Post-event Connected Vehicle Data Exploration - Lessons Learned

Travel Monitoring and Surveys
Office of Highway Policy Information
2024

U.S. Department
of Transportation

Technical Report Documentation Page

<table>
<tr><td colspan="2">1. Report No.
FHWA-HPL-24-012</td><td>2. Government Accession No.
N/A</td><td>3. Recipient's Catalog No.
N/A</td></tr>
<tr><td colspan="3">4. Title and Subtitle
Post-event Connected Vehicle Data Exploration - Lessons Learned</td><td>5. Report Date
May 2024</td></tr>
<tr><td colspan="3" rowspan="2"></td><td>6. Performing Organization Code
N/A</td></tr>
<tr><td colspan="3">7. Author(s)
Shuqing Wang, Tianjia Tang, Brian Brotsos, and David Winter</td><td>8. Performing Organization Report No.
N/A</td></tr>
<tr><td colspan="3">9. Performing Organization Name and Address
Federal Highway Administration
Office of Highway Policy Information , HPPI30
1200 New Jersey Avenue, SE
Washington, DC 20590</td><td>10. Work Unit No.
N/A</td></tr>
<tr><td colspan="3" rowspan="2"></td><td>11. Contract or Grant No.
N/A</td></tr>
<tr><td colspan="3">12. Sponsoring Agency Name and Address
Federal Highway Administration
Office of Policy and Governmental Affairs
1200 New Jersey Avenue, SE
Washington, DC 20590</td><td>13. Type of Report and Period
Technical Illustration 2023</td></tr>
<tr><td colspan="3" rowspan="2"></td><td>14. Sponsoring Highway
FHWA/HPPI-30</td></tr>
<tr><td colspan="4">15. Supplementary Notes
N/A</td></tr>
<tr><td colspan="4">16. Abstract
Traditional traffic monitoring sensors on roadways provide valuable information about travel demands and patterns. However, they do not capture detailed data on how vehicles are driven and interact with the road and the environment. To address this gap, the Office of Highway Policy Information in the FHWA has explored connected vehicle (CV) data from Wejo and the US DOT JPO Connected Vehicle pilot project. The CV data analysis is termed post-CV data analysis as the analysis is not done in site and in real time vehicles are traveling. Through this exploration, a wealth of highly desired information has been extracted from the post-CV data. This includes data on hard vehicle acceleration and deceleration accompanied by specific geospatial locations on the roadway, vehicle speed, seat belt usage, and windshield wiper status. The geolocation data is particularly significant as it helps identify areas where potential geometric and pavement inadequacies may exist. The availability of seat belt information, including when and under what conditions they are buckled, is unprecedented. The post-CV data offers not only the distance vehicle travelers buckled but also the length of time travelers buckled along with microlevel information on when and where buckle/unbuckling occurred during their journeys. To effectively utilize post-CV data, it is crucial to have a suitable platform for data storage, access, and analytics. The choice of a data platform should be based on the programming language it supports and the expertise of an organization's analysts. Given the size of the data and the potential presence of Personal Identifiable Information (PII), the focus should be on accessing and utilizing the data rather than owning it. From a cost perspective, accessing the data is more economical than owning it. It is important to acknowledge that CV data may have quality issues. While they are primarily machine-generated, they are still prone to errors. Analysts are advised to perform data quality checks before utilizing the data to ensure its reliability and accuracy.
The present paper provides an overview on post-CV data analysis related issues including its significant and unprecedent value, tools needed, expertise desired, and the awareness of potential data quality present. The goal of this paper is to encourage team and cooperative effort in acquiring and utilizing CV data to facilitate strategies for a safe and efficient highway travel.</td></tr>
<tr><td colspan="2">17. Key Words
Connected vehicle data, hard brake, sudden acceleration, seatbelt usage, geolocations, post-CV data analysis.</td><td colspan="2">18. Distribution Statement
No restrictions.</td></tr>
<tr><td>19. Security Classif. (of this report)
Unclassified</td><td>20. Security Classif. (of this page)
Unclassified</td><td>21. No. of Pages
33</td><td>22. Price
N/A</td></tr>
</table>

Notice

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for use of the information contained in this document. This report does not constitute a standard, specification, or regulation.

The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

Contents

1. Abstract

Traditional traffic monitoring sensors on roadways provide valuable information about travel demands and patterns. However, they do not capture detailed data on how vehicles are driven and interact with the road and the environment. To address this gap, the Office of Highway Policy Information in the FHWA has explored connected vehicle (CV) data from Wejo and the US DOT JPO Connected Vehicle pilot project. The CV data analysis is termed post-CV data analysis as the analysis is not done in site and in real time vehicles are traveling. Through this exploration, a wealth of highly desired information has been extracted from the post-CV data. This includes data on hard vehicle acceleration and deceleration accompanied by specific geospatial locations on the roadway, vehicle speed, seat belt usage, and windshield wiper status. The geolocation data is particularly significant as it helps identify areas where potential geometric and pavement inadequacies may exist. The availability of seat belt information, including when and under what conditions they are buckled, is unprecedented. The post-CV data offers not only the distance vehicle travelers buckled but also the length of time travelers buckled along with microlevel information on when and where buckle/unbuckling occurred during their journeys. To effectively utilize post-CV data, it is crucial to have a suitable platform for data storage, access, and analytics. The choice of a data platform should be based on the programming language it supports and the expertise of an organization's analysts. Given the size of the data and the potential presence of Personal Identifiable Information (PII), the focus should be on accessing and utilizing the data rather than owning it. From a cost perspective, accessing the data is more economical than owning it. It is important to acknowledge that CV data may have quality issues. While they are primarily machine-generated, they are still prone to errors. Analysts are advised to perform data quality checks before utilizing the data to ensure its reliability and accuracy.

The present paper provides an overview on post-CV data analysis related issues including its significant and unprecedent value, tools needed, expertise desired, and the awareness of potential data quality present. The goal of this paper is to encourage team and cooperative effort in acquiring and utilizing CV data to facilitate strategies for a safe and efficient highway travel.

2. Introduction

The Travel Monitoring and Surveys division within the FHWA Office of Highway Policy Information (HPPI) gathers data related to traffic flows, flow patterns and travel behaviors. On the traffic flow front, the data collection includes traffic volume, vehicle classification, speed, and vehicle axle weight, acquired through sensors maintained by State highway agencies. However, due to the limited availability of these sensors, the collected data primarily exists at a macro-level, necessitating statistical adjustments to accurately represent specific roads or regions. These datasets offer invaluable insights into understanding travel demands and overall productivity and serve as essential resources for planning, safety analysis, and policy evaluation purposes.

While the embedded pavement and roadside sensors offer valuable insights, they lack intricate micro-level details concerning vehicles, drivers, and their interactions with the road. Information regarding acceleration, deceleration, speed consistency, and seatbelt usage is notably absent. Availability of such micro-level data could facilitate integrated analysis, pinpointing geospatial areas associated with frequent hard braking and acceleration and understanding their correlation with drivers and roadway conditions, and ultimately facilitating solutions to make travel safer and more efficient.

The emergence of Connected Vehicles (CV) has revolutionized transportation safety and efficiency by enabling the real-time exchange of operational statuses and geospatial information among vehicles and infrastructure. While real time in-situ CV data usage are the bottom-line for CV operations, archived CV data termed post-CV data may also prove valuable. The authors gained an opportunity to analyze post-CV data from four OEMs and the US DOT Intelligent Transportation System Joint Program Office Connected Vehicle Pilot projects (JPO Pilot). The analysis of the post-CV data demonstrates the capability of post-CV data to bridge the gap in micro-level vehicle travel data. This post-CV data offers otherwise nonexistent information on enhancing highway infrastructure planning, safety measures, and operational enhancements.

This article intends to introduce the basic concept of post-CV data to analysts and managers, offering insights into the fundamentals of CV data, available variables, essential data platforms, tools, and other spectrum of information that can be derived. It is hoped that this article will encourage further exploration of post-CV data usage to enhance travel safety and efficiency through team effort and broad cooperation among different offices and entities.

3. CV Data and Size

CV data emerges from vehicles communicating with each other and infrastructure, enhancing travel safety and operational efficiency by exchanging safety and mobility information. CV data encompasses vehicle and infrastructure-generated data, detailing vehicle and infrastructure states, including traffic control devices.

CV data poses unique challenges due to its size. The three JPO Pilot projects (Tampa Hillsborough Expressway Authority, I-80 in Wyoming, and New York City) generate over 33 GB per day. Florida's four OEM CV vehicles, representing less than 10% of its registered 19 million vehicles, generate over 60 GB of data in a single day with more than a billion data records. Extrapolating, a year's coverage in Florida could produce data over 21 terabytes.

For refence and comparison purpose, FHWA's National Bridge Inventory (NBI) annual data is approximately 1.5 GB and the Highway Performance Monitoring System (HPMS) annual data is only about 4 GB.

> CV data demands substantial storage and bandwidth for transmission. Decision makers need to be aware of the data size challenge.

CV Data Variables

Understanding the CV data variables is crucial as these variables determine the extractable information. The OEM CV data comprises approximately 60 variables, split into vehicle movement and driving event datasets. The vehicle movement data set has information on the journey of a vehicle's movement with geolocations and speed, heading, and ignition state with timestamps. The driving event dataset records various events along with corresponding geolocations, speeds, headings, and timestamps.

The following is a list of vehicle movement data variable names:

- journey_id
- datapoint_id
- captured_time_local
- captured_time_utc
- local_captured_date
- geohash
- latitude
- longitude
- state_code
- country_code
- postal_code
- heading
- speed
- ignition_state

The following is a list of driving event data variable names:

- acceleration_type
- anti_lock_braking_system_status
- autonomous_emergency_braking_type
- captured_time_local
- captured_time_utc
- country_code
- datapoint_id
- door_identifier
- door_status_change_type
- electronic_stability_status
- event_type

- exterior_temperature
- fuel_consumption
- fuel_level
- geohash
- heading
- ignition_state
- journey_event_type
- journey_id
- lateral_acceleration
- latitude
- light_identifier
- light_state_change_type
- local_captured_date
- longitudal_acceleration
- longitude
- odometer
- parking_brake_identifier
- parking_brake_status_change_type
- postal_code
- seat_identifier
- seat_occupancy_status
- seatbelt_status
- seatbelt_warning_status_change_type
- signal_identifier
- signal_state_change_type
- speed
- speed_threshold_change_type
- state_code
- wiper_identifier
- wiper_interval
- wiper_state_change_type

Data variables are reasonable indications of a dataset's value. Data variables determine what information can be extracted from the data.

The variable event_type has the following critical choices:

Acceleration_Change, Anti_Lock_Braking_System_State_Change, Autonomous_Emergency_Braking_Change, Door_State_Change, Electronic_Stability_State_Change, Journey, Light_State_Change, Parking_Brake_State_Change, Seat_Belt_Change, Seat_Belt_Warning, Seat_Occupancy_Change, Signal_State_Change, Wiper_State_Change, Speed_Threshold_Change

All event data include their corresponding geolocations, speeds, headings, fuel levels, odometer readings, and timestamps when events occurred.

The JPO Pilot CV data, obtained through custom-installed safety devices (On-Board Unit) on pilot vehicles and roadside units (RSU), include Basic Safety Messages (BSM), Traveler Information Messages (TIM), Signal Phase and Timing (SPaT), and EVENT categories.

The pilot data are organized into four key category files: the basic safety message (BSM), traveler information messages (TIM), the signal phase and timing (SPaT), and EVENT. BSM contains each acceleration/deceleration point in 3 axes (longitudinal, lateral, and vertical acceleration) plus the yaw rate. TIM contains roadside sign information. EVENT is a log collection of BSM, TIM and SPaT with time and location obfuscated. This CV pilot data analysis focuses on BSM and TIM.

While the JPO CV data published follow the SAE International Surface Vehicle Standard J2540, J2735 and J2945 specifications, there are significant differences between these data files regarding data variables and the exact meaning of such variables. Data variables, format, structures, data availability and units vary between pilot sites or even between different RSUs. Also, unlike the commonly used relational tabular data model, the JPO pilot post-CV data takes

a hybrid approach to represent and save its complex data variables. At the table level, it saves each data point as a regular record. At the record level however, data are saved in hierarchical manner as the JSON format. Both the BSM and TIM contain many data variables (over 60). Certain variables such as PathHistoryPoint of BSM and SEQUENCE_item_itis of TIM are array type, containing multiple entries.

The following is a list of BSM data variables:

- RSUID
- recordGeneratedBy
- recordGeneratedAt
- coreData_id
- coreData_secMark
- coreData_lat
- coreData_long
- coreData_elev
- coreData_speed
- coreData_heading
- coreData_angle
- coreData_accelSet_long
- coreData_accelSet_lat
- coreData_accelSet_vert
- coreData_accelSet_yaw
- PathHistoryPoint_latOffset
- PathHistoryPoint_lonOffset
- PathHistoryPoint_elevationOffset
- PathHistoryPoint_timeOffset

The following is a list of TIM data variables:

- RSUID
- recordGeneratedBy
- recordGeneratedAt
- TravelerDataFrame_frameType_ roadSignage
- TravelerDataFrame_msgId_ roadSignId_position_lat
- TravelerDataFrame_msgId_ roadSignId_position_long
- TravelerDataFrame_msgId_ roadSignId_position_elevation
- TravelerDataFrame_msgId_ roadSignId_viewAngle
- TravelerDataFrame_msgId_ roadSignId_mutcdCode
- TravelerDataFrame_content_speedLimit_SEQUENCE_item_itis
- TravelerDataFrame_content_ advisory_SEQUENCE_item_itis

These BSM and TIM data variables listed above are only the key variables used during the analyses. The names shown are simplified from their hierarchical levels. For more information about this pilot program and CV data, readers may visit https://datahub.transportation.gov/stories/s/Connected-Vehicle-Pilot-CVP-Open-Data/hr8h-ufhq/.

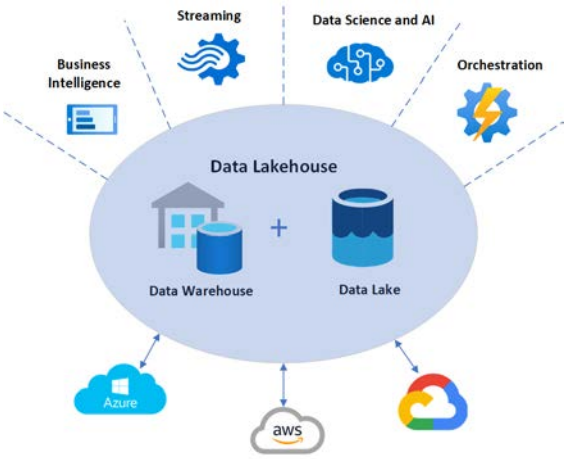
4. Platform to Conduct Post-CV Data Analysis

Data platforms, including data storing, accessing and analytical tools, are critical to carry out efficient and productive data analysis, especially for post-CV data due to the size of the data, complex data structures, and heterogeneous data formats.

4.1: Data Storage Concept

Currently, the practice of complex data storage is through the so-called Data Lakehouse architecture. As shown in Figure 1, a Data Lakehouse has two critical parts: Data Warehouse and Data Lake, reflecting the long evolution history of data storing technologies. A Data Warehouse handles traditional structured relational data from all sources but with the same data format. A Data Lake on the other hand provides the capability to store data structured, semi-structured and unstructured data coming in all data formats such as JSON, CSV, Parquet. etc.

Figure 1: Databricks architecture illustration



Source: FHWA Office of Highway Policy Information.

4.2: Data Analytics

Based on the Data Lakehouse architecture, the big data industry has adopted the open analytics platform known as the Databricks Lakehouse for building, deploying, sharing enterprise level data, analytics, and AI solutions at scale. It integrates many data processing components such as Apache Spark, a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters, and MLFlow, an open-source platform for machine learning. The Databricks Lakehouse have been customized by many companies including Microsoft (Azure Databricks), Amazon (Amazon Databricks), and Google (Google Databricks) to provide commercial deployments in different cloud environments. In addition to Databricks Lakehouse, there are some other alternative platforms such as Snowflakes and Cloudera. They provide data platforms as Software-As-A-Service (SaaS) with each strengthening on different aspects of data storing, accessing, and processing.

The abovementioned big data platforms offer broad similar capacities and compete with their unique specificities. When deciding on a platform, the programming languages a platform supports should be a pivotal factor. Primary programming languages may include Python, SQL, R, Scala, and Java. Agency users' familiarity and knowledge with any supported programming language are critical to utilizing a platform's capacity and achieving the information extraction goal.

Primary programming languages (e.g., Python, SQL, R, Scala, and Java) proficiency is one of the factors in platform selection.

The OEM CV data provider uses the AWS Databricks Lakehouse platform. The authors analyzed the CV data via an evaluation Databrick account using self-developed Databricks codes in Python and SQL.

The JPO Pilot post-CV data are stored on the AWS cloud as an S3 Bucket. The authors analyzed the pilot CV data via FHWA Turner Fairbank Highway Research Center's Path to Advancing Novel Data Analytics (PANDA) laboratory AWS Databricks platform. Unlike the OEM CV data that is structured and stored in relational database tables, JPO Pilot post-CV data are unstructured and saved as many JSON files in S3 Bucket. For each Roadside Unit, data of each hour is saved as one JSON file, which results in a tremendous number of individual files. The author developed a process in Databricks to automatically enumerate all daily and weekly files and loaded each JSON file into Apache Spark data frames for further processing. Another issue in the Pilot CV data is data format. Data types and even data units are different from one file to another. A set of Python codes developed resolved all the format issues.

More platforms an organization having accesses to does not necessarily mean more productivity or higher efficiency.

5. Data Insights Explored

5.1: Identification of Vehicle Hard Brake and Acceleration Geolocations

CV data provide insights into vehicle acceleration and deceleration events, pinpointing specific geolocations with timestamps. While a single individual event may not highlight potential infrastructure or operational issues, clusters of such events with high frequency of occurrence in specific roadway areas could signal the potential of infrastructure inadequacies such as, curvature, grade, sight distance, and pavement conditions. In addition, operating challenges, such as congestion and varying speeds among vehicles, may also be potential issues.

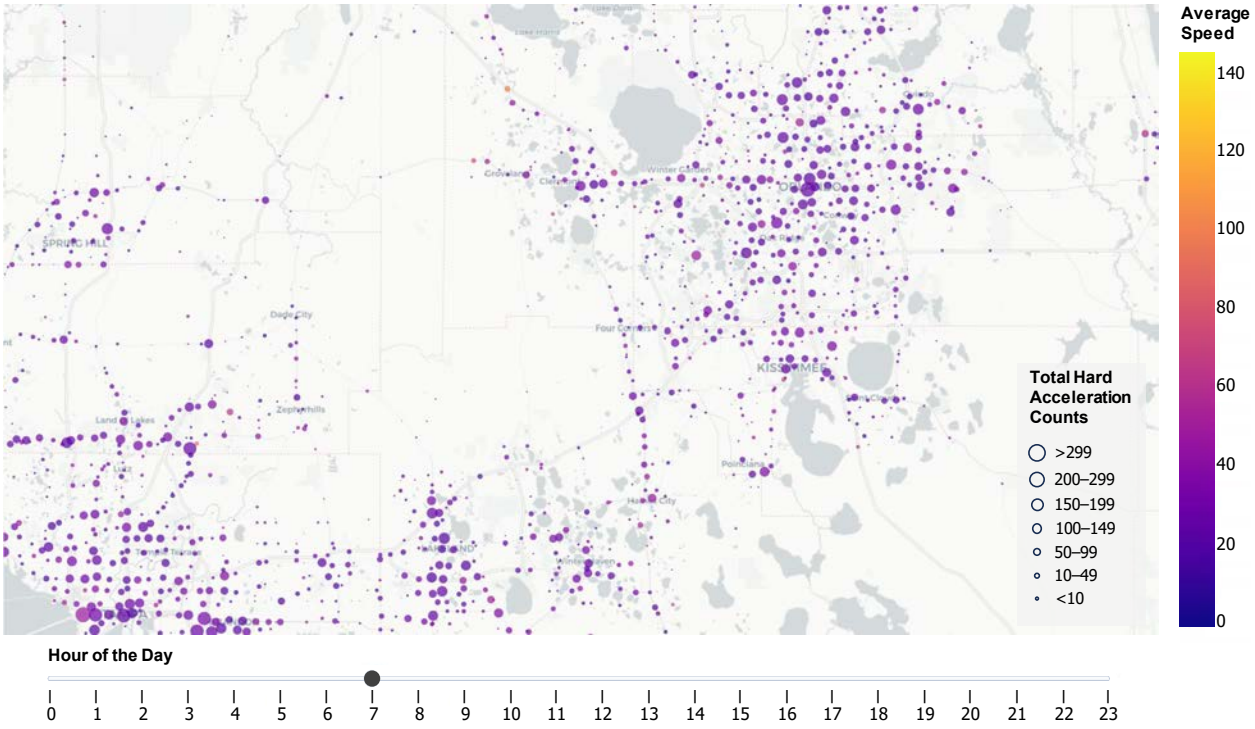
5.1.a: OEM CV Data

The OEM post-CV data explored have brake and acceleration event data only for these events exceeding a threshold value associated with longitudinal maneuvers. While the OEM's specification of a hard brake is not available, AASHTO's A Policy on Geometric Design of Highways and Streets specifies a maximum of deceleration of 3.4 m/s^2 for sight distance computation.

Several approaches were used to analyze the OEM post-CV brake maneuver data. The first one plotted out all individual events on a roadway GIS map. This enables direct observation of such events visually. While this approach offers visual cues on geolocations related to where such events are, it lacks a quantitative measure (e.g., number of events per unit area or unit roadway length) for systematic comparison and ranking. The second approach developed to overcome the issue was through density computation for each predefined zone. In this case, a 2 kilometer (km) by 2 km square (4 km^2) zone is defined and used for Florida. The entire state was divided into many thousands of such square zones. Events that occurred in each zone were counted and the density of such events was visually displayed. Obviously, the size of the zone can be structured in any size analyst's desire. The zone establishment facilitates event frequency computation and identification of high frequency locations. The third approach was computing event density per roadway length. To enable comparison on a common base, a fourth approach can further be developed to divide the third approach result by their corresponding AADT values. In other words, the measure is the result of the number of events per AADT for a unit length of roadway segments. This normalization process has enabled direct comparisons across different roadways.

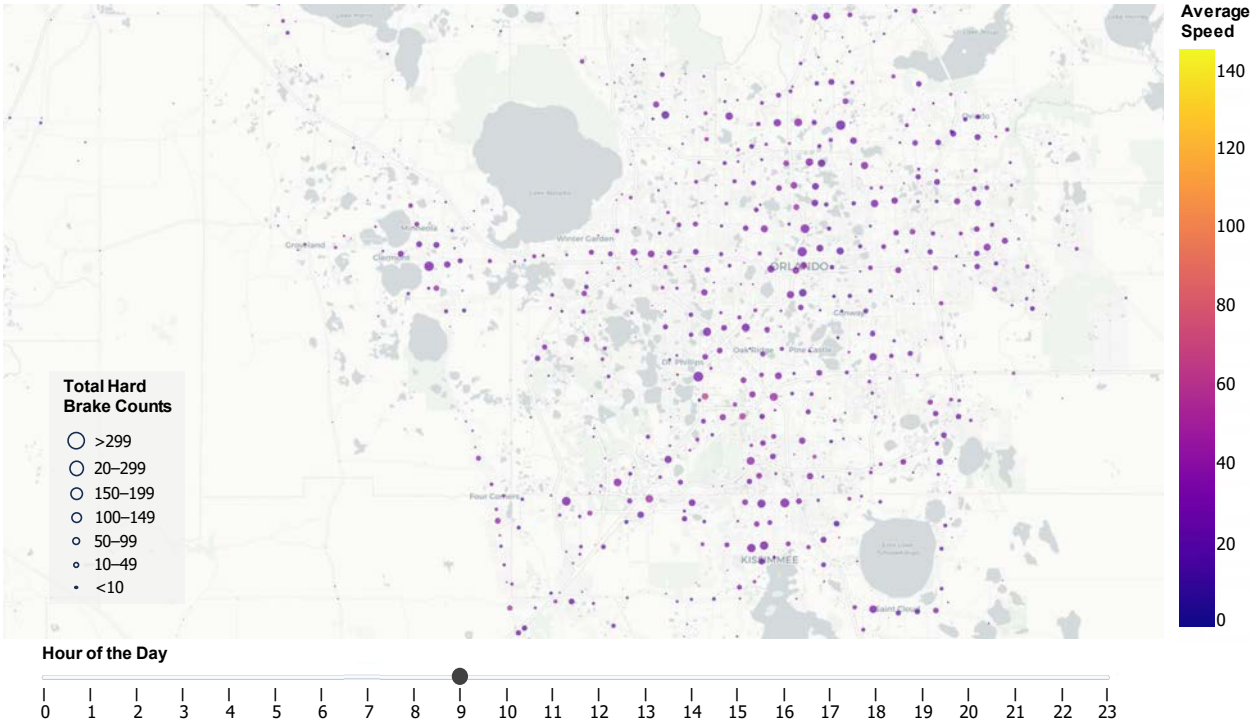
Figures 2 and 3 illustrate the zonal hard acceleration and hard brake event densities, respectively. The bigger the dot, the higher the number of events that occurred in the 4 km^2 zone during the hour. Visual maps like these help analysts expedite the geolocating of high frequency locations. In addition, geolocation density data can be further analyzed by hour of the day together with average speeds. Figure 4 illustrates the hard brake event counts and average speed distribution of the highest frequency area over hours of the day by solid blue and hatched yellow bars, respectively.

Figure 2: Illustrating hard acceleration densities in 4 square km by hour of the day and average speed



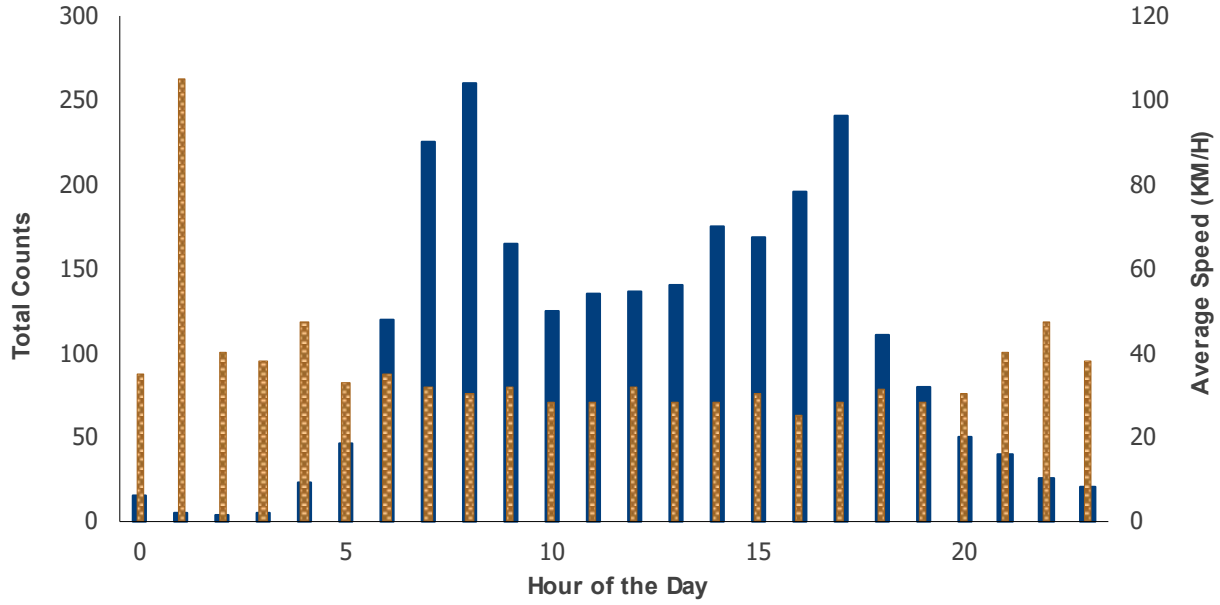
Source: FHWA Office of Highway Policy Information.

Figure 3: Hard brake densities per 4 square km by hour of the day and average speed



Source: FHWA Office of Highway Policy Information.

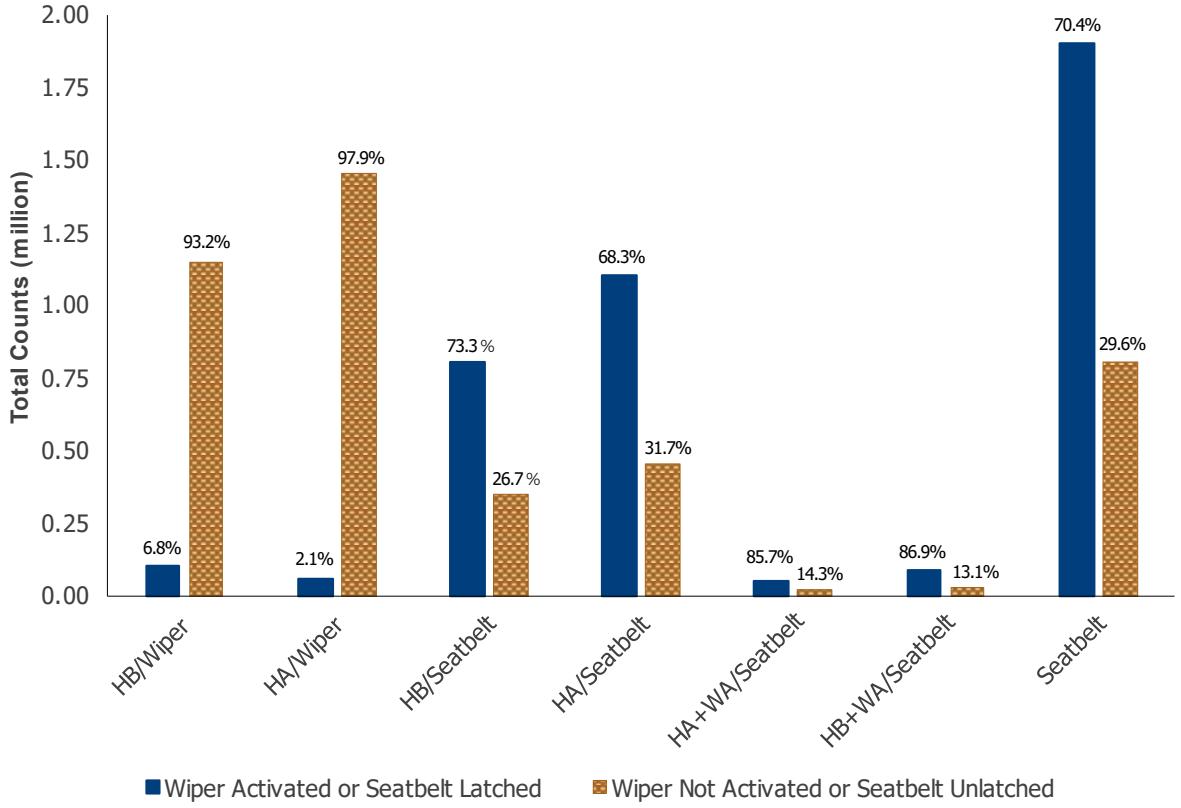
Figure 4: Hard brake counts and average speed in the highest frequency area (4 km²) by hour of the day



Source: FHWA Office of Highway Policy Information.

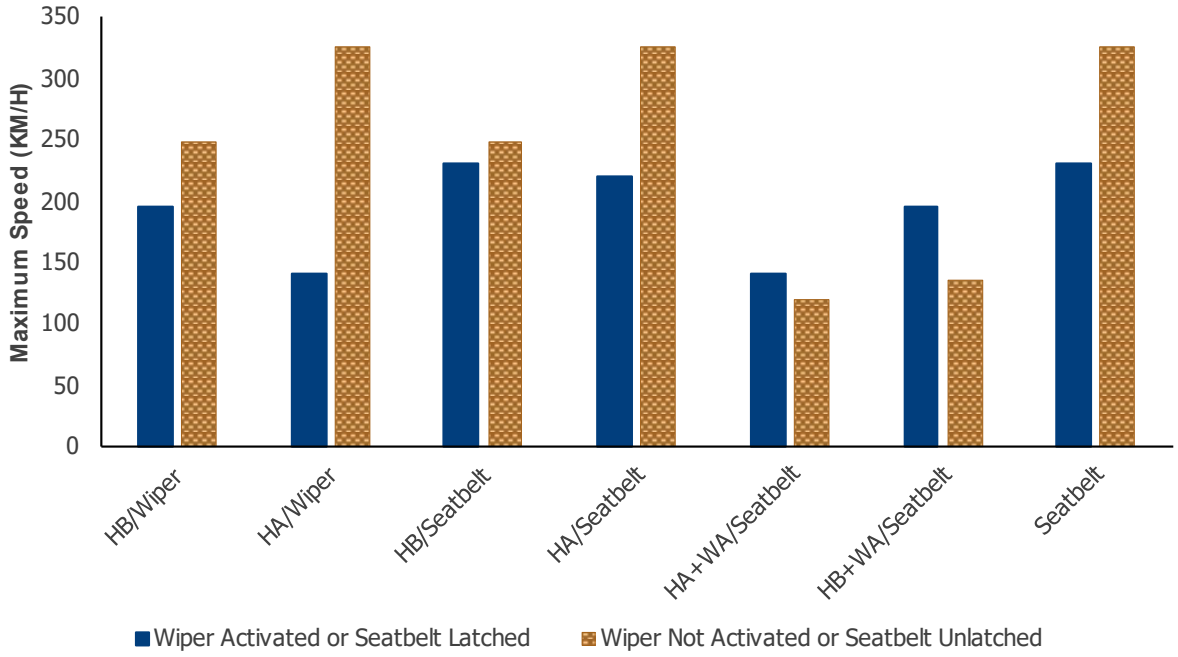
CV data events can be durational and transitional. For example, seatbelts latched/unlatched, wipers on/off, headlight on/off are durational as they typically sustain a long period of time. Acceleration and brake are transitional as they would sustain for only a short period of time. Acceleration and brake analysis can be integrated with seatbelt, wiper, and headlight status, yielding additional information. Figures 5 and 6 shows hard brake and acceleration counts and maximum speed for each event combination.

Figure 5: Counts of HA and HB events by wiper and seatbelt states



Source: FHWA Office of Highway Policy Information.

Figure 6: Maximum speed of HA and HB events by wiper and seatbelt states



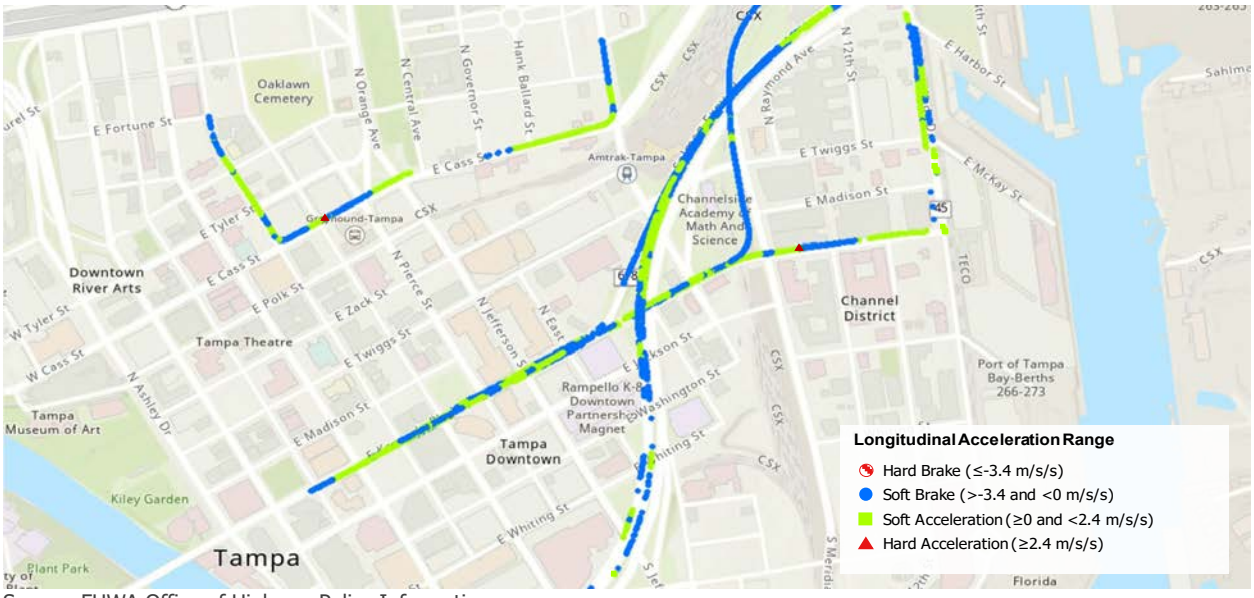
Source: FHWA Office of Highway Policy Information.

5.1.b: JPO CV Pilot Data

Unlike the OEM CV acceleration or deceleration event data, the JPO Pilot post-CV data contains all acceleration and deceleration maneuvering data covering 3 axes (longitudinal, lateral, and vertical) plus yaw rate. This analysis focuses only on longitudinal and lateral (turning) movements. The lateral acceleration data offer additional information related to vehicle rollover tied with roadway superelevation, curvature, and pavement condition. With the JPO Pilot post-CV data exploration, hard acceleration for longitudinal maneuvers is defined as acceleration equal to or exceeding 3.4 m/s^2. Excessive lateral acceleration is defined as 2.4 m/s^2 (0.25g).

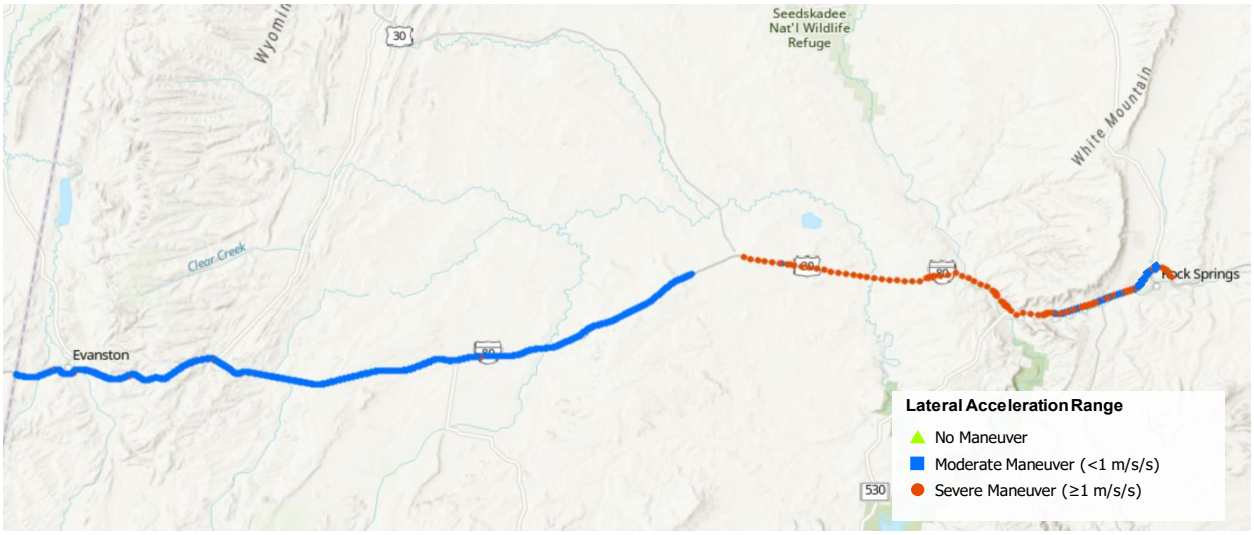
Figure 7 shows different maneuvers based on longitudinal acceleration states. And Figure 8 shows different maneuvers based on lateral acceleration states.

Figure 7: Longitudinal acceleration points with different maneuver categories



Source: FHWA Office of Highway Policy Information.

Figure 8: Lateral acceleration points with different maneuver categories



Source: FHWA Office of Highway Policy Information.

5.2: Seatbelt Usage Exploration

Seatbelts save lives. While seatbelt usage data lays the foundation for safety work, seatbelt data gathering has always been challenging as the predominant survey method lacks microlevel information. For example, few drivers could tell when they are buckled, unbuckled, or re-buckled during a trip, and percentages of travel time and travel distance that they buckled or unbuckled.

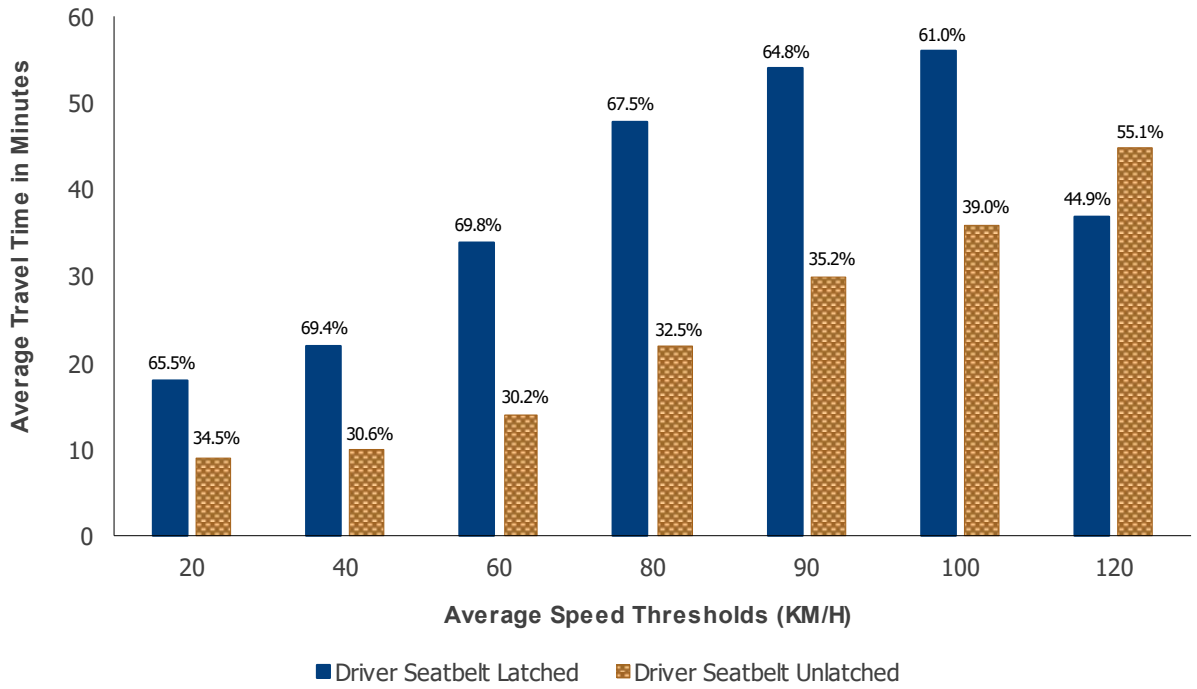
CV event data covers seatbelt usage, enabling the analysis of seatbelt latching status by time of the day, day of the week, and month of the year. In addition, CV data offer information on seatbelt usage status based on vehicle miles traveled (VMT) and vehicle hours traveled (VHT), vehicle speed, and journey progress (e.g., the start of a trip, the end of a trip).

5.2.a: OEM CV Seatbelt Data

Seatbelt status includes latched and unlatched while a vehicle's engine is on for drivers and front passengers. A simple count of latched and unlatched events can produce a summary of basic parameters, such as percentages of latched and unlatched in terms of the number of such activities. Frequency alone information on seatbelt latched or unlatched status is not sufficient to measure seatbelt usages. For example, a single latch/unlatch action does not necessarily indicate that the seatbelt is used less than multiple latch/unlatch actions. To provide a fuller picture, seatbelt latched/unlatched event data was transformed into seatbelt latched/unlatched state information with associated trips. This transformation enables the analysis of seatbelt usage by VMT, VHT, speed, and vehicle journey progress.

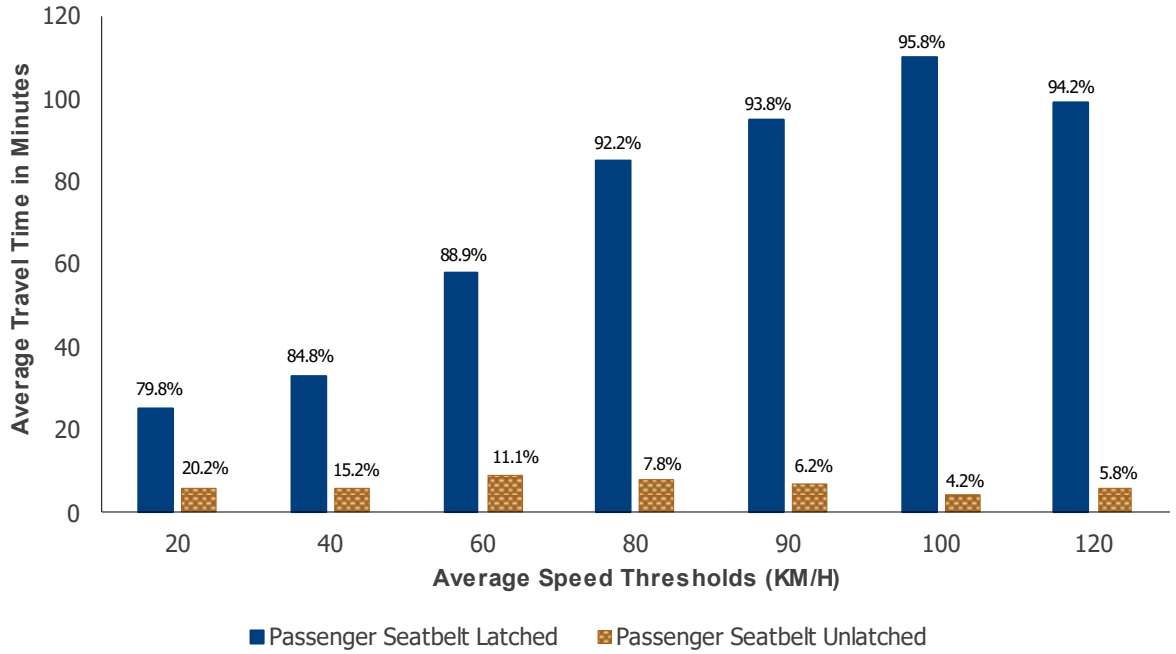
Figures 9 and 10 show the driver and front passenger seatbelt latched/unlatched durations by average travel speed, respectively.

Figure 9: Driver seatbelt usage by average speed thresholds



Source: FHWA Office of Highway Policy Information.

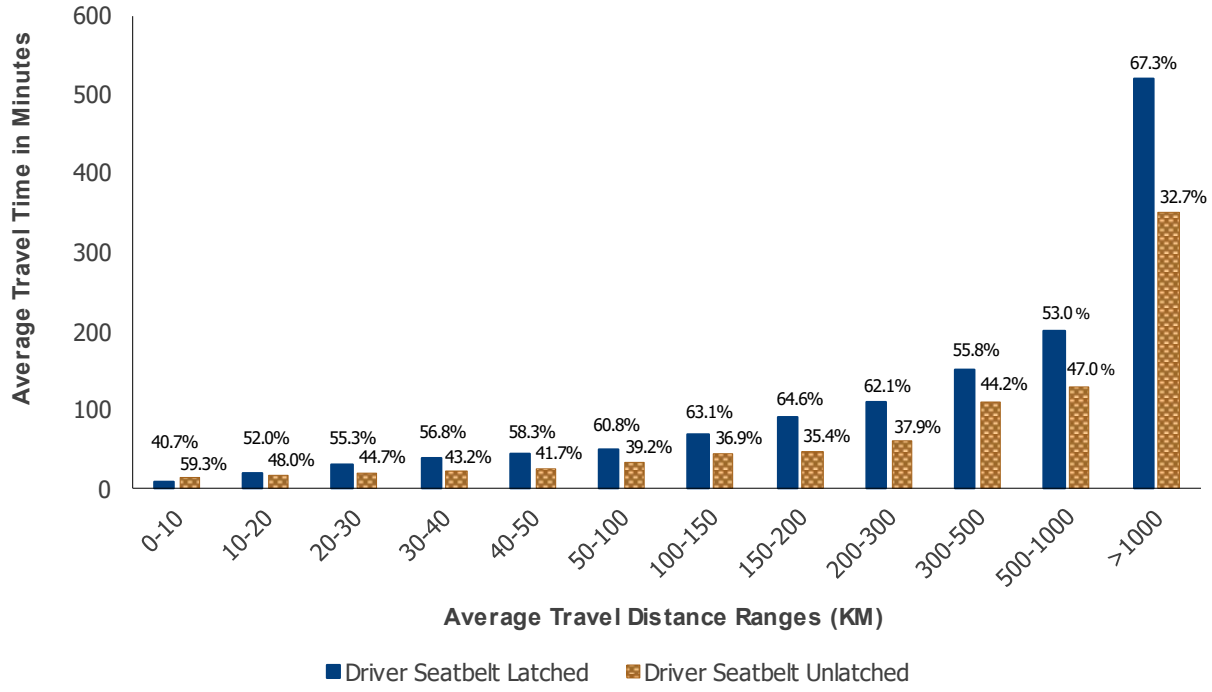
Figure 10: Passenger seatbelt usage by average speed thresholds



Source: FHWA Office of Highway Policy Information.
Note: Unoccupied passenger journeys excluded.

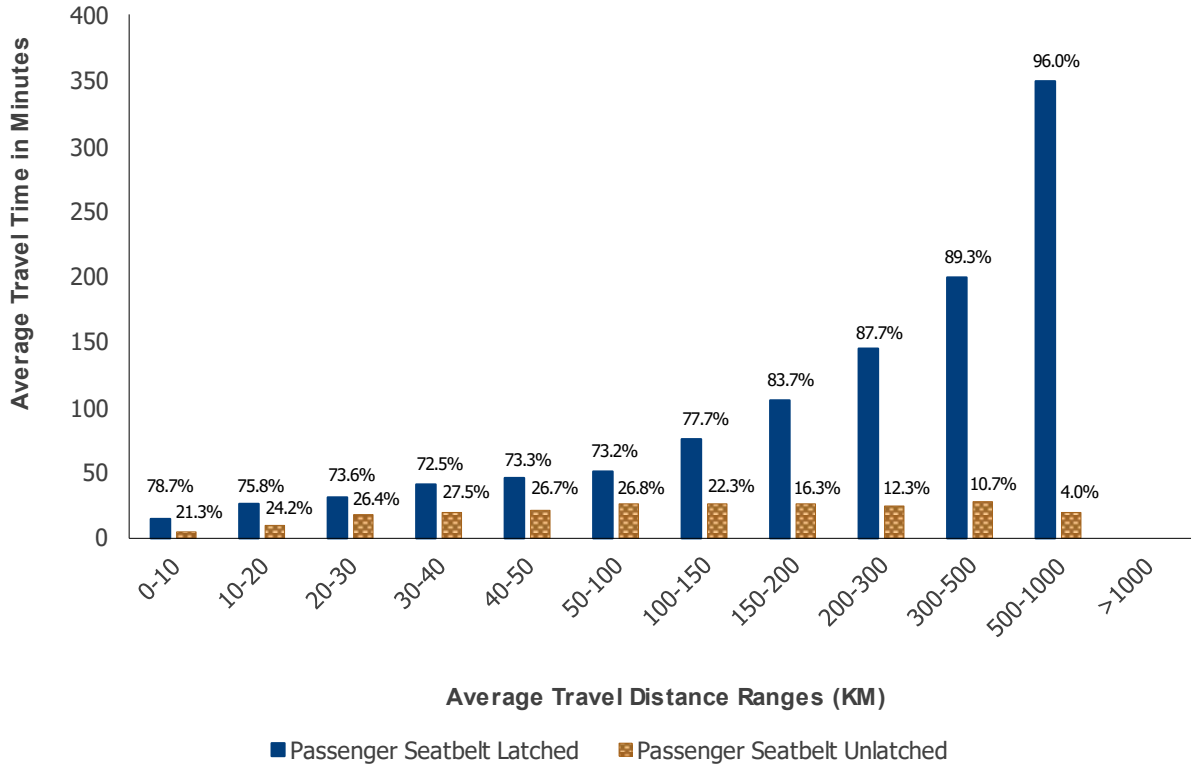
Similarly, Figures 11 and 12 show the driver and front passenger seatbelt durations by average travel distances, respectively.

Figure 11: Driver seatbelt usage by travel distances



Source: FHWA Office of Highway Policy Information.

Figure 12: Passenger seatbelt usage by travel distances



Note: Unoccupied passenger journeys excluded.
Source: FHWA Office of Highway Policy Information.

Tables 1 and 2 show the percentages of seatbelt usages by average travel speed thresholds and travel distances, respectively.

Table 1: Seatbelt usages in percentage by average travel speed thresholds

Seatbelt	Average Travel Speed Thresholds (KM/H)						
	20	40	60	80	90	100	120
Driver	65.5	69.4	69.8	67.5	64.8	61	44.9
Passenger	79.8	84.8	88.9	92.2	93.8	95.8	94.2

Table 2: Seatbelt usages in percentage by travel distances

Seatbelt	Travel Distances (KM)											
	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 100	100 - 150	150 - 200	200 - 300	300 - 500	500 - 1000	> 1000
Driver	40.7	52	55.3	56.8	58.3	60.8	63.1	64.6	62.1	55.8	53	67.3
Passenger	78.7	75.8	73.6	72.5	73.3	73.2	77.7	83.7	87.7	89.3	96	NA

5.2.b: JPO CV Pilot Data

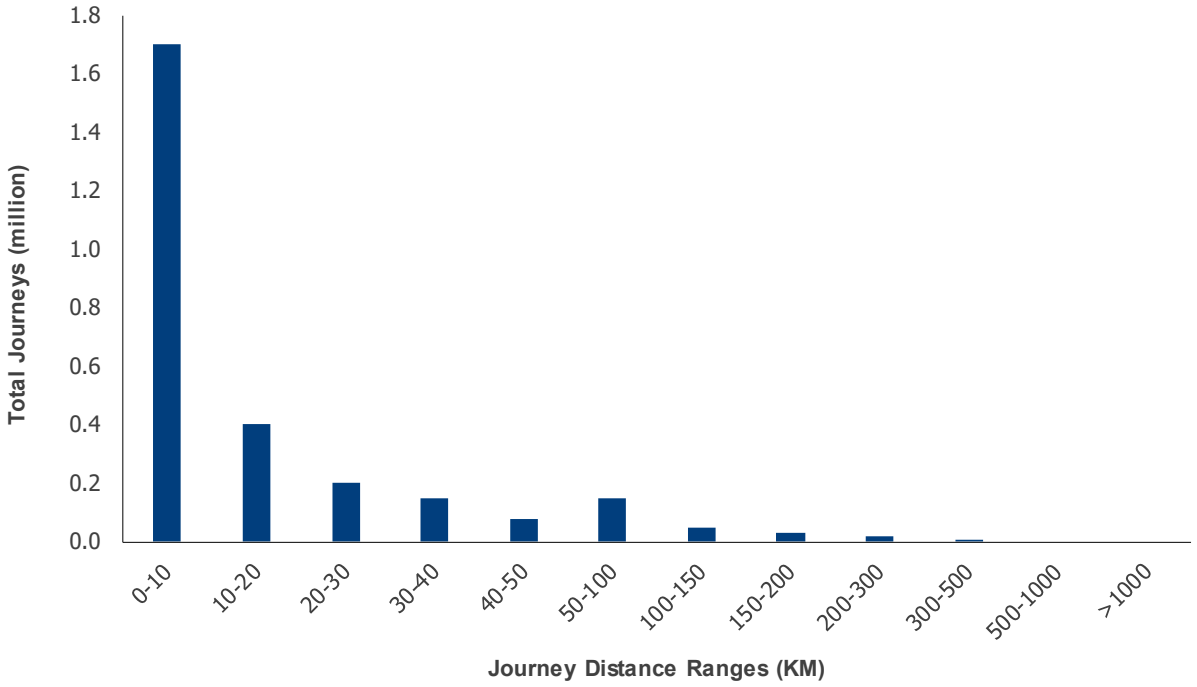
Seatbelt data associated with the JPO Pilot CV data is not available.

5.3: Trip Distribution by Length

5.3.a: OEM CV Data

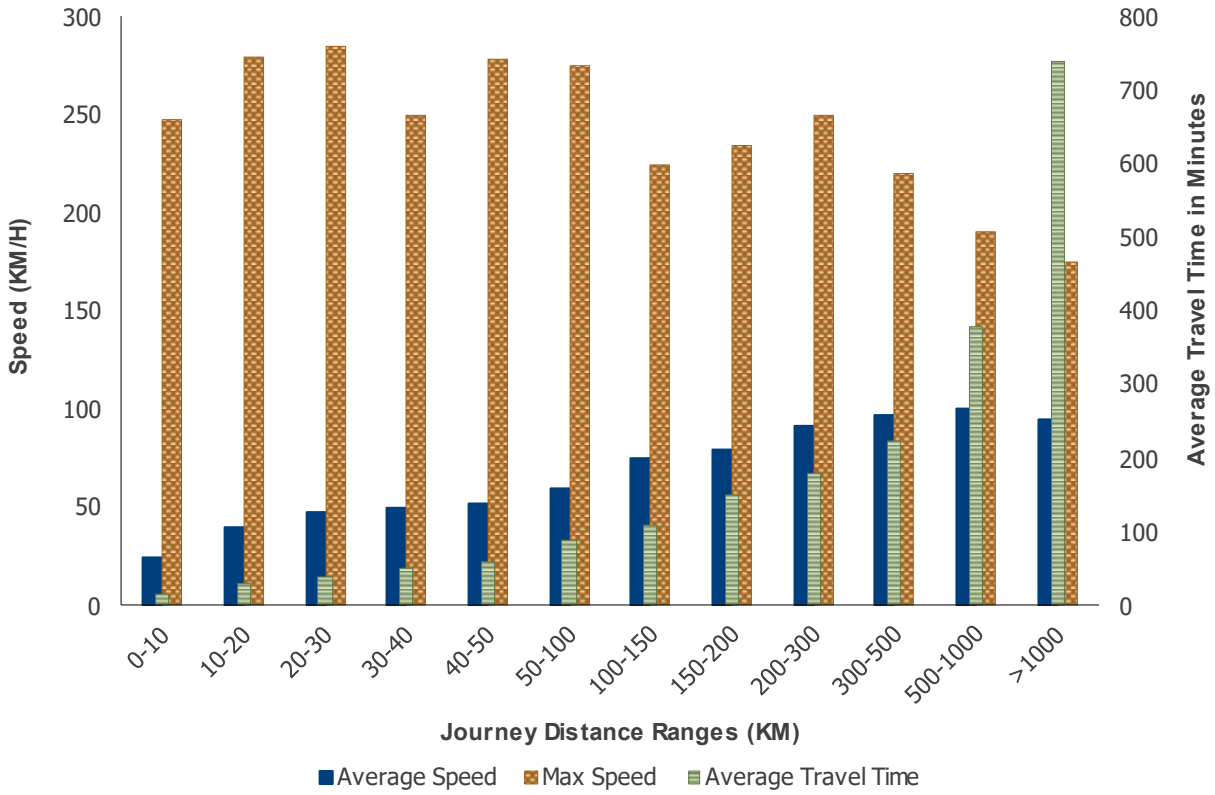
Trip distribution by trip length information is critically needed for travel demand modeling associated with transportation planning. The post-CV data analysis of the OEM CV data offers such information as illustrated in Figures 13-15. Due to privacy considerations, OEM CV data do not provide trip information, which is defined by the starting and ending points of travel. Instead, it provides only journey information, which is defined by the points where a car's engine starts and stops. Figure 13 shows the total number of journeys by travel distance. From the distribution, we can understand that journeys of less than 10 KM are predominant. Figure 14 shows the average speed, maximum speed, and average travel time for different travel distances. Figure 15 shows the journey frequency at which the front passenger seat is occupied.

Figure 13: Number of journeys by travel distances



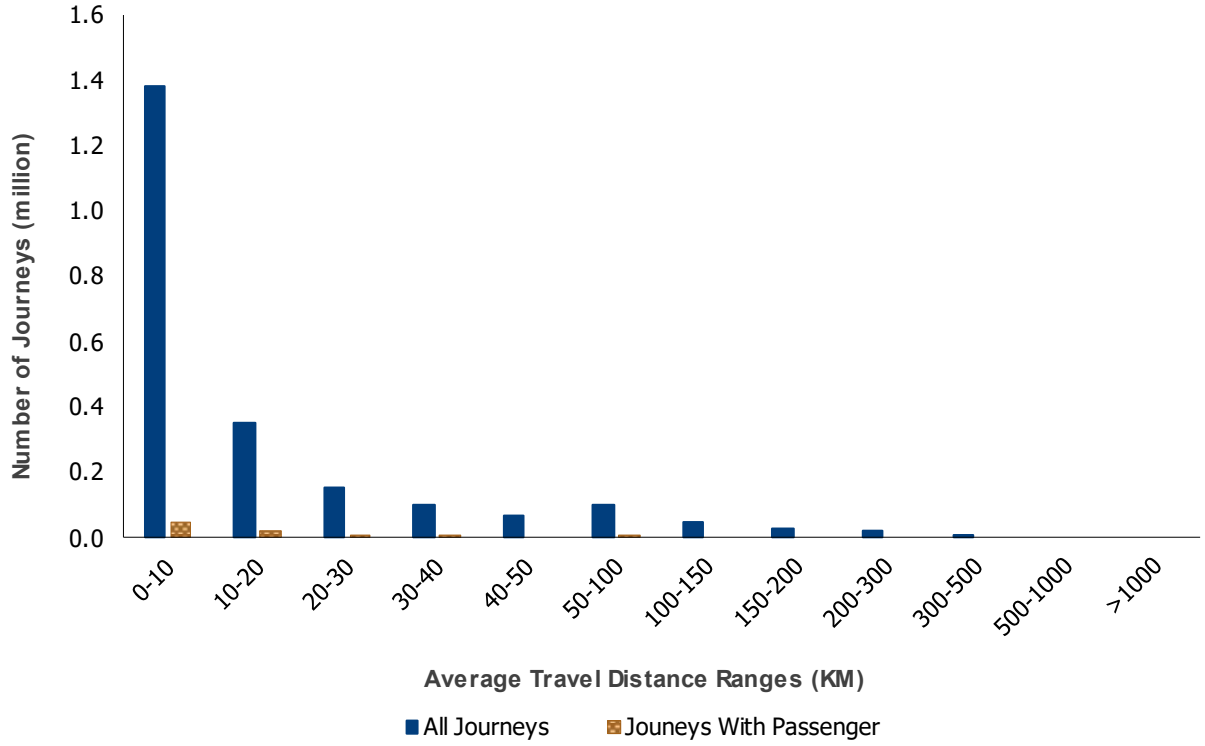
Source: FHWA Office of Highway Policy Information.

Figure 14: Travel speed and travel time by travel distances



Source: FHWA Office of Highway Policy Information.

Figure 15: Frequencies of journeys occupied with front passenger by travel distances.



Source: FHWA Office of Highway Policy Information.

5.4: Posted Speed Limits vs. the 85th Actual Travel Speed

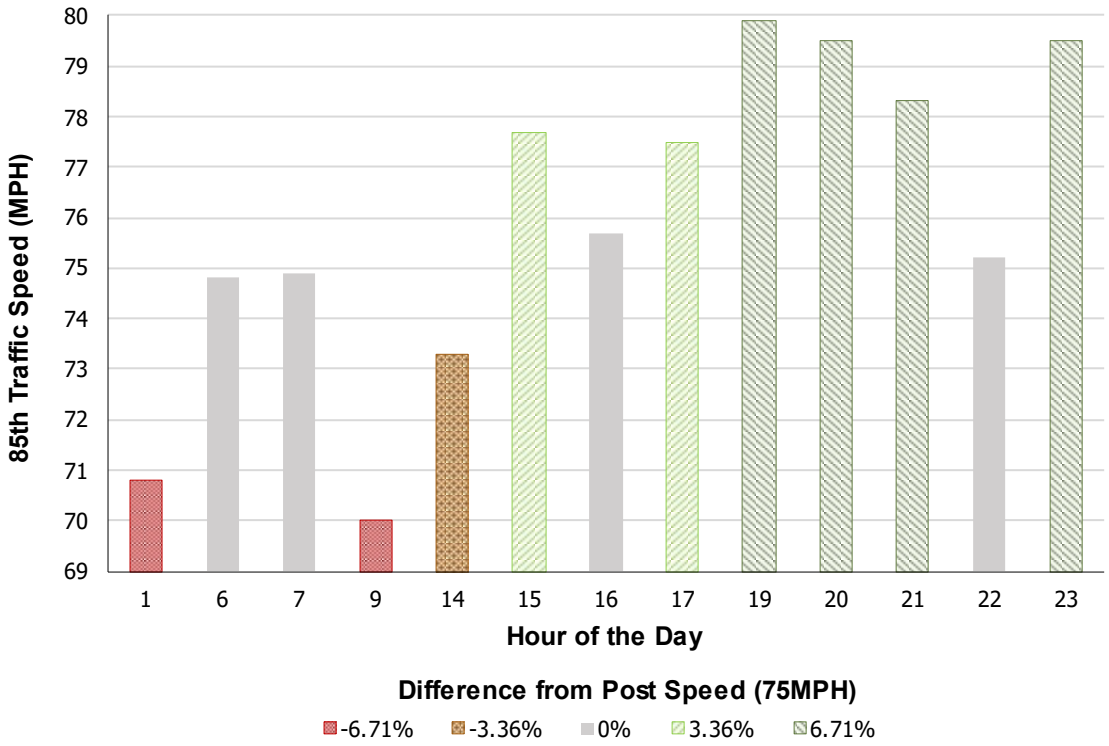
5.4.a: JPO Pilot Post-CV Pilot Data

JPO Pilot post-CV TIM data file has roadside speed limit sign information, which is broadcasted to the connected vehicles. The BSM contains actual vehicle travel speed. By conflating the speed limit sign data and the actual vehicle speed information with roadway segments, the exploration further analyzed the similarities between the posted speed limits and the 85th percentile actual speeds. Figure 16 illustrates the results of the analysis along a segment of Interstate 80. Figure 17 illustrates the 85th percentile speed and the posted speed by hour of the day for a sample roadway segment.

Figure 16: Speed differences between the speed limits in TIM and the 85th percentile of actual speeds

Source: FHWA Office of Highway Policy Information.

Figure 17: Actual 85th percentile traffic speed vs posted speed by hour of the day.



Source: FHWA Office of Highway Policy Information.

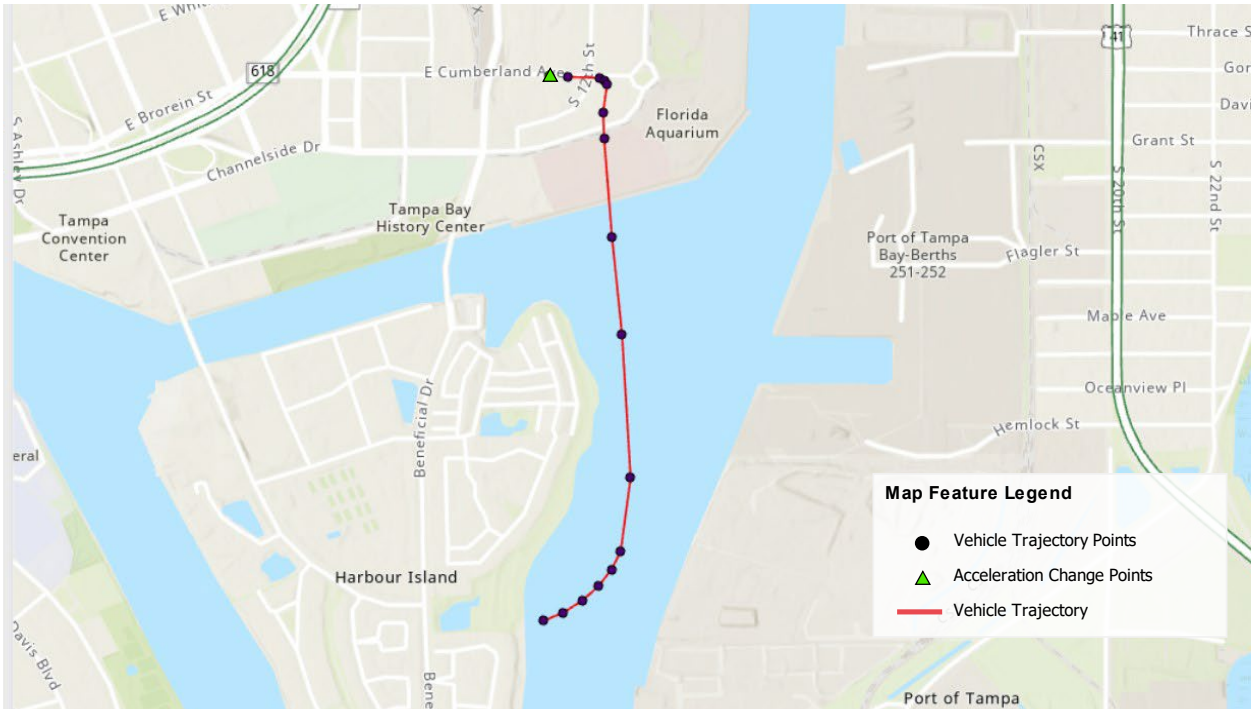
5.5: Roadway Curvature and Frequency of Vehicle Maneuvers

The BSM data is used to count the frequencies of vehicle maneuvers characterized by acceleration and deceleration changes. These maneuver changes are integrated into the roadway geospatial alignment data. The roadway curvature analyses focused only on JPO's Pilot WYDOT data that has coverage of many different curvature roadway segments. The hypothesis that the sharper a roadway curve is, the more maneuvers a driver may perform is not observed, though. The limited data set, for example, shows that the sharper a roadway curve is, the steadier the driver (fewer maneuvers) maintains its vehicle operation.

6. CV Data Quality

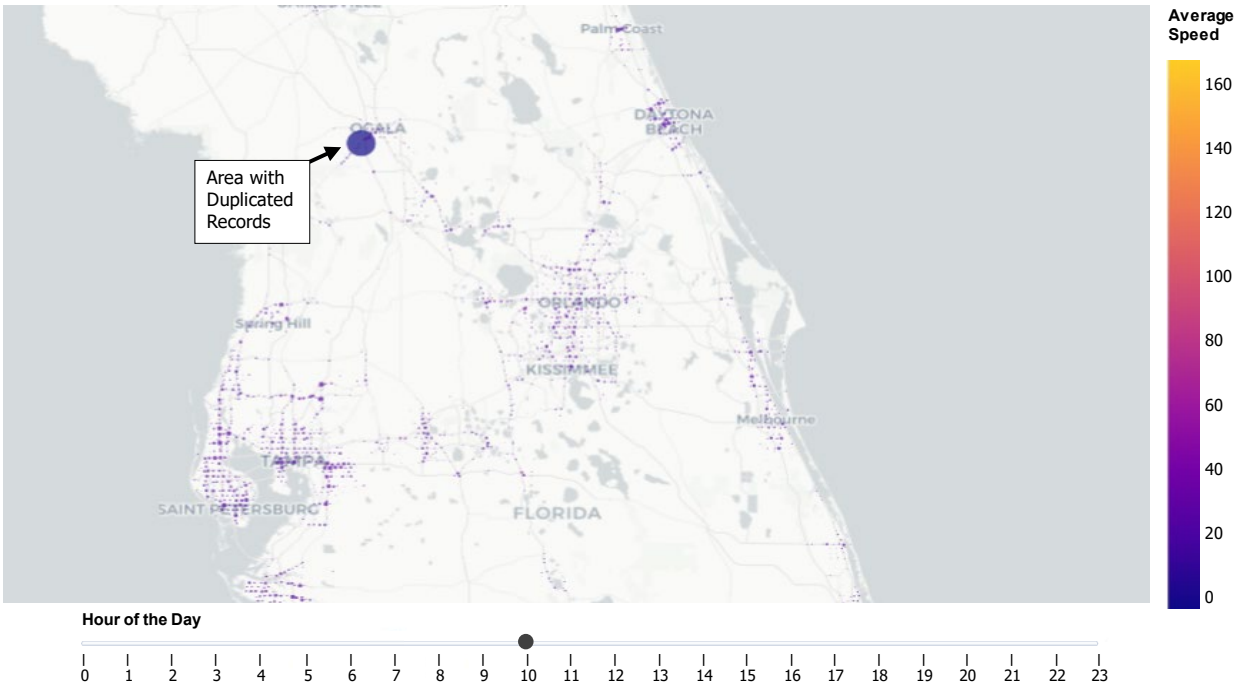
CV data are machine-generated by vehicles or roadway and roadside-based traffic control devices. Common assumption is that machine-generated data tend to have fewer data quality issues than manually manipulated data. The analysis of the post-CV data work indicated that CV data should not be assumed error free for post-CV usages. For example, as shown in Figure 18, almost all vehicle path history data elements in the JPO Pilot post-CV BSM file are not coded correctly. In addition, duplicative records for the same events exist. Figure 19 illustrates this type of geospatial coding error. Table 3 shows an example that locations in vehicle movement of the OEM CV data are not correct as two consecutive points are spatially separated too far away, resulting in an unrealistic calculated travel speed.

Figure 18: Wrong Path history of BSM data



Source: FHWA Office of Highway Policy Information.

Figure 19: OEM CV data records duplicated in one area.



Source: FHWA Office of Highway Policy Information.

Table 3: Wrong locations in vehicle movements

Journey_id	Start_Time	End_Time	Start_Lat	Start_Long	End_Lat	End_Long	Speed_Avg	Speed_Max	Distance	Dura-tion	Calculated Speed
45a69c	2021-07-01T11:13:31.000	2021-07-01T11:46:18.000	28.53476	−82.40806	26.53838	−80.07574	41.40	111.74	393258741	1967	719741
8cc24d	2021-07-01T10:51:50.000	2021-07-01T11:08:12.000	28.53249	−82.40768	28.53475	−82.40805	6.92	88.7	193574091	982	709640
4a9832	2021-07-01T10:34:06.000	2021-07-01T10:43:53.000	28.49272	−82.42498	26.74275	−80.06567	17.73	85.24	113989705	587	699085
8bb95c	2021-07-01T12:00:22.000	2021-07-01T12:00:52.000	55.03106	−162.50066	28.49251	−82.4249	6.678	16.12	5405219	30	648626

Source: FHWA Office of Highway Policy Information.

QA/QC process for post-CV data is not only important but also very challenging due to its sheer amount of data records and complicated data items.

7. Summary

As connected and autonomous vehicles are deployed, CV data availability and post-CV data analysis are becoming a reality. Post-CV data analysis offers information that transportation professionals never have had before. This new information includes specific roadway geospatial locations where different maneuvers such as acceleration and deceleration occurred and how such maneuvers are tied with roadway physical and operational conditions (e.g., pavement, congestion) and weather (e.g., wet pavement). It will offer decision-makers new tools and information to tackle safety and operational issues. Further information regarding driver seatbelt usage is unprecedented. The seatbelt information covering drivers and front passengers offers comprehensive information on when and under what conditions seatbelts are buckled or unbuckled.

To take advantage of post-CV data, it is imperative to have a suitable platform for data storage, data transmission, accessibility, and analytics. In the authors' experience, the election of a data platform should be based on the specific programming language the platform supports and the language expertise an organization's analysts possess. Given the size of the data and potential Personal Identifiable Information presence, the authors believe that ownership of the data is significantly less critical than the ability to access and utilize rights of such data. From a cost standpoint, accessing the data would also be significantly lower than owning the data.

Lastly, the authors would like to point out CV data quality issues. Even though CV data are mainly machine-generated, they are also prone to error. Analysts are cautioned that data quality checks should be performed before utilizing the data.

8. Acknowledgements

Wejo provided sample CV data and a Databricks platform for the Florida case evaluation.

The US DOT Intelligent Transportations Systems Joint Program Office (JPO) provided a portion of the CV data used in the analysis in this exploration.

FHWA Turner Fairbank Highway Research Center provided the Databricks platform known as the Path to Advancing Novel Data Analytics (PANDA) Laboratory for the JPO CV data analysis.

FHWA and all its offices and units do not endorse products or manufacturers. Trademarks or names appear in this report only because they are considered essential to the objective of the document.