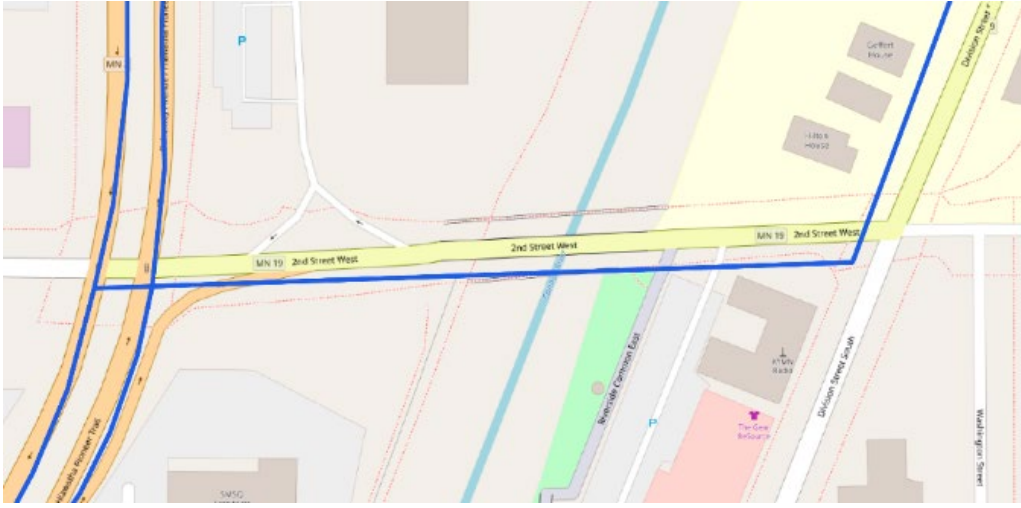
APPENDIX I

LRS Data Detailed Assessment Outcomes

Table 33 details the results of the quality assessment of the roadway inventory based on data from these states.

Table 9. LRS Data Detailed Assessment Outcomes

Assessment Criteria	Assessment																																																			
<p>Completeness</p>	<p>LRS data are typically complete in terms of the main highways in a state, but local roads are often not included. This limits how these data can be used. Often data for local and arterial roads are inaccurate (e.g., segment labeling and “from”/”to” measures are incorrect or missing). Figure 72 shows a simple visualization of missing values by column in the Minnesota LRS data.</p>  <table border="1" data-bbox="412 751 1458 1192"> <caption>Columns in Minnesota DOT LRS Data</caption> <thead> <tr> <th>Column</th> <th>Percentage</th> <th>Number</th> </tr> </thead> <tbody> <tr><td>ROUTE_ID</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_NAME</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_LABEL</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_SIGN_LABEL</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_SYSTEM</td><td>1.0</td><td>389</td></tr> <tr><td>GNS_FEATUREID</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_NUMBER</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_SUFFIX</td><td>1.0</td><td>389</td></tr> <tr><td>TRAFFIC_DIRECTION</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_SIGNING</td><td>1.0</td><td>389</td></tr> <tr><td>OWNERSHIP</td><td>1.0</td><td>389</td></tr> <tr><td>ROUTE_QUALIFIER</td><td>0.15</td><td>58</td></tr> <tr><td>FROM_MEASURE</td><td>1.0</td><td>389</td></tr> <tr><td>TO_MEASURE</td><td>1.0</td><td>389</td></tr> <tr><td>SHAPE_LENGTH</td><td>1.0</td><td>389</td></tr> <tr><td>GEOMETRY</td><td>1.0</td><td>389</td></tr> </tbody> </table> <p>Figure 28. Counts per column for Minnesota LRS data. Most fields are complete.</p>	Column	Percentage	Number	ROUTE_ID	1.0	389	ROUTE_NAME	1.0	389	ROUTE_LABEL	1.0	389	ROUTE_SIGN_LABEL	1.0	389	ROUTE_SYSTEM	1.0	389	GNS_FEATUREID	1.0	389	ROUTE_NUMBER	1.0	389	ROUTE_SUFFIX	1.0	389	TRAFFIC_DIRECTION	1.0	389	ROUTE_SIGNING	1.0	389	OWNERSHIP	1.0	389	ROUTE_QUALIFIER	0.15	58	FROM_MEASURE	1.0	389	TO_MEASURE	1.0	389	SHAPE_LENGTH	1.0	389	GEOMETRY	1.0	389
Column	Percentage	Number																																																		
ROUTE_ID	1.0	389																																																		
ROUTE_NAME	1.0	389																																																		
ROUTE_LABEL	1.0	389																																																		
ROUTE_SIGN_LABEL	1.0	389																																																		
ROUTE_SYSTEM	1.0	389																																																		
GNS_FEATUREID	1.0	389																																																		
ROUTE_NUMBER	1.0	389																																																		
ROUTE_SUFFIX	1.0	389																																																		
TRAFFIC_DIRECTION	1.0	389																																																		
ROUTE_SIGNING	1.0	389																																																		
OWNERSHIP	1.0	389																																																		
ROUTE_QUALIFIER	0.15	58																																																		
FROM_MEASURE	1.0	389																																																		
TO_MEASURE	1.0	389																																																		
SHAPE_LENGTH	1.0	389																																																		
GEOMETRY	1.0	389																																																		
<p>Timeliness</p>	<p>Updates to LRS data are infrequent; since it is unlikely for agencies to have a strong need for real-time updates of changes to routes, this is often delayed. This is a limiting factor when using LRS data for real-time analyses. Stale LRS location data are difficult and sometimes impossible to match to data collected on new road segments that do not align with old ones.</p>																																																			
<p>Consistency</p>	<p>Most LRS data are consistent in naming of routes and format of geometries, but LRS datasets may be a combination of state-level and local-level geodata, and some inconsistencies can be found in how metadata are expressed between highways, arterials, and local roads.</p>																																																			
<p>Conformity</p>	<p>LRS data conform to best practices in how they represent geometries and metadata about routes. The data can come in different file formats, but are usually readable using the Python Geopandas package, QGIS, ESRI, or similar software. The geometries within an LRS dataset may use a different coordinate referencing system than the typical WGS 84 system, but it can be easily converted without too much loss. For example, Minnesota uses the NAD 83 / UTM zone 15N system.</p>																																																			

Assessment Criteria	Assessment
Accuracy	<p>LRS datasets provide a reasonably accurate representation of route geometries but may be off by tens of meters in some cases. In the case of Minnesota’s LRS data, there are cases where the geometry of a route does not line up exactly with the road—especially when there are curves or sharp changes in direction, but even sometimes when the road is straight. This can be observed in Figure 73. This causes problems when trying to snap points with a certain range and the distance of the LRS offset exceeds the snap range.</p>  <p>©OpenStreetMap – Basemap used under the Creative Commons Attribution-ShareAlike 2.0 License (CC BY-SA 2.0), https://creativecommons.org/licenses/by-sa/2.0/legalcode (no changes made).</p> <p>Figure 29. LRS route geometry (in dark blue) overlaid on a map.</p>
Integrability	<p>LRS datasets should have good integrability within themselves and across datasets. Care should be taken to ensure the same coordinate referencing system (CRS) is used across datasets.</p>

APPENDIX J

Third-Party Road Network API Detailed Assessment Outcomes

Table 34 details the results of the quality assessment of the third-party road network API.

Table 10. Third-Party Road Network API Data Detailed Assessment Outcomes

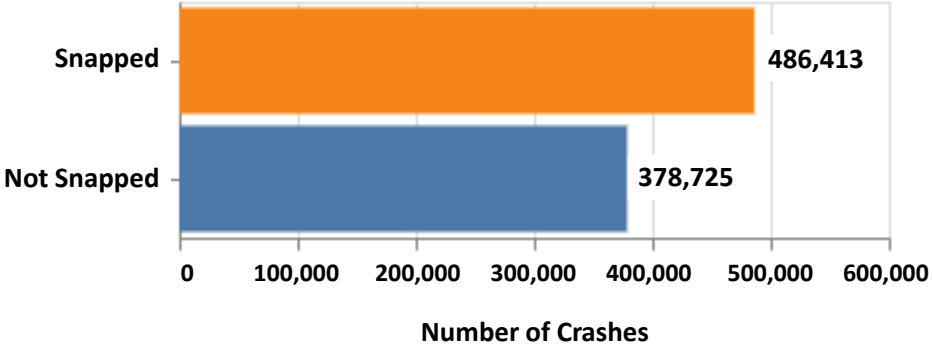
Assessment Criteria	Assessment
Completeness	The third-party road network data provider is known to have wide and up-to-date coverage of roads throughout the U.S.; however, they do not provide details about this coverage. The team was not able to perform an assessment of the completeness of the data, as queries to the API are limited for non-paying users.
Timeliness	The API returns results in near real time. The third-party road network provider is incentivized to provide up-to-date maps and therefore must make a certain amount of effort to ensure maps and roads are current. It is unknown how long it takes for new road segments to be reflected in the data. In 2021, the API started accepting date/time arguments in queries, so it is unknown how the API will deal with historical data queries if drastic changes have been made to the road system.
Consistency	The data are consistent across the different API services. The main data point of interest is the “place ID,” which represents a third-party-defined place within its ecosystem.
Conformity	The API conforms to the WGS84 standard for all its location data of data typing. Snapped points returned by the API include latitude and longitude coordinates with 15-decimal precision. The third-party has its own proprietary standard to express PlaceIDs (road segment identifier).
Accuracy	Since the “nearest road” service does not accept a heading parameter, it may be difficult to attach a coordinate to the correct road segment in some places. Outside of this limitation and the limitations inherent in other data sources (e.g., crash reports), the API itself is highly accurate.
Integrability	Since the data returns a proprietary “place ID” as a result, its data are mostly useful only within the third-party’s ecosystem. The place ID does not translate well to other datasets, nor is it easy to try to map TMC codes or other road encodings to the third-party’s place ID. This limits the usefulness of these services.

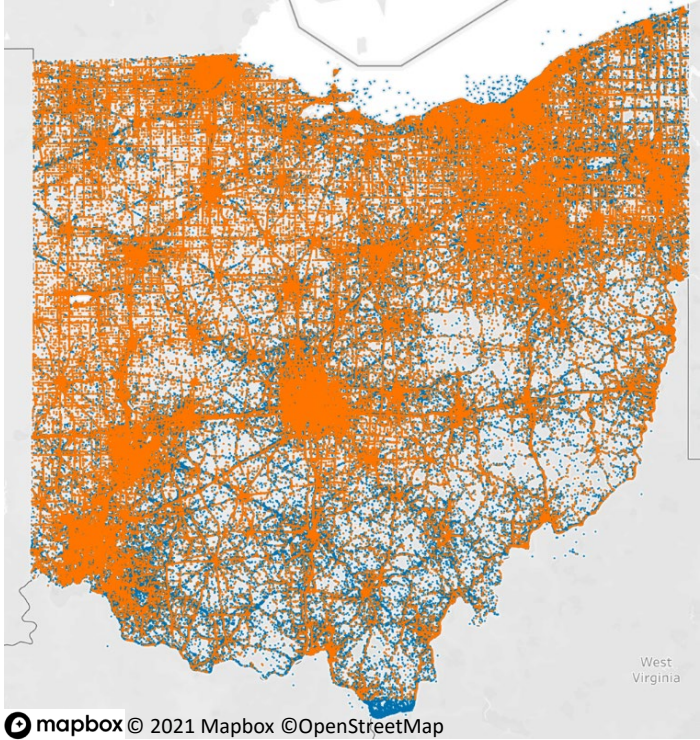
APPENDIX K

SharedStreets Referencing System/OpenStreetMap Detailed Assessment Outcomes

Table 35 details the results of the quality assessment of the SharedStreets Referencing System/OSM.

Table 11. SharedStreets Referencing System/OSM Data Detailed Assessment Outcomes

Assessment Criteria	Assessment						
Completeness	<p>SharedStreets data are as complete as OpenStreetMap, as they are based on that dataset. OpenStreetMap is a crowdsourced effort, meaning updates are made by a vested community of users. Whether or not an arbitrary geographic point from another dataset can be matched with a point in SharedStreets, however, is dependent on how close that point is to a road and whether an accurate heading is provided.</p> <p>The SharedStreets Referencing System has the potential to provide data that can help analysts tie together disparate data sources in the development of big data pipelines. Therefore, the team assessed the SharedStreets Referencing System/OSM. To assess SharedStreets, the team used the SharedStreets toolkit to integrate the location of known crash locations in Ohio to an OSM road segment.</p> <p>SharedStreets has good coverage, yet there are gaps. Figure 74 and Figure 75 illustrate the coverage and the gaps. Figure 74 shows the portion of the Ohio crash records that could (orange) and could not (blue) be snapped to a SharedStreets point. Only about 56 percent of the crash records were snapped to SharedStreets. Figure 75 maps these points – what is in orange represents a crash location that was snapped to SharedStreets (good overall coverage across the state), and what is in blue are the crash locations that could not be snapped to a SharedStreets point (many still show through). In other words, if the team were to use SharedStreets as reference data, many of the crash/incident locations would be lost.</p>  <table border="1" data-bbox="492 1360 1419 1703"> <thead> <tr> <th>Category</th> <th>Number of Crashes</th> </tr> </thead> <tbody> <tr> <td>Snapped</td> <td>486,413</td> </tr> <tr> <td>Not Snapped</td> <td>378,725</td> </tr> </tbody> </table> <p>Figure 30. Ohio Crash records “Snapped” to a SharedStreets Point</p>	Category	Number of Crashes	Snapped	486,413	Not Snapped	378,725
Category	Number of Crashes						
Snapped	486,413						
Not Snapped	378,725						

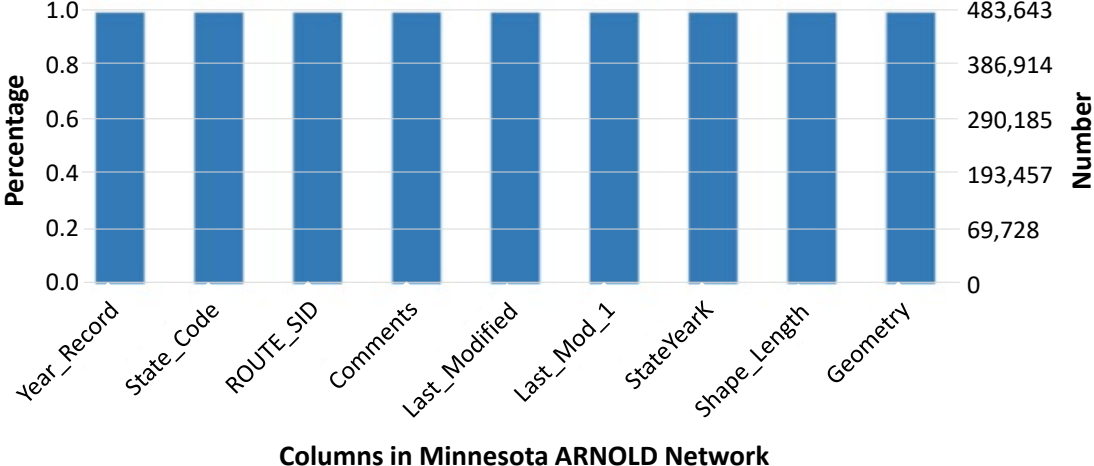
Assessment Criteria	Assessment
	 <p data-bbox="537 1052 1370 1125">Figure 31. Crashes in Ohio with SharedStreets Points in Orange Overlaying the Original Coordinates in Blue</p>
Timeliness	A timestamped version of OpenStreetMap data can be used when making calls to the SharedStreets API. This allows use of specific versions of OpenStreetMap, which is updated nightly.
Consistency	Results from the SharedStreets API are consistent with each other and a particular version of OpenStreetMap yet may change with different versions of OSM as it is updated.
Conformity	The results of the SharedStreets API are formatted as JSON and follow conventions for well-formatted JSON and geometries.
Accuracy	When snapping or matching arbitrary points to SharedStreets points, the accuracy of results is highly dependent on the accuracy of the original coordinates' bearing. For instance, if the bearing is north instead of northeast and the road's bearing in SharedStreets is northeast, then it may not match at all or it may match to a point farther away from where it should match. Due to this, SharedStreets may not be the best choice for datasets that do not have accurate bearings specified.
Integrability	Since SharedStreets uses OSM and includes latitude/longitude coordinates and geometries in its results, it is quite easy to integrate with other datasets.


APPENDIX L

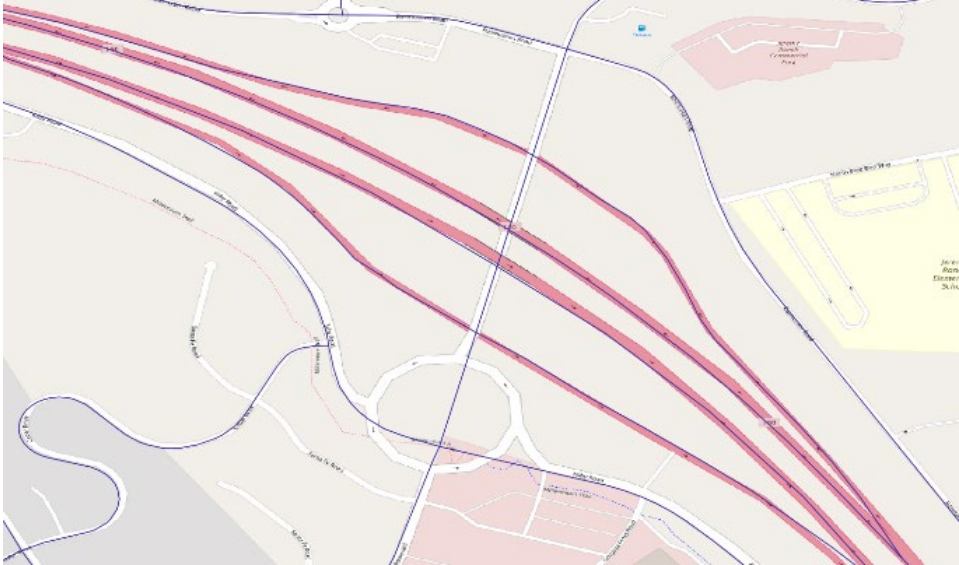
ARNOLD Detailed Assessment Outcomes

Table 36 details the results of the quality assessment of the ARNOLD data.

Table 12. ARNOLD Data Detailed Assessment Outcomes

Assessment Criteria	Assessment																														
Completeness	<p>Completeness of ARNOLD data depends on the state submissions. There are situations where not all fields are consistently populated across roadway types or locations, such as rural arterials vs. urban primary classes.</p> <p>The ARNOLD data reviewed by the team are complete in the sense that there are no missing attributes for any row; however, the data are less complete than the ARNOLD description, particularly regarding the metadata associated with road segments. Figure 76 shows a count of all missing values in each column of the Minnesota ARNOLD network (no missing data).</p>  <table border="1" data-bbox="397 829 1485 1291"> <caption>Columns in Minnesota ARNOLD Network</caption> <thead> <tr> <th>Column</th> <th>Percentage</th> <th>Number</th> </tr> </thead> <tbody> <tr> <td>Year_Record</td> <td>1.0</td> <td>483,643</td> </tr> <tr> <td>State_Code</td> <td>1.0</td> <td>386,914</td> </tr> <tr> <td>ROUTE_SID</td> <td>1.0</td> <td>290,185</td> </tr> <tr> <td>Comments</td> <td>1.0</td> <td>193,457</td> </tr> <tr> <td>Last_Modified</td> <td>1.0</td> <td>69,728</td> </tr> <tr> <td>Last_Mod_1</td> <td>1.0</td> <td>0</td> </tr> <tr> <td>StateYearik</td> <td>1.0</td> <td>0</td> </tr> <tr> <td>Shape_Length</td> <td>1.0</td> <td>0</td> </tr> <tr> <td>Geometry</td> <td>1.0</td> <td>0</td> </tr> </tbody> </table> <p>Figure 32. Ratio of non-null data in each column of the Minnesota ARNOLD dataset</p> <p>The US DOT website states that ARNOLD includes all roads in the US, and only a few local roads are missing from ARNOLD. It includes all highways and arterials but lacks some sections of road segments in the local road system, which can be observed when comparing ARNOLD road segments to the OpenStreetMap base map. Figure 77 shows this kind of comparison in the area around Lynchburg, VA. Blue lines represent ARNOLD segments overlaid on the OpenStreetMap base map where highways are in red, arterials are in yellow, and local roads are in gray. OSM roads can barely be seen because ARNOLD covers most of them in blue.</p>	Column	Percentage	Number	Year_Record	1.0	483,643	State_Code	1.0	386,914	ROUTE_SID	1.0	290,185	Comments	1.0	193,457	Last_Modified	1.0	69,728	Last_Mod_1	1.0	0	StateYearik	1.0	0	Shape_Length	1.0	0	Geometry	1.0	0
Column	Percentage	Number																													
Year_Record	1.0	483,643																													
State_Code	1.0	386,914																													
ROUTE_SID	1.0	290,185																													
Comments	1.0	193,457																													
Last_Modified	1.0	69,728																													
Last_Mod_1	1.0	0																													
StateYearik	1.0	0																													
Shape_Length	1.0	0																													
Geometry	1.0	0																													

Assessment Criteria	Assessment
	 <p data-bbox="396 907 1464 961">© OpenStreetMap – Basemap used under the Creative Commons Attribution-ShareAlike 2.0 License (CC BY-SA 2.0), https://creativecommons.org/licenses/by-sa/2.0/legalcode (no changes made).</p> <p data-bbox="425 974 1451 1045">Figure 33. ARNOLD Network in Blue Overlaid on the OSM Base Map Around Lynchburg, VA</p>
Timeliness	<p data-bbox="396 1075 1471 1318">The ARNOLD data are dependent on the input of all states in the nation and, as such, it is difficult to synchronize the collection and verification of data to allow for rapid publishing of network updates. State agencies have different road networks on which they collect data with varying levels of precision, often only on the road segments for which they are responsible. They also have varying levels of resources to collect and prepare new or not yet collected data. The last ARNOLD update used data sent by states in 2017 and was published in 2018. While road changes are not frequent, this frequency is too slow to accurately support most modern real-time application.</p>
Consistency	<p data-bbox="396 1335 1419 1474">The ARNOLD data are consistent across states with similar content being used to describe road segments. This is not surprising as there is little metadata associated with road segments in ARNOLD and all columns, including the comments column, are well-defined and standardized. Unfortunately, this standardized content is minimal.</p>
Conformity	<p data-bbox="396 1491 1455 1696">The ARNOLD dataset is entirely standardized. This is true for the date-time and location fields expressing the state and recency of the data and for route IDs (both local and national), which are expressed using a specific ARNOLD format including a positive and negative distinction on separated roadways. Road types are expressed using standardized text in the comments fields to identify main roads, ramps, etc. All geometries are expressed using the WGS84 referential system for every state.</p>
Accuracy	<p data-bbox="396 1713 1458 1881">The geometric accuracy of ARNOLD is not ideal. It is as good as the state data it depends on. Its geometric accuracy is also affected by the fact that ARNOLD is published several years after the data are collected, and new construction projects may be started and completed during that time. Figure 78 shows the ARNOLD network in purple. There is a missing roundabout in ARNOLD at an exit on I-80 in Utah.</p>

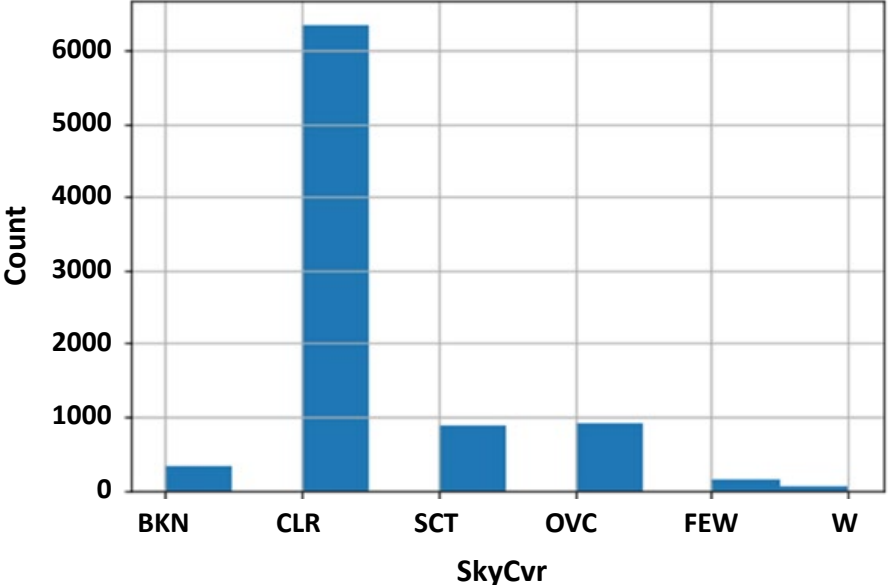
Assessment Criteria	Assessment
	 <p data-bbox="477 877 1479 930">© OpenStreetMap – Basemap used under the Creative Commons Attribution-ShareAlike 2.0 License (CC BY-SA 2.0), https://creativecommons.org/licenses/by-sa/2.0/legalcode (no changes made).</p> <p data-bbox="550 942 1325 974">Figure 34. Missing Roundabout in ARNOLD Off I-80 in Utah</p> <p data-bbox="396 997 1471 1060">The ARNOLD metadata is also generally accurate; however, these data are so minimal, restricted to year, state code, automatically calculated shape length, and simplistic road type.</p>
Integrability	<p data-bbox="396 1100 1471 1520">Integrating the ARNOLD dataset may pose some challenges. While the ARNOLD geometry is accurate enough to be conflated to other geo-datasets with “close enough” road segments, it will be much more difficult to integrate ARNOLD with datasets that have less accurate and more simplistic road geometries. Integrating with such datasets would require some additional steps to match road segment metadata, such as common road name, directional/flow indicator, road type, mileage/road measure, which as of now are missing in ARNOLD despite the plan to incorporate them. Also, since ARNOLD is a compilation of LRS routes, topology as a measure of connected segments is completely missing. This means any process relying on connecting nodes is suspect and every attempt will need to be made via fuzzy tolerances to find subsequent segments. ARNOLD would definitely benefit from adding a flow direction/heading attribute to the road segments. This would avoid having to calculate the flow direction for each heading to be matched during integration efforts.</p>

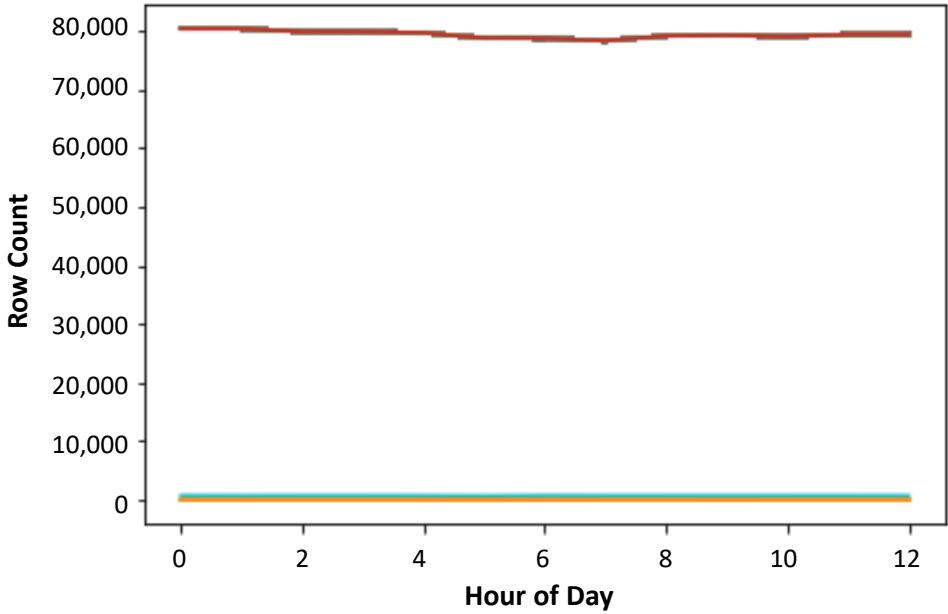
APPENDIX M

MADIS Data Detailed Assessment Outcomes

Table 37 provides more details on the quality assessment of the MADIS data.

Table 13. MADIS Data Detailed Assessment Outcomes

Assessment Criteria	Assessment														
Completeness	<p>The MADIS data sample provided on the NCEP website (after being serialized and flattened from its NetCDF format to a tabular format) contains 1,032,107 records, each composed of 192 columns. Out of these 192 columns, only four appear to be missing data. Below are the measures with missing data and the percentage of the missing data in the dataset:</p> <ul style="list-style-type: none"> • Elevation is missing in 0.0013 percent of the records. • Station is missing in 0.0027 percent of the records. • Station type is missing in 0.48 percent of the records. • Skycvr is missing for 99 percent of the records. <p>Three out of these four – elevation, station name, and station type – could be easily fixed using data imputation, but Skycvr cannot, as most data for that measure are missing. Figure 79 shows the breakdown of the few thousand records in the sample dataset with non-null Skycvr values in over one million records.</p>  <table border="1" data-bbox="532 1010 1414 1591"> <caption>Data for Figure 35: Count of SkyCvr Measure in the Entire MADIS Sample Dataset</caption> <thead> <tr> <th>SkyCvr Measure</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>BKN</td> <td>~300</td> </tr> <tr> <td>CLR</td> <td>~6500</td> </tr> <tr> <td>SCT</td> <td>~800</td> </tr> <tr> <td>OVC</td> <td>~900</td> </tr> <tr> <td>FEW</td> <td>~100</td> </tr> <tr> <td>W</td> <td>~50</td> </tr> </tbody> </table> <p>Figure 35. Count of SkyCvr Measure in the Entire MADIS Sample Dataset</p>	SkyCvr Measure	Count	BKN	~300	CLR	~6500	SCT	~800	OVC	~900	FEW	~100	W	~50
SkyCvr Measure	Count														
BKN	~300														
CLR	~6500														
SCT	~800														
OVC	~900														
FEW	~100														
W	~50														
Timeliness	<p>The MADIS data are not meant to be used for real-time analysis, and out of the many data sources, while some may be in near-real-time (e.g., weather station broadcasting data every 5 to 30 minutes) others (e.g., satellites) take several hours before they send data. To assess timeliness of this single day MADIS sample, the team tried to detect if measures with non-null values were consistent across all hours of the day. Figure 80 shows that for most of the measures, about 80,000 of them were reported consistently every hour.</p>														

Assessment Criteria	Assessment
	<p data-bbox="488 268 1422 331">Some measures containing nulls, such as Skycvr, are shown at the bottom with almost zero record counts.</p>  <p data-bbox="683 978 1273 1010">Figure 36. Distribution of MADIS values per hour</p>
Consistency	See Accuracy
Conformity	<p data-bbox="488 1083 1435 1255">The MADIS data are strictly standardized and provided using a NetCDF file format. Data and associated metadata, including level-1 quality metrics, provide information on the conformity of the data in each record. Using NetCDF data analysis tools, it is easy to quickly identify nonconforming data and filter them out. In the MADIS data sample, all records passed the level-1 quality check.</p>
Accuracy	<p data-bbox="488 1276 1446 1549">The team was not able to perform comparisons on the MADIS dataset to establish the accuracy and consistency of the sample observations. Instead, the team used the quality control metadata provided as part of the MADIS dataset and assessed how many of the records did not pass the quality control. The MADIS quality control is composed of two categories of quality control (QC) checks – static and dynamic. The static QC checks are single-station, single-time checks for validity, internal consistency, and vertical consistency. The dynamic QC checks include position consistency, temporal consistency, and spatial consistency. The results of these tests are combined into three levels:</p> <ul data-bbox="537 1566 1143 1661" style="list-style-type: none"> • level 1 = validity • level 2 = internal consistency, temporal consistency • level 3 = spatial consistency check <p data-bbox="488 1671 1455 1839">The team observed that in the sample, all but 126,645 records (12 percent) passed all quality controls. Non-null measures that did not vary across the entire day were also identified. They were mostly data fields such as IDs, documentation references, elevation and quality and consistency measures that were not expected to change over the course of the day. Non-varying measures included:</p> <ul data-bbox="537 1856 704 1883" style="list-style-type: none"> • isOverflow

Assessment Criteria	Assessment
	<ul style="list-style-type: none"> • handbook5Id • homeWFO • numericWMOid • pressChangeChar • pressChange3Hour_pascal • pressChange3HourDD • pressChange3HourQCA • pressChange3HourQCR • pressChange3HourQCD_pascal • pressChange3HourICA • pressChange3HourICR • windDirICA • windDirICR • windSpeedICA • windSpeedICR • windDirMaxQCD_degree • visibilityQCD_meter • visibilityICA • visibilityICR • precipAccumQCD_mm • precipAccumICA • precipAccumICR • precipRateQCD_meter_second • timeSinceLastPcp_second • seaSurfaceTempICR • roadLiquidIcePercent4_percent • soilMoisturePercentQCD_percent • mobileElev_meter
Integrability	<p>The MADIS data are meant to be available for the widest range of users, but primarily for NOAA scientists. As such, MADIS uses the NetCDF file format, which is designed to store scientific data. Originally created by NASA, NetCDF, which stands for network Common Data Form, is a file format for storing multidimensional scientific data (array oriented) such as temperature, humidity, pressure, wind speed, and direction. This file format is supported by many libraries and tools created by the scientific community that make it easy to extract and filter data out of each NetCDF file. Unfortunately, these tools were also mostly created for programming languages used by scientists (e.g., FORTRAN, C/C++) not for the current data science and GIS toolkits, which are built around the Python and R programming languages. This makes data stored in NetCDF files a bit more challenging to extract and use with common data science and GIS tools.</p> <p>A few libraries have been developed in Python and R to work with NetCDF files, but they are not as complete as the ones in FORTRAN and C/C++. This means that for data analysts to use NetCDF data, they need to learn the NetCDF data structure, understand its multi-dimensional array structure and where the data and metadata are located and related, and then develop custom code before they can convert the NetCDF data into a format they can easily use with Python or R. This is not a trivial task, and the project team had to</p>

Assessment Criteria	Assessment
	<p>develop custom code using Jupyter Notebook (a Python data analysis tool) before assessing the MADIS data sample.</p> <p>NetCDF and the data contained in the MADIS dataset are beyond the needs of most transportation agencies, and it would be preferable for such data to be simplified and available in a format that could be imported without any complex extraction process. Commercial weather data services have already identified this as an opportunity, and some weather data services use the MADIS data in combination with a few other weather data sources, combine them, simplify them, and make them available to the public for a fee in formats that are much easier to consume.</p> <p>NOAA provides access to its MADIS real-time and archived NetCDF files via file transfer protocol (FTP), which is now considered a legacy file sharing service and is not ideal for sharing large datafiles. MADIS data are also accessible in two other ways – a client and a server tool called Local Data Manager (LDM),² which can provide event-based MADIS data or through a web-based data retrieval tool following the open-source project for Network Data Access Protocol (OPeNDAP).³ The latter has client libraries in Python and R but is still a scientific tool. While it may be easier than parsing NetCDF files, it will still require some work to integrate the MADIS with a common datastore or system and will be better suited for real-time/event-based data processing than for historical data analysis.</p>

² <https://www.unidata.ucar.edu/software/ldm/>

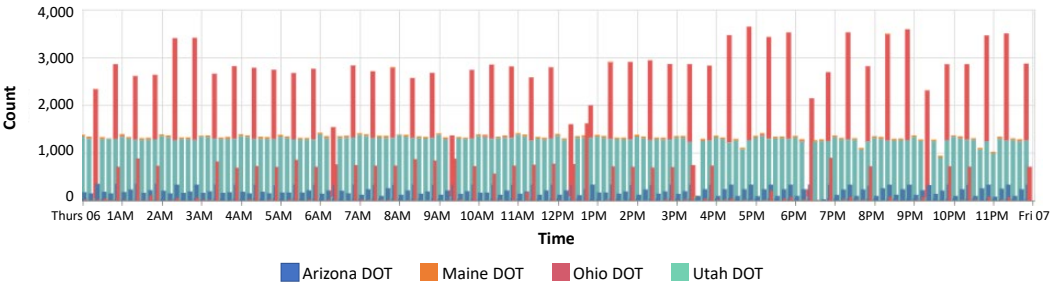
³ <https://www.opendap.org/>

APPENDIX N

Road Weather/Weather Data Environment (WxDE) Data Detailed Assessment Outcomes

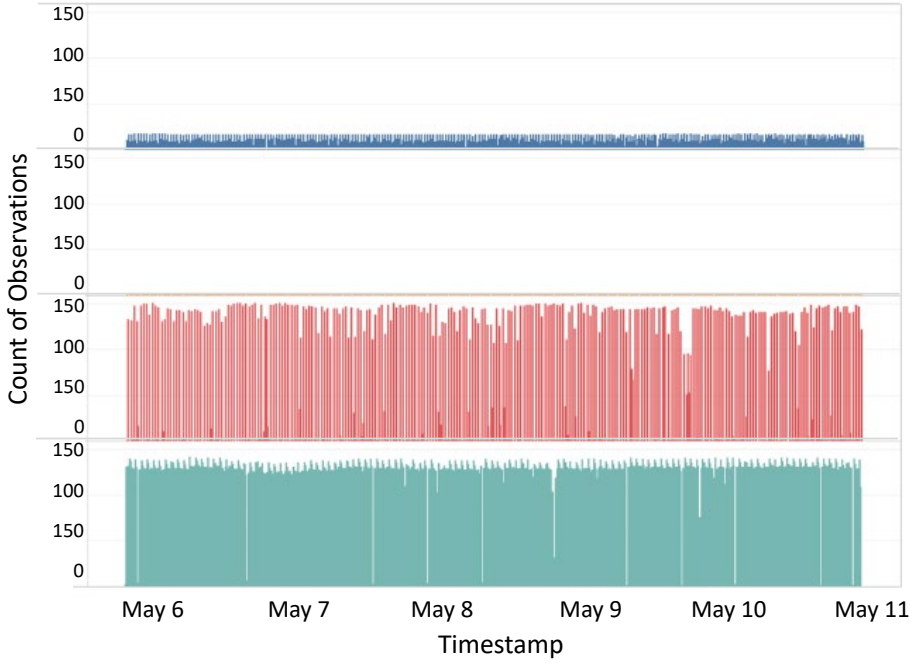
Table 38 details the results of the quality assessment of the WxDE data.

Table 14. WxDE Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>WxDE data are assembled from weather data collected by contributing states from distinct types of weather stations equipped with different sensors measuring different weather observations (based on the need and climate of each state). While the team was able to subscribe for data feeds from seven states, the data received did not include weather observations for every state and the amount of data received was inconsistent for each interval and each state where data were available. Figure 81 shows the number of records received every five minutes from four states over five days.</p>  <p>Figure 37. Count of WxDE Records per State per 5-Minute Interval over 3-Hours</p> <p>Figure 82 shows a count of all observations collected in a data sample from Arizona, Maine, Ohio, and Utah and illustrates that the data collected in one state can be different than the data collected by another, both in terms of the measures and the quantity of measurements. For example, Utah collects data pertaining to snow, while other states have very sparse or non-existent snow data. Figure 82 also shows that the reviewed states are equipped with weather stations and sensors unequally, with states like Utah covering most of their main roadways, while Maine only shows two weather stations for the entire state. This makes the WxDE data incomplete geographically, as not enough fixed weather stations are present in some states to get sufficient weather observations along their road network.</p> <p>Completeness of the WxDE data with regard to time is good for some states in the data sample. Figure 83 shows that the state of Utah (green) consistently reports observations every five min with minimal misses, while states like Ohio (red) have wider variations in the data reported, the numbers of observations they report every 5 minutes, and even how frequently they report observations.</p>

Obs Type Name	Contributor			
	AZ_State_DOT	ME_State_DOT	OH_State_DOT	UT_State_DOT
essAirTemperature	212	22	757	1,985
essAtmosphericPressure	190			
essDewpointTemp	209	22	757	1,985
essIceThickness	149			
essInstantaneousSolarRadiation				907
essMobileFriction	149	22		1,279
essPavementTemperature				1,352
essPrecipitation24Hours		11		
essPrecipitationOneHour	156	11		
essPrecipitationSixHours	156	11		
essPrecipitationThreeHours	156	11		
essPrecipitationTwelveHours	156	11		
essPrecipRate	156	11	530	
essPrecipSituation	179	11		
essPrecipYesNo	179	11		
essRelativeHumidity	212	22	757	1,985
essRoadwaySnowDepth		22		371
essSubSurfaceMoisture				29
essSubSurfaceTemperature	189	22	458	1,845
essSurfaceConductivityV2				29
essSurfaceFreezePoint				15
essSurfaceIceOrWaterDepth	10	22		
essSurfaceStatus			1,036	1,309
essSurfaceStatus2	218	22		
essSurfaceTemperature	211	22	1,034	37
essSurfaceWaterDepth	149	22		
essVisibility	201	22	586	1,641
essWetBulbTemp				1,985
icePercent			1,350	
precip10min				947
precipIntensity			757	
precipType			757	
windSensorAvgDirection			757	
windSensorAvgSpeed			757	
windSensorGustDirection	201	11	564	
windSensorGustSpeed	201	11	757	1,977
windSensorSpotDirection	201	11		
windSensorSpotSpeed	201	11		

Figure 38. Observations Counts per Observation Type Per State

Assessment Criteria	Assessment
	 <p data-bbox="391 982 1425 1056">Figure 39. Number of Air Temperature Observations per 5 Minutes, May 5-10, 2021 for Arizona, Maine, Ohio, and Utah</p> <p data-bbox="370 1066 1455 1381">One of the main drawbacks of the reviewed WxDE data sample is that it only contains observations from fixed weather station measures and therefore can only be used reliably to associate atmosphere and pavement conditions with traffic incidents that occur near weather stations. According to the WxDE documentation, the WxDE data should also contain observations from mobile/telematics weather stations. Such data would definitely improve the geographical and temporal completeness (some mobile/telematics datasets are listed on the WxDE website, but none were accessible at the time the data were assessed). Should this data become available, WxDE would become a much more complete dataset, especially if mobile/telematics reporting are numerous enough to report on most road segments at regular intervals.</p>
Timeliness	<p data-bbox="370 1398 1455 1644">Weather observation data from the WxDE are captured and reported at different frequencies depending on the state, station, sensor, and observations. As such, the same observation measure in WxDE can be updated every 5 minutes in one state and every 15 or 30 minutes in another. This is also observed within states, as weather stations can be of different generations, having different communication capabilities. While WxDE allows for its subscribers to pull data every five minutes, the data received are not aligned with this requirement, which can result in missing data. This can be seen when reviewing the WxDE archived observation data in Figure 82 and Figure 83.</p> <p data-bbox="370 1654 1455 1822">The team also found that in some states, such as Utah, while decent quality real-time data were available, historical data were unavailable. This is problematic as any long-term analysis would require the analyst to capture the data for months/years before being able to perform the analysis. The team also discovered that certain states did not return any data in the real-time feed for over three hours, which would make the assembly of a historical dataset difficult.</p>

Assessment Criteria	Assessment
Consistency	<p>Data across the WxDE system are consistent with each other in format and content. This is due in part to the standard developed by WxDE and the conversion and quality control applied to every observation measure received. It can be noted that WxDE only publishes the best data received (the ones that pass quality control). For some states, this means that the number of observation measures is inconsistent because too many observation measures are missing or fail quality control. These states may have legacy equipment that needs to be replaced or low-quality communication lines between stations and the DOT.</p>
Conformity	<p>WxDE has developed its own specification for the data it publishes. The sample dataset reviewed conformed to the specified WxDE data types and formats, which align with industry best practices. Observation data were provided in the expected units of measurement and the units of measurement were always specified for each observation. Two measures and units, International System of Units (SI) and imperial system of units, were provided for most measurements. Location data were provided using latitude/longitude coordinates expressed using the WSG 84 coordinate referencing system. The timestamps were expressed in the UTC) time zone to eliminate the need to convert time data between time zones and to adjust to seasonal time zone changes.</p>
Accuracy	<p>ESSs may not always be maintained or monitored to counter sensor failure and sensor drift, which can lead to data quality issues (e.g., missing data, erroneous data). To circumvent this problem, quality control is performed by WxDE on collected data. Quality control metrics are computed with each collected observation and provide an indication of the quality of that observation, including checking if the observation is significantly different from historical norms for the observation area or checking if the observation is significantly different from similar nearby observations. Quality control metrics are composed of two categories: Weather Data Environment quality control metrics and Vehicle Data Translator quality control metrics, as well as an overall quality control metric called "Complete." The "Complete" quality control metric can take the following values:</p> <ul style="list-style-type: none"> • "P" – The observation passed the quality check. • "N" – The observation did not pass the quality check. • "-" – The quality check was not run for the observation, usually due to insufficient background field data. • "/" – The quality check was not configured for the observation type. <p>In the data sample collected during May 5 to 10, 2021 from Arizona, Maine, Ohio, and Utah, all observation records for all observation types have a "Complete" quality control metric value of "P," indicating a high accuracy of all records provided by WxDE.</p> <p>It is interesting to note that for Maine, WxDE shows observations coming from two weather stations in the data sample; the MADIS data, on the other hand, indicate the presence of 18 different weather stations in Maine. This suggests that the WxDE may be removing observations that do not pass the quality checks rather than publishing them.</p>
Integrability	<p>The WxDE is a research project built as an open-source platform meant to collect and share transportation-related weather data, particularly CV application weather data. It stores observation measures in a relational database and makes the data accessible using the CSV or XML format, either as a subscription to a real-time data feed or as an archive download. The data available from WxDE include latitude/longitude coordinates in the WSG 84 coordinate referencing system, which allow the sensor data to be conflated with other datasets. It also includes</p>

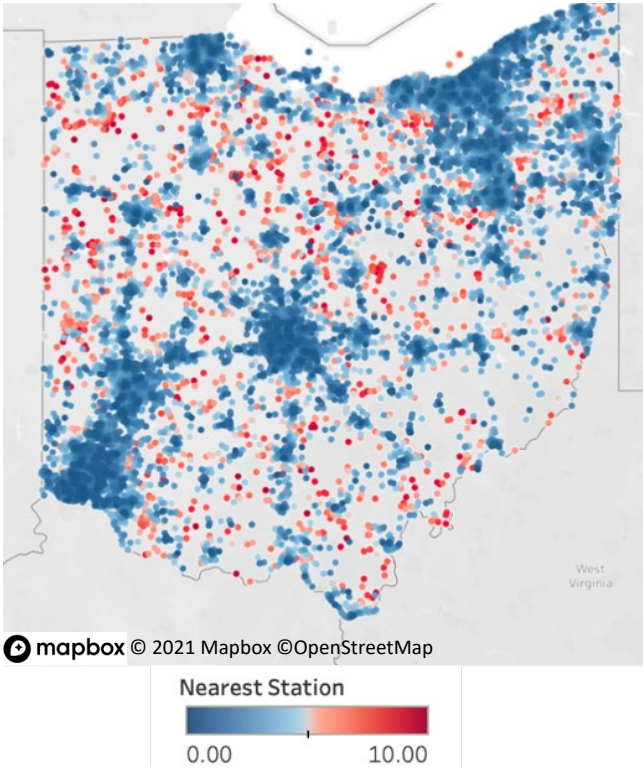
Assessment Criteria	Assessment
	<p>timestamps in the UTC time zone as well as measures expressed in standard units for each observation type. These aspects of the WxDE data make it easily integrated with other datasets. But there are concerns in practice when trying to consume/access WxDE data. These may be due to the research nature and current initial stages of the project. At the time of assessment, there was no way to download WxDE historical data or a robust real-time data feed. Data feed setup required some light “hacking” to work unreliably (some data feeds did not return any data in the real-time feed for over 3 hours), and downloading historical data failed due to an internal website error.</p> <p>These findings made the WxDE platform impossible to use reliably to integrate weather data with other systems/datasets. When considering (at the time) that WxDE only collected and shared fixed weather station data, the prospect of having the platform able to easily collect and share aspiring mobile/telematics weather seems unlikely, especially when considering the legacy technology stack selected for the platform, which will be ill-suited to handle the very large flow of sensor data that will come from mobile/telematics weather sensors.</p>

APPENDIX O

Third-Party Weather API Data Detailed Assessment Outcomes

Table 39 details the results of the quality assessment of third-party weather data.

Table 15. Third-Party Weather API Detailed Quality Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>The third-party weather API data are meant to be hyperlocal, that is, very precise in terms of location and time. To assess the completeness of the data, the team assessed how close the third-party weather measures were to the locations of eight million crashes in seven states. The team found that 99.7 percent of all crashes had a weather station source within ten miles or less of the crash location. Figure 84 shows a map of a sample of crashes in Ohio color coded by their distance to the nearest weather station. The map shows that most crashes in Ohio were associated with weather data from a weather station that was five miles away or less (blue). The crashes that were associated with data from weather stations that were between 5 and 10 miles away (red) are less numerous and located in rural areas.</p>  <p>Figure 40. Distance from Nearest Weather Station</p>
Timeliness	<p>Third-party weather API requests can be made in real-time to obtain data at a specific location from the past, present, or future (forecasted). When queried, the weather API returns a result immediately, and this result may contain weather data at three levels of precision:</p> <ul style="list-style-type: none"> • Currently – at the exact requested time • Hourly – within the hour of the requested time

Assessment Criteria	Assessment
	<ul style="list-style-type: none"> • Daily – the day of the requested time <p>Issues in timeliness, outside of service outages, occur when the third-party service cannot generate weather data exactly at the requested time. In the more than eight million crash locations and times that the team submitted to the weather API over the course of 20 hours, no API call failed, and none of the returned results missed the “currently” weather data.</p>
Consistency	<p>Due to the specific location/time nature of the third-party weather data, it is difficult to test data consistency without collecting a large amount of data over a long period of time at multiple locations. Understanding that MADIS is the main data source used by the third-party weather data provider, it can be assumed that the observation measures coming into the third-party model are consistent, but no information was available regarding the influence of the given modeling algorithm processing the data. The third-party weather data provider does maintain and publish a service status page containing a list of its downtime and incidents since 2012. According to the provider, between January 2019 and April 2021, the service encountered ten downtime events ranging from 3 to 100 minutes with an average of 33 minutes.</p>
Conformity	<p>The third-party developed its own specification, in an analogous way to the free navigation app data provider. It provides data using the REST API protocol, which is the most common API standard in modern data systems. It returns data in JSON format, which is also the most common data object standard in modern systems. Data in the weather API is expressed using a mix of U.S. Imperial and International System (SI) units. Observation narratives are standardized as well, but no documentation is provided on them. Across the eight million data records collected, the team counted 64 unique weather “summary” types.</p>
Accuracy	<p>One of the ways to assess the accuracy of the third-party weather data would have been to compare its measurements to measurements taken at the same time and location by another source, such as the WxDE. Unfortunately, the team was not able to gather enough WxDE data to perform such a comparison. As a substitute, the team calculated the distance to the nearest weather station provided by the third-party weather API in each query result. It can be assumed that the farther the weather station is from the requested location the less accurate the weather measurements may be. Overall, 99.7 percent of all crashes had a weather station within less than 10 miles, which would provide more accurate results for most of the queries. The farthest weather station found in the crash dataset was located 62 miles away from the crash location.</p> <p>Also, the team compared the weather narrative provided by the third-party weather data provider for each crash location and time to the weather information provided in each crash report. The comparison was done using the Levenshtein distance method, which measures the proximity between two texts by calculating the number of characters they do not have in common. Table 40 shows the results of this comparison with Levenshtein distance (number of uncommon characters between the two texts) group by ranges. It can be noted that 72 percent of all third-party weather data records between 20 and 40 characters long have less than 10 uncommon characters with crash report weather narratives. This means that about 70 percent of the third-party weather narratives are a partial match to the crash report narratives.</p>


Assessment Criteria	Assessment												
	<p data-bbox="521 289 1338 317" style="text-align: center;">Table 16. Comparison of Crash Report and Third-Party Weather Narrative</p> <table border="1" data-bbox="680 338 1179 611" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th data-bbox="680 338 964 426">Number of uncommon characters</th> <th data-bbox="964 338 1179 426">Percent of records</th> </tr> </thead> <tbody> <tr> <td data-bbox="680 426 964 464">0 to 5</td> <td data-bbox="964 426 1179 464">40%</td> </tr> <tr> <td data-bbox="680 464 964 501">5 to 10</td> <td data-bbox="964 464 1179 501">32%</td> </tr> <tr> <td data-bbox="680 501 964 539">10 to 20</td> <td data-bbox="964 501 1179 539">27%</td> </tr> <tr> <td data-bbox="680 539 964 577">20 to 30</td> <td data-bbox="964 539 1179 577">21%</td> </tr> <tr> <td data-bbox="680 577 964 611">over 30</td> <td data-bbox="964 577 1179 611">0%</td> </tr> </tbody> </table>	Number of uncommon characters	Percent of records	0 to 5	40%	5 to 10	32%	10 to 20	27%	20 to 30	21%	over 30	0%
Number of uncommon characters	Percent of records												
0 to 5	40%												
5 to 10	32%												
10 to 20	27%												
20 to 30	21%												
over 30	0%												
Integrability	<p data-bbox="418 625 1443 1010">The weather API data are meant to be consumed by modern data systems through a REST API returning JSON formatted data. The weather API does not provide access to a historical dataset; instead, it allows historical, current, and forecasted data to be requested for a specific location and time. This is ideal for integration with modern real-time data systems if the service is reliably available. The use of JSON format, the WGS84 referential system, and standard timestamp and time zone information makes it easy to enrich commercial or public datasets using the weather API in batch or in real-time. The third-party also provides a consistent weather summary narrative. The team counted 64 unique “summary” types across the eight million data points collected. This also allows the third-party weather data to be integrated with other datasets containing weather information in text form, such as crash reports, without too much difficulty.</p> <p data-bbox="418 1024 1378 1121">The third-party also offers more than 50 API client libraries allowing calls to the API to be embedded into applications. Many languages are supported, from C++ to Swift, covering desktop, web, mobile, and backend applications.</p>												

APPENDIX P

Third-Party CV Data Detailed Assessment Outcomes

Table 41 details the results of the assessment of the third-party CV data obtained by the team.

Table 17. Third-Party CV Data Detailed Assessment Outcomes

Assessment Criteria	Assessment
Completeness	<p>The team performed a review of third-party CV data to determine their completeness for the selected geography and time. The attributes of the data were all provided and there were no missing attributes within the rows for either the CV driver events data or the CV movement data. It was also noted that no gaps were detected within the provided time. The provided data was complete, as expected.</p> <p>The third-party CV dataset included CV movement data on all levels of roadway in the study area for the provided time, but the actual market penetration is unknown. However, the completeness of the data on all available roads allows the data to be used to represent the conditions of an average vehicle on the network. Figure 85 shows a depiction of the provided data, including the spatial representation of the latitude and longitude points representing the individually reported movements that occur independent of a specific roadway allowing for increased accuracy to represent the location of the vehicle or event at each reported interval.</p>  <p>Source of aerial photo: Esri</p> <p>Figure 41. CV Movement Data dataset excerpt</p>
Timeliness	<p>The third-party CV datasets represent vehicle events and movements across the world with a timestamp based on Universal Time Coordinated (UTC), which allows the data to fluidly cross geographies and other boundaries. The precision of the data collected is based on determined latitude and longitude derived from either satellite connections or triangulation, allowing good precision and does not restrict the capture of this data to any singular roadway network. As this data is not a complete 100 percent capture of the traffic on the network, the capture relies on market penetration which varies by geography, roadway type, and time of day due to the number of vehicles on the network. This gap is expected to continue to close over time as</p>

Assessment Criteria	Assessment
	adoption increases. While this did not impact the use of the CV data in the use cases of this study, the collection rate variation did warrant noting.
Consistency	As these data are provided by a third-party to represent vehicle movements and driver event data, there are no consistency concerns between various data collection or vehicle manufacturers contained within the data. Any required normalization or standardization is completed by the third part and therefore provides a consistent representation of the data across geographies and time periods.
Conformity	<p>The data are highly uniform and standardized based. The CV movement and the CV driver event data have not undergone a standardization across third-party providers. This is observed in several of the more detailed attribute categories of the driver event data where certain attributes, such as road type, are described which, while consistent within the provided data, do not yet have national standards established. The third-party CV data conform to the expected standards of UTC for time and displaying latitude and longitude in a ready-to-read format in separate attribute columns.</p> <p>The data also conforms to expected data transfer using parquet and CSV, which are recognized standards and are readable by various automated data tools.</p>
Accuracy	<p>The provided latitude and longitude of the third-party CV datasets are populated directly from the vehicle or collection device and represent the location of the reported record at the time of collection. These datasets are not snapped or spatially positioned to a nearby roadway feature. This preserves the most precise location information of the provided data. This does require the user to undertake a post-process of the data if the information is needed to be snapped to a roadway network for analysis.</p> <p>The provided precision does allow for the raw location-based information to be used; however, two important considerations are noted:</p> <ol style="list-style-type: none"> 1) The beginning and ending points of the vehicle movement are provided at a higher scale to ensure anonymity of the provided vehicle. This typically includes the first and last quarter mile or less of the trip path. 2) While the precision is detailed to determine path, turns, and other movements, the current accuracy does not allow for depiction of roadway specifics, such as which lane the vehicle used on a multilane highway. The grouping of the points suggests this may be possible in the future as precision across all geographic coverage areas increases, but currently the data does not yet consistently provide this level of detail.
Integrability	The third-party CV data have potential for integrability with other datasets. The data location and time stamp information is provided in a near raw point format allowing the data to consistently represent the information across geographies regardless of the base network that is used (ARNOLD or other). The use of this data will require some additional post processing work to further perform analyses, but as this data provides consistent highly detailed data and attributes it is considered highly integrable.