Pre-Print Manuscript:
Zhou, Y., and Bridgelall, R. (2020). A Review of Usage of Large-Scale Connected Vehicle Data. *Transportation Research Record*, DOI: 10.1177/0361198120940996.

A Review of Usage of Large-Scale Connected Vehicle Data

Yun Zhou*
Doctoral Graduate Research Assistant
Department of Transportation, Logistics, and Finance, North Dakota State University
NDSU Department 2880
P.O. Box 6050, Fargo, ND 58108
Phone: (218) 289-7182; Email: yun.zhou@ndsu.edu
ORCID: 0000-0002-2782-0282

Raj Bridgelall, Ph.D.
Assistant Professor and Program Director
Department of Transportation, Logistics and Finance, North Dakota State University
NDSU Department 2880 P.O. Box 6050, Fargo, ND 58101-6050
Phone: (408) 607-3214; Email: raj@bridgelall.com
ORCID: 0000-0003-3743-6652

*Corresponding Author

Word count: 5,331 + 1,750 (7 Tables 250 each) = 7081 (7500 max)
Abstract: 208 words (250 max)

Funding Acknowledgements
The authors received no financial support for the research, authorship, and/or publication of this article.

Data Accessibility Statement
This paper (the *SPMD Dataset* subsection of the *Methodology* section) reviewed data that is available on the USDOT public data portal (https://data.transportation.gov/ or https://www.its.dot.gov/data/) and articles available as referenced.

ABSTRACT
GPS loggers and cameras aboard connected vehicles can produce vast amounts of data. Analysts can mine such data to decipher patterns in vehicle trajectories and driver-vehicle interactions. An ability to process such large-scale data in real time can inform strategies to reduce crashes, improve traffic flow, enhance system operational efficiencies, and reduce environmental impacts. However, connected vehicle technologies are in the very early phases of deployment. Hence, related datasets are extremely scarce, and the utility of such emerging datasets is largely unknown. Subsequently, this paper provides a comprehensive review of studies that used large-scale connected vehicle data from the United States Department of Transportation Connected Vehicle Safety Pilot Model Deployment program. It is the first and only dataset available to the public. The data contains real-world information about the operation of connected vehicles that organizations are testing. The authors provide a summary of the available datasets, their organization, the overall structure, and other characteristics of the data captured during pilot deployments. Subsequently, the authors classify the data usage into three categories: driving pattern identification, development of surrogate safety measures, and improvements in the operation of signalized intersections. Finally, the authors identify some limitations experienced with the existing dataset.

Keywords: Connected vehicle, Intelligent Transportation Systems, Smart cities, Signal phase and timing, Surrogate safety measures, Risky driving pattern, Intersection safety.

INTRODUCTION

Connected vehicle (CV) technology enables real-time communications among users, vehicles, and the multimodal infrastructure. Producers and transportation agencies assert that CV technology will dramatically reduce the number of fatalities and serious injuries caused by accidents on our roadways. The technology will achieve this by notifying and alerting drivers about potentially dangerous driving situations. Examples include pedestrians or bicyclists approaching an intersection, vehicles in blind areas beyond a curve, and oncoming cars swerving into a lane to avoid an object or pothole on the road (*1*). CV technology can also smooth out traffic flows, diminish congestion, and reduce travel time. Agencies can analyze CV data to inform eco-friendly transportation planning (*2*).

GPS loggers and cameras aboard CVs can produce abundant travel data. CV data exchanges include vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I). CVs communicate basic safety messages (BSM) among vehicles and with roadside equipment (RSE) (*3*-*4*). BSMs contain information such as vehicle location, speed, acceleration, and time. In addition to BSMs, CVs also produce trajectory data, various driver-vehicle interaction data, and contextual data. Analysts can mine such data to inform strategies that can reduce crashes, improve traffic flow, enhance system operational efficiencies, and reduce environmental impacts.

CV development is still in its infancy. Hence, there is very little information about practical challenges and quantifiable benefits of real-world deployments. Therefore, the United States Department of Transportation (USDOT) launched a one-year Connected Vehicle Safety Pilot Model Deployment (SPMD) in August 2012 in Ann Arbor, Michigan to advance knowledge about practical deployments. The deployment included 30 roadside equipment (RSE) installations along approximately 75 lane-miles of roadway and approximately 3,000 equipped vehicles (*4*). The data collected is now available to the public via the USDOT's public data portal (https://data.transportation.gov/ or https://www.its.dot.gov/data/) (*4*). As part of the USDOT CV pilot programs, the agency awarded in 2015 a three-phase pilot to two cities and one state—New York City, Tampa, and Wyoming (*5*). All three sites finished a 12-month period of concept development (phase I) and a 24-month period of deployment design, build, and test (Phase II). Each site is now entering the final deployment phase to operate and test deployed CV systems for a minimum of 18 months. The sites will monitor key performance measures of the CV pilot. Some datasets from this program are available at the USDOT's public data portal.

Based on the current literature, analysts can use the CV data by itself or combine it with other data sources to identify driving patterns, develop surrogate safety measures, evaluate location-based intersection safety, and improve operations at signalized intersections. The **contributions** of this paper are a comprehensive snapshot-in-time review of previous studies that used connected vehicle data, and potential areas that researchers can advance using the newly released large-scale CV data from the USDOT Pilot Deployment Program. This paper contains all the information in one place to facilitate ongoing research about the potential value and utility of emerging CV data.

The remainder of this paper includes: a methodology section which describes the approach to the literature review and the strategy to classify the usage of the CV data; a results section which summarizes the data usage and applications of the CV dataset, a conclusion section which provides overall remarks about the usability and limitations of the CV dataset and briefly discusses future work.

METHODOLOGY

The authors first researched the newly released large-scale CV datasets from the United States Department of Transportation (USDOT) Pilot Deployment Program to understand its organization, structure, and content. After conducting extensive literature searches using all the traditional scientific databases of research output, the authors organized the data usage into three categories of application development. The publication sources included the Transportation Research Information Services (TRIS) Database and the International Transport Research Documentation (ITRD) Databases.

SPMD Dataset

The SPMD data comprises of the following datasets: two driving datasets that include the Data Acquisition System 1 (DAS1) and Data Acquisition System 2 (DAS2), a BSM dataset, an RSE dataset, and three contextual dataset that include weather, network, and schedule. Each table in a database contains comma-separated value (CSV) formatted information collected during a 24-hour period (*6*).

DAS1 and DAS2 contain data from the DAS that the University of Michigan Transportation Research Institute (UMTRI) and the Virginia Tech Transportation Institute (VTTI) developed, respectively. The BSM dataset includes messages that a participating vehicle transmitted and/or received, irrespective of the DAS installed. The RSE dataset contains data that roadside units transmitted and/or received. The contextual datasets contain information about conditions at the time of data collection. Contextual data include information about network configuration and performance, weather, schedules (transit and special events), roadwork activity, and traffic incidents.

Several studies indicated that the CV data was available at the Research data Exchange (RDE, https:www.its-rde.net/home), but this link is no longer active. The CV data is now only available at the USDOT's public data portal (https://data.transportation.gov/ or https://www.its.dot.gov/data/). The *DataWsu* and *DataFrontTargets* are considered as part of the DAS1 dataset but they available as separate data files at the USDOT's public data portal. The SPMD Sample Data Handbook provides a detailed introduction for each dataset and all the data files under that dataset (*6*). Table 1 summarizes the datasets and a list of their accompanying files (6). The DAS1 dataset was mentioned most as a data source in the reviewed studies. Table 2 further describes each file in the DAS1 dataset. The *DataWsu* and *DataFrontTargets* files are most commonly used to capture the position and motion information of host vehicles. The *DataWsu* file contains 27 fields, which is the most of all the DAS1 data files. Most of the data logged in the *DataWsu* file comes from the onboard Wireless Safety Unit (WSU) that produces GPS and inertial sensor data, and the Controller Area Network (CAN) that communicates vehicle performance and status information. Table 3 describes the data elements in the *DataWsu* file.

The Mobileye system the Intel Corporation is a vision-based system that enables various Advanced Driver Assistance System (ADAS) capabilities. The *DataFrontTargets* file contains information from the installed Mobileye system that collects information from the scene ahead of the vehicle. The system uses communicates measures and warnings based on a serious of proprietary algorithms. Table 4 briefly describes the data elements of *DataFrontTargets* file.

CV Pilot Deployment Program Dataset

The program scale of the three-phase pilot sites is much larger than that of the SPMD program. Furthermore, the estimated program duration of 54 months is more than four times that of the SPMD program. At the time of this writing, there were only six sample RSE data files available from the USDOT's public data portal. They are:

1. Wyoming
    a. BSM one-day sample file
    b. Two traveler information message (TIM) sample files
2. Tampa
    a. One signal phasing and timing (SPaT) sample file
    b. One BSM sample file
    c. One TIM sample file
3. New York
    a. None.

Because the data from the three CV Pilot Deployment sites are limited, this paper focuses on studies that use the SPMD data or other CV data sources or probe data to determine the research areas that can be advanced using emerging CV datasets from the three CV Pilot Deployment sites.

Table 1 Files Associated with the SPMD Dataset. Source: SPMD Sample Data Handbook (*6*).

Driving Data		Message	Infrastructure	Contextual		
DAS1	**DAS2**	**BSM**	**RSE**	**Weather**	**Network**	**Schedule**
AudioTimes*	HV_Radar	BrakeByte1Events	BSM	Weather/ climatic data	Pointer to Resources	Pointer to* Resources
DataFrontTargets	HV_Primary	BrakeByte2Events	Geometry			
DataLane	DAS2_Trip_Summary*	BsmP1	Lane			
DataWsu		ExteriorLightsEvents	LaneNode			
DAS1_Trip_Summary*		PosAccurByte1Events	MAP			
		PosAccurByte2Events	Packet			
		PosAccurByte3Events	PCAPFile			
		PosAccurByte4Events	SPAT			
		SteerAngleEvents	SPATMovement			
		ThrottlePositionEvents	TIM			
		TransStateEvents	TIMRegion			
		WiperStatusFrontEvents	TIMRegionNode			
		BSM_Trip_Summary*				

*Not available at USDOT's public data portal (https://data.transportation.gov/).

Table 2 Description of the Files in the DAS1 Dataset. Source: SPMD Sample Data Handbook (*6*).

File Number	File	Description	Update Frequency
1	DataFrontTargets	Data collected by the Mobileye sensor. It captures information about the (vehicle) or object that is in front of the host vehicle.	10Hz
2	DataLane	Stores lane marking quality adjacent to the host vehicle and the distances between each side of the vehicle and each lane line.	10Hz
3	DataWsu	Logs of the GPS and CAN Bus data produced by an onboard device.	10Hz
4	DAS1_Trip_Summary*	A list of summary measures for each vehicle trip.	1 per trip

*Not available at USDOT's public data portal (https://data.transportation.gov/).

Table 3 Data Elements of the DataWsu File. Source: SPMD Sample Data Handbook (*6*).

Field Name	Type	Units	EnumId	Description
Device	Integer	none	-	A unique numeric ID assigned to each DAS. This ID also doubles as a vehicle's ID.
Trip	Integer	none	-	Count of ignition cycles—each cycle starts and ends when the ignition is in the on and off positions, respectively.
Time	Integer	centiseconds	-	Time (centiseconds) since the DAS started, which (generally) starts when the ignition is in the on position.
GpsValidWsu	Integer	none	1	Indicates whether or not a GPS data point is valid.
GpsTimeWsu	Integer	millisecond		Epoch GPS time received from the remote vehicle that has been targeted by the host vehicle's WSU.
LatitudeWsu	Float	deg	-	Latitude from WSU receiver .
LongitudeWsu	Float	deg	-	Longitude from WSU receiver.
AltitudeWsu	Real	m	-	Altitude from WSU receiver.
GpsHeadingWsu	Real	deg	-	Heading from WSU GPS receiver .
GpsSpeedWsu	Real	m/sec	-	Speed from WSU GPS receiver.
HdopWsu	Real	none	-	Horizontal dilution of precision.
PdopWsu	Real	none	-	Position dilution of precision.
FixQualityWsu	Integer	none	-	GPS Fix Quality.
GpsCoastingWsu	Integer	none	-	GPS Coasted.
ValidCanWsu	Integer	none	1	Valid Vehicle CAN Bus message to WSU.
YawRateWsu	Real	deg/sec	-	Yaw rate from vehicle CAN Bus via WSU.
SpeedWsu	Real	kph	-	Speed from vehicle CAN Bus via WSU.
TurnSngRWsu	Integer	none	11	Right turn signal from vehicle CAN Bus via WSU.
TurnSngLWsu	Integer	none	11	Left turn signal from vehicle CAN Bus via WSU.
BrakeAbsTcsWsu	Integer	none	-	Brake, ABS, and traction control from vehicle CAN Bus via WSU.
AxWsu	Real	m/sec^2	-	Longitudinal acceleration from vehicle CAN Bus via WSU.
PrndlWsu	Integer	none	403	Current transmission state (Park, Reverse, Neutral, Drive, Low) from vehicle CAN Bus via WSU.
VsaActiveWsu	Integer	none	-	Stability control active from vehicle CAN Bus via WSU.
HeadlampWsu	Integer	none	-	Headlamp state from vehicle CAN Bus via WSU.
WiperWsu	Integer	none	-	Wiper state from vehicle CAN Bus via WSU.
ThrottleWsu	Real	none	-	Throttle position from vehicle CAN Bus via WSU.
SteerWsu	Real	deg	-	Steering angle/position from vehicle CAN Bus via WSU.

Table 4 Data Elements of the DataFrontTargets File. Source: SPMD Sample Data Handbook (*6*).

Field Name	Type	Units	EnumId	Description
Device	Integer	none	-	A unique numeric ID for each DAS, which also doubles as a vehicle's ID.
Trip	Integer	none	-	Count of ignition cycles that begins and ends when the ignition is in the on and off position, respectively.
Time	Integer	centiseconds	-	Time (centiseconds) since DAS started, which (generally) starts when the ignition is in the on position.
TargetId	Integer	none	-	Numeric ID that the Mobileye sensor assigns to different objects being tracked, with a value of 1 assigned to the closest.
ObstacleId	Integer	none	-	ID of new obstacle that the Mobileye sensor assigns—the value is the last used free ID.
Range	Integer	m	-	Longitudinal position of an object (typically the closest) relative to a Mobileye defined reference point on the host vehicle.
RangeRate	Real	m/sec	-	Rate of change of the Range variable.
Transversal	Real	m	-	Mobileye assigned lateral position of the obstacle.
TargetType	Integer	none	409	Object classification (car, truck, pedestrian, etc.)
Status	Integer	none	410	Motion classification (stopped, moving, etc) of an identified obstacle/target.
CIPV	Integer	none	1	Indicates whether an obstacle in the vehicle's path is the closest.

RESULTS

Several studies reported on the use of CV data from SPMD. This section classifies them into three categories: driving pattern identification, development of surrogate safety measures, and improvement of signalized intersection operation.

Driving Pattern Identification

Driving pattern is one of the key factors that affect traffic safety. Vehicle speed, acceleration, and deceleration are primary factors in the classification of driving patterns. Agencies consider that driving above the speed limit is hazardous and risky. Speed restrictions can be a dynamic function of road conditions and traffic situations. Researchers proposed several acceleration thresholds to classify driving behavior as calm, normal, and aggressive (*7-11*). Risky driving happens when longitudinal or lateral accelerations exceed certain thresholds. Researchers found that risky driving patterns are highly correlated with the likelihood of crashes or near-crash events (*8-9*) (*11-13*).

Three recent studies used the CV data from the SPMD program to propose methodologies that use critical information from the instantaneous BSM exchanges between CVs and roadside equipment to determine repeatable driving behaviors (*14-16*). Study (*14*) investigated the longitudinal and lateral motion of the driving decision from BSMs and established reasonable thresholds to identify potentially dangerous events such as hard accelerations or braking, and quick lane changes. They used the DAS datasets and the 10-Hertz motion data that contain speed along with longitudinal and lateral acceleration to visualize the driving behavior. They investigated the relationships between speed and acceleration by visualizing the distributions of acceleration in longitudinal and lateral directions. The results validated that instantaneous driving decisions could provide valuable information to identify extreme driving events such as sudden lane changes. The distributions of directional variations in acceleration informed the thresholds of extreme acceleration in different directions. The study presented valuable information on establishing context-relevant alerts, warnings, and control assistance to nearby vehicles.

Study (*15*) conducted a time-series analysis to categorize driving patterns into different regimes based on their volatility and average duration. The study explored correlations to examine dwell times and switching times between regimes. The study used Google Earth to visualize vehicle trajectories from the DAS dataset and to classify trips based on roadway type. Aggregating the DAS datasets into one-second groups enabled a detailed econometric analysis of instantaneous driving decisions. They analyzed the data using an expectation-maximization and Dynamic Markov switching models of two-regimes and three-regimes. The results revealed that acceleration and deceleration are two distinct regimes. The rate of deceleration was higher than the rate of acceleration, and braking was more volatile during deceleration than during acceleration.

Machine learning methods can identify the importance of motion-related variables in classifying driving data into aggressive and normal driving patterns (*17-18*). The authors of (*16*) applied machine learning to the CV data from the SPMD program to identify aggressive driving patterns on horizontal curves. They used the random forest method of machine learning to develop an aggressive driving detection model based on a time-to-lane crossing (TLC) metric, under three scenarios. This detection model provided high classification accuracy in all three scenarios, and it ranked the importance of the variable candidates in identifying aggressive driving behaviors.

A common limitation for these three studies that used the SPMD dataset was the limited number of observations. Study (*14-15*) used the one-day sample datasets but deleted some observations that contained errors. Study (*16*) used a one-month dataset. Larger datasets require more processing capacity to produce results in a reasonable amount of time, but they can facilitate the removal of outlier situations, such as extremely bad weather conditions that could bias the results.

Surrogate Safety Measures Development

Historical crash data can identify high-risk locations. However, because historical crash data usually take a significant amount of time to collect, researchers developed surrogate safety measures (SSMs). They are proactive solutions to assess safety risks by capturing near-crash events when crash data is absent or limited. SSMs can quantify safety-related performance at a road segment or evaluate the effectiveness of a safety treatment more efficiently (*19*). Data collected from sensors can inform the development of SSMs to identify high-risk locations accurately. The newly available CV data collected from CV devices and RSEs give researchers a new opportunity to conduct SSM research. Some researchers attempted to develop SSMs from the vehicle trajectory data of the SPMD program (*20-21*).

Study (*20*) proposed a framework to process CV data, calculate SSMs and their safety indices, and analyze the correlation between crash records and the calculated safety indices. They calculated three SSMs, which were time-to-collision (TTC), modified time-to-collision (MTTC), and deceleration rate to avoid collision (DRAC), at the vehicle-level and their safety indices at both the trip-level and the link-level. They used a negative binomial model to analyze the relationship between crash records and the safety indices of three SSMs. They found that the MTTC model provided the best overall performance. The study concluded that using SSMs with motion-related CV data could improve overall safety evaluation.

Study (*21*) developed a new SSM called time-to-collision with disturbance (TTCD) which can capture rear-end conflict risks in car-following scenarios. They used the CV data to access the risk identified with the TTCD model by comparing that risk with the risk identified by historical crash data. The result was that, among all accessed SSMs, TTCD can capture risk data at the highest level of correlation with historical rear-end crash data. The associated high-risk locations identified by TTCD were very similar to those identified by historical crash data. This study suggested that researchers could use real-world CV data to identify high-risk locations as crash predictors. This result validated the results from study (*22*), which presented a framework and used simulated CV data to examine surrogate measures for evaluating the risk of secondary crashes on highways.

Both studies (*20-21*) presented detailed procedures for cleaning and processing the CV data from the SPMD program. They determined that CV data could help predict high-risk locations in a very short period (one month and two months) before a substantial number of crashes occur. Thus, CV data can help to develop proactive safety measures to improve road safety management (*21*).

Intersections are one of the most dangerous locations on roadways based on their annual crash history (*23-24*). Traditionally, the safety evaluation of an intersection depends on historical crash frequency data and survey feedback from active users. Analysis of CV data can detect high-risk intersections where historical crash data is limited. Three studies (*25-27*) combined historical crash data and real-world CV data from the SPMD program to evaluate location-based intersection safety. The purpose of this type of study is to conduct proactive safety measures on specific intersections before the occurrence of crashes, and to seek solutions for improving their safety. Studies (*20-21*) attempt to identify high-risk locations which are not limited to intersections, whereas studies (*25-27*) focused on evaluating location-based intersection safety due to the high crash frequency detected at intersections.

One study (*25*) provided an example of using CV data to assess location-based risk by detecting extreme driving decisions. Researchers used the results of previous work that identified extreme driving events to estimate crash risk as a function of instantaneous driving decisions at specific locations. They introduced the concept of location-based volatility (LBV) to calculate the coefficient of variation as a standardized measure of dispersion. The coefficient of variation indicates the fluctuation of the longitudinal acceleration and deceleration at a specific location. The study used rigorous fixed- and random-parameter Poisson regression models to investigate the relationship between LBV and crash frequency. Results suggested that there is a statistically significant relationship between LBV and crash frequencies for signalized intersections. One limitation of the study was the limited number of variables available after the deletion of inaccurate observations.

Another study (*26*) combined the CV BSM data with the crash and inventory data at several intersections to investigate the relationship of different measures of volatility with crash frequency. The study evaluated thirty-seven different measures of volatility. The researchers used fixed- and random-parameter Poisson regression models in two levels of bulk passing and individual passing to investigate the relationship between each measure of volatility and crash frequency. Several methods investigated showed positive and statistically significant association with crash frequency. They methods were the three measures at bulk passing level, the time-varying stochastic volatility of speed, the percent laying beyond threshold-bonds of speed created using mean plus two standard deviation at intersections, and the percent laying beyond threshold-bonds of acceleration created using mean plus two standard deviation at intersections.

One research group (*27*) proposed a methodology to quantify driving volatility at each intersection to assess intersection-based crash risk based on CV BSM data, crash data, and traffic and intersection inventory data. They proposed to quantify driving volatility based on speed, acceleration/deceleration, vehicular jerk, eight different volatility measures, coefficient of variation, mean absolute deviation around mean, percentage of outliers, and time-dependent dynamic volatility, at both aggregate intersection level and trip level. Subsequently, they used Poisson and Poisson-lognormal regressions models to test the correlations between crash frequency and intersection-based volatility with consideration of unobserved heterogeneity. They used Full Bayesian estimation method and Markov Chain Monte-Carlo Gibbs Sampler techniques to estimate the parameters. They calculated Moran's I statistics to investigate the correlation between crash frequency and spatial factors. The results suggested that the crash frequency significantly correlated with three measures: the two standard deviations threshold at the intersection level, the coefficient of variation of speed at the pass level, and the mean absolute deviance of vehicular jerk at the passing level.

The common limitation of the three studies was the limited CV data available during the research period. The one-month or two-month CV data were relatively small sample sizes to explain 5-year average crash rates. In addition, all studies only considered crash frequency but not crash severity as a risk factor.

Signalized Intersection Operation Improvement

Signalized intersections are usually hot spots for traffic congestion, especially during rush hour. They cause significant hours of delay and crash volume every year. Agencies often consider adaptive signal control to accommodate varying demands. However, their significant cost to install and maintain is a deterrent to deployment. A less-expensive alternative is to re-time the signal by analyzing CV data from RSEs to estimated traffic volume (*28*-*30*). A downside of this approach is the currently low penetration rate of CVs. Study *(31)* conducted a proof-of-concept by using CV data in a low penetration rate environment to optimize signal coordination. They could not use vehicle trajectories because the data was from a fixed location. Estimating traffic volume from vehicle trajectories is an essential input for signal operation and algorithm design optimization.

Study (*32*) proposed a method to estimate traffic volumes by using trajectory data from CVs or trajectory data from navigation devices at locations with low CV penetration rates. The study used the BSM data from the RSE in the SPMD program to capture trajectory variables such as motion and position, and data from the signal phase and timing (SPaT) data to capture the timing periods and signal status. They combined the data to produce a space-time trajectory dataset. They modeled traffic arrivals as a time-dependent Poisson process. An expectation maximization (EM) procedure provided an estimate of the arrival rate. They tested the estimation procedure with CV trajectory data and vehicle trajectory data from navigation service users. The results suggested that their proposed approach was effective and that agencies could use it to improve signal control operation in environments with low CV penetration rates.

Monitoring queue status at signalized intersections could help optimize the available capacity (*33*-*35*). For example, CV data can enable improved route selection through real-time notifications of traffic status. Study (*36*) proposed an integrated macroscopic and microscopic traffic flow model to estimate time-space queueing dynamics at signalized intersections using RSE and SPaT data from the CV BSM repository. They identified the three regions of queue formation region, queue region, and queue dissipation under the normal scenario and oversaturated scenario based on vehicle deceleration, stop, and acceleration behaviors. This integrated method estimated queue process in both queue length and queue time at signalized intersections. Thus, this method could help to improve real-time traffic status estimation at the signalized intersections equipped with connected vehicle technologies.

Data Manipulation

Error Checking

Best practices in the use of large-scale data begin with quality evaluation and cleaning before conducting any analysis or data mining tasks. However, relatively few studies reported experience with error-checking and deep-cleaning SPMD datasets. The U.S. National Highway Traffic Safety Administration published an independent evaluation of safety applications for passenger vehicles in the SPMD program (*37*). They found several errors in the programing of Volkswagen-Audi’s forward-collision warning (FCW), intersection movement assist (IMA) applications, and issues with a GPS on one vehicle that led to inaccuracies in some data records in the SPMD datasets. Other researchers found errors in the DAS dataset such as speeds faster than 200 mph and altitudes greater than 30,000 ft (*14*) (*21*). Another research found that a significant portion (42%) of the “lateral acceleration” observations exceeded the maximum value that the wireless communications device could record (*25*). Duplicated records and invalid messages were found in DAS dataset (*21*).

Studies (*15*, *26*) checked dataset and claimed that no error has occurred. Study (*15*) indicated that they conducted error-checking process by linking microscopic trip data with a trip-summary file to check the information consistency at trip-level and didn't state error-checking for data values. Study (*26*) performed an error-checking process for its spatial data by mapping data and resulted with a good match with the real map.

Among all the studies, study (*21*) presented a detailed data process procedure by a detailed data preparation description and a data flow chart, which provides great guidelines for both data cleaning and data process for later researches. Study (*20*) presented data process and indicated the software applied for data process. Study (*26*) presented data integration and process steps in a data flow chart. The authors of (*38*) demonstrated an automated method of cleaning the data of a taxi probe dataset that utilizes known distributions of vehicle operations to detect possible outliers for removal. Table 5 summarizes data shortcomings and provides recommendations for data cleaning based on knowledge synthesized from the literature search. Table 6 summarizes the data cleaning and processes, and the software tools used.

Data Mining Approaches

Data mining techniques extract patterns from large-scale data that are interesting (*39*). Common data mining approaches include statistical regression models and machine learning methods (*39*). Statistical regression models estimate the numerical relationships between variables and can predict new values, whereas machine learning methods recognize complex patterns and facilitate decision-making based on data (*39*).

Statistical regression models are the most commonly used methods among all the studies reviewed. As shown in Table 7, seven studies (*14-15*, *20*, *25-27*, *32*) applied different regression models to investigate the numerical relationships between factors. Study (*21*) calculated Pearson's correlation coefficients as test statistics to quantify correlations between developed measures and crash frequency. One study used the machine learning method of Random Forest to identify aggressive/risky driving (*16*). Additional studies suggested potential application of machine learning methods in CV data. Random forest and Support Vector Machines have been applied to the driving style classification and transportation mode recognition problem (*39-40*). Study (*33*) used CV data, without machine learning, to demonstrate the developed model.

Data Mining Challenges

As summarized in Table 5, a common data cleaning recommendation is to detect and remove erroneous records from the dataset, and to abandon data fields where there is a considerable number of data records with errors. However, data elimination may reduce the data size, which could result in lower model accuracy (*15*). Computation capacity may become another issue when the dataset is very large. Studies (*27*) indicated that the workstation level computer (Dell Precision T7600, 3.1 GHZ (32 CPUs) took a considerable amount of time to compute data models for a data size of 230 million observations.

Table 5 Data Shortcomings and Recommendation for Data Cleaning

Type	Description	Examples	Recommendations for Data Cleaning
Outliers	Data values exceed maximum allowable value that can be recorded, or does not represent fact.	Value exceed maximum allowable value: Speeds > 200 mph [14]. Speeds > 415 mph and acceleration rate > 10 m/s/s [21]. Do not represent fact: Altitude > 30,000 ft [14].	Removed outliers [14][21]. Avoid to use all the values in the variable if outliers take a large portion of the data values (42% of data) [25].
Duplicated records	Duplicated records	Duplicated records in *DataWsu* data [21].	Check for duplicated data records and removed such records if they exist [21].
Invalid message	Invalid message	Invalid WSU or CAN Bus Message in *DataWsu* [21].	Filter records to remove invalid messages, e.g. filter with criterion "*GpsVaildWsu* = 1 and *VaildCanWsu*=1" [21].
Improperly recorded message	Activity recorded out of the scale that a sensor is designed for, thus recorded data values didn't fall into a normal data range.	Mobileye sensors may record the speed of vehicles in opposite directions if the road is narrow and doesn't have a median. Thus, the *Rangerate* values (speed of leading vehicle - speed of the following vehicle) in *DataFrontTargets* data may be negative and its absolute value is greater than the host vehicle's speed, indicating that the leading vehicle is backing up at a speed of *Rangerate-GpsSpeedWsu* [21].	Filter records to improperly recorded message, e.g. filter with criterion "*Rangerate-GpsSpeedWsu*>1" [21].

1 **Table 6 Summary of Data Processing**

Topics	Studies	Dataset	Data Sample Information	Data Cleaning and Processing	Software
Driving Pattern Identification	[14]	DAS	968,522 records of basic safety messages, from 155 trips made by 49 vehicles	Removed observations with errors. (Speed >200 mph, and altitude > 30,000 ft). Data visualization to show the extent of instantaneous driving volatility.	R, MATLAB, and Google Earth for data processing and visualization. Stata for modeling
	[15]	DAS	1,399,084 records of basic safety messages, from 184 trips made by 71 vehicles	Error-checked by linking microscopic trip data with a trip-summary file. Two datasets matched in terms of trip-level information. Data aggregation from 10 BSM per second to 1 BSM per second.	R, MATLAB, and Google Earth for data processing and visualization. Stata14.1 for modeling
	[16]	BsmP1 (BSM)	1.5 billion rows of data, data size 204 GB	Data records on the eastbound of a horizontal curve were selected. East of (42.299469, -83.724666) were eliminated for study design.	R for data processing and extract information, Google Earth for extracting GPS coordinates.
Surrogate Safety Measures	[20]	*DataWsu* and *DataFront Targets* (DAS1)	*DataWsu* file of 12 GB and *DataFrontTargets* file of 4.34 GB from nearly 100 vehicles.	Two datasets were read by Python to check data type and data organization. Then import to Hadoop for query using Apache Hive. Next, datasets were exported into small files and joined in PostgreSQL database. Fourthly, ArcGIS were used to ingrate link and intersection information. Fifthly, the data points around the intersections were removed by a 75-ft buffer zones created in PostgreSQL. Data processing framework figure in page 8	Python, Hadoop, Apache Hive, PostgreSQL, and ArcGIS,
	[21]	*DataWsu* and *DataFront Targets* (DAS1)	62,589,725 BSMs from 90 vehicles.	*DataFrontTargets* file were filed for observations with vehicles in front of the host vehicle. Duplicates were removed from *DataWsu* file. Then *DataWsu* file were filtered remove invalid bus messages. Next, two cleaned datasets were merged, and cleaned to remove outliers. Finally, the datasets filtered out the vehicle movement in the opposite direction. Data process procedure figure in page 314	R for data manipulation and ArcGIS for spatial processing. R package ggmap for data visualization.
	[25]	Not stated	Not stated	Checked data values, 42% of data, 27,240,788 data point, had the lateral acceleration values exceeded the maximum value that the wireless communications device could record. Lateral acceleration variable was not used in this study.	Not stated

	[26]	BSM (RSE)	215,000,000 BSMs at selected intersections.	Data examination and error-checking process before data integration. Extra intersection data using geocodes to map the intersection location data from BSM, well matched with the real map. Appropriate geocoded polygons are used to filter BSM data for each selected intersection. A data integration and processing steps showed as a figure in page 295	Not stated
	[27]	Not stated	230 million BSMs	Appropriate geocodes are used to filter BSM data for each selected intersection. Zero speeds are removed from BSM data.	Stata's MATA language for modeling, and WinBUGS software for MCMC Gibbs sampling. noted that computations took long time at workstation level computer (Dell Precision T7600, 3.1 GHZ (32CPUs)).
Signalized Intersection Operation Improvement	[32]	BSM data and SPAT (RSE)	Not stated	First select an interested movement and select GPSdata associated with the movement and time period based on direction of CV trajectories and prepare corresponding signal status data. Then, based on road geometry, calculate CVs' longitudinal position along the road from GPS positions, and generate time-space trajectories. Map GPS time into signal clock time and then aggregated trajectories to calculate the time dependent factor.	Not stated
	[36]	SPAT and V2I driving records data (inferred to be BSM in RSE database	2150 vehicle's daily trajectories	Not stated	Not stated

Table 7 Data-Driven Methods Applied to CV Data

Topics	Studies	Technique and Purpose
Driving Pattern Identification	[14]	Negative binomial regression model: Correlation of extreme event frequency.
	[15]	Markov-switching dynamic regression model (time series analysis): Quantification and prediction of driving patterns.
	[16]	Random forest classification of risky driving behaviors.
Surrogate Safety Measures	[20]	Negative binomial regression model: Statistical relationship between the link developed safety surrogate measures and crash frequency.
	[21]	Pearson's correlation coefficients: correlation between developed safety surrogate measures and rear-end crashes.
	[25]	Fixed-and random-Poisson regression models: Quantification of the relationship between intersection-specific violations and crash frequency.
	[26]	Fixed-and random-Poisson regression models: Quantification of the relationship between intersection-specific violations and crash frequency.
	[27]	Hierarchical fixed- and random-parameter Poisson and Poisson log-normal models: Model crash function of intersection-specific volatilities and other factors. Full Bayesian estimation method and Markov Chain Monte-Carlo Gibbs sampler techniques: estimate parameter in Poisson models. Moran's I statistics: investigate correlation between crash frequency and spatial factors.
Signalized Intersection Operation Improvement	[32]	Time-dependent Poisson process: Model of traffic arrivals. Expectation Maximization: estimate parameter.
	[33]	Developed a mathematical model (without data mining) and used CV data to demonstrate the model.

CONCLUSION

This paper surveyed the literature and identified ten studies that used the only real-world connected vehicle dataset currently available to the public—the large-scale CV datasets recently released from the United States Department of Transportation (USDOT) Pilot Deployment Program. This paper first provides a summary of the available datasets and describes their organization, overall structure, and characteristics of the data captured during pilot deployments. Secondly, the authors presented a summary of studies that used the data and classified the usage into three categories involving application development, identifying driving patterns, developing surrogate safety measures, and improving the operation of signalized intersections.

One common limitation indicated in some of studies is that only part of the dataset is useful for analyses because of errors in the data collection processes and the large percentage of erroneous attributes. All studies used one-day, one-month, or two-month CV data from the USDOT program. A primary contribution of this review is a summary of current usage and applications of the first and only dataset available to the public that contains real-world CV data. This summary is in one place, thus providing a convenient reference to the research community. Future work will extend the survey as more data becomes available.

AUTHOR CONTRIBUTION STATEMENT:

The authors confirm contribution to the paper as follows: study conception and design: Yun Zhou and Raj Bridgelall; data collection: Yun Zhou and Raj Bridgelall; analysis and interpretation of results: Yun Zhou and Raj Bridgelall; draft manuscript preparation: Yun Zhou and Raj Bridgelall. All authors reviewed the results and approved the final version of the manuscript. The authors do not have any conflicts of interest to declare.

REFERENCE

1. JPO, ITS. What Are Connected Vehicles and Why Do We Need Them? https://www.its.dot.gov/cv_basics/cv_basics_what.htm. Accessed on March 22, 2019
2. JPO, ITS. How Will Connected Vehicles Be Used? https://www.its.dot.gov/cv_basics/cv_basics_how_used.htm. Accessed on March 22, 2019
3. Greer, L., J., Fraser, D. Hicks, M. Mercer, and K. Thompson. Intelligent Transportation Systems Benefits, Costs, and Lessons Learned: 2018 Update Report. No. FHWA-JPO-18-641. United States Department of Transportation. Intelligent Transportation Systems Joint Program Office, 2018.
4. JPO, ITS. Safety Pilot Model Deployment Data. https://datahub.transportation.gov/Automobiles/Safety-Pilot-Model-Deployment-Data/a7qq-9vfe. Accessed on April 25, 2020
5. JPO, ITS. Connected Vehicle Pilot Deployment Program – Program Overall. https://www.its.dot.gov/pilots/pilots_overview.htm. Accessed on March 22, 2019
6. JPO, ITS. Safety Pilot Model Deployment Data, Safety Pilot Model Deployment Sample Data Handbook, Safety_Pilot_Model_Deployment_Sample_Data_Handbook.docx. Accessed on March 22, 2019.
7. De Vlieger, I., D. De Keukeleere, and J.G. Kretzschmar. Environmental Effects of Driving Behaviour and Congestion Related to Passenger Cars. *Atmospheric Environment.* 2000. 34(27): 4649-55.
8. Simons-Morton, B.G., M.C. Ouimet, Z. Zhang, S.E. Klauer, S.E. Lee, J. Wang, P.S. Albert, and T.A. Dingus. Crash and Risky Driving Involvement Among Novice Adolescent Drivers and Their Parents. *American Journal of Public Health*. 2011.101(12) :2362-7.
9. Simons-Morton, B.G., Z. Zhang, J.C. Jackson, and P.S. Albert. Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes? *American journal of epidemiology*. 2012. 175(10):1075-9.
10. Simons-Morton, B.G., K. Cheon, F. Guo, and P. Albert. Trajectories of Kinematic Risky Driving Among Novice Teenagers. *Accident Analysis & Prevention*. 2013. 51: 27-32.
11. Kim, E., and E. Choi. Estimates of Critical Values of Aggressive Acceleration from a Viewpoint of Fuel Consumption and Emissions. Transportation Research Board Annual Meeting, Washington, DC. 2013.
12. Paleti, R., N. Eluru amd C.R. Bhat. Examining the Influence of Aggressive Driving Behavior on Driver Injury Severity in Traffic Crashes. *Accident Analysis & Prevention.* 2010. 42(6): 1839-1854.
13. Guo, F., and Y. Fang. Individual Driver Risk Assessment Using Naturalistic Driving Data. *Accident Analysis & Prevention*. 2013. 61:3-9.
14. Liu, J., and A.J. Khattak. Delivering Improved Alerts, Warnings, And Control Assistance Using Basic Safety Messages Transmitted Between Connected Vehicles. *Transportation research part C: emerging technologies*. 2016. 68: 83-100.
15. Khattak, A.J., and B. Wali. Analysis of Volatility in Driving Regimes Extracted from Basic Safety Messages Transmitted Between Connected Vehicles. *Transportation research part C: emerging technologies*, 2017. 84: 48-73.
16. Jahangiri, A., V.J. Berardi, and S.G., Machiani.. Application of Real Field Connected Vehicle Data for Aggressive Driving Identification on Horizontal Curves. *IEEE Transactions on Intelligent Transportation Systems*, 2018. 19(7): 2316-2324.

17. Wang, W. and J. Xi. A Rapid Pattern-Recognition Method for Driving Styles Using Clustering-Based Support Vector Machines. In 2016 American Control Conference (ACC), IEEE, 2016. 5270-5275.
18. Meiring, G. and H. Myburgh. A Review of Intelligent Driving Style Analysis Systems and Related Artificial Intelligence Algorithms. *Sensors*, 2015. 15(12): 30653-30682.
19. Laureshyn A, C. Johnsson, T. De Ceunynck, Å. Svensson, M. de Goede, N. Saunier, and S. Daniels. Review of current study methods for VRU safety. Appendix 6–Scoping review: surrogate measures of safety in site-based road traffic observations. 2016.
20. He, Z., X. Qin, P. Liu and M.A. Sayed. Assessing Surrogate Safety Measures Using A Safety Pilot Model Deployment Dataset. Transportation Research Record, 2018. 2672(38): 1-11.
21. Xie, K,. D. Yang, K. Ozbay, and H. Yang. Use of Real-World Connected Vehicle Data in Identifying High-Risk Locations Based on A New Surrogate Safety Measure. *Accident Analysis & Prevention*. 2019. 125:311-319.
22. Yang, H., Z. Wang, and K. Xie. Impact of Connected Vehicles on Mitigating Secondary Crash Risk. *International Journal of Transportation Science and Technology*. 2017. 3: 196-207.
23. Abdel-Aty, M., and K. Haleem. Analyzing Angle Crashes at Unsignalized Intersections Using Machine Learning Techniques. *Accident Analysis & Prevention*, 2011. 43(1): 461-470.
24. Persaud, B, T. Nguyen. Disaggregate Safety Performance Models for Signalized Intersections on Ontario Provincial Roads. *Transportation Research Record*. 1998. 1635(1): 113-20.
25. Kamrani M., B. Wali, and A.J. Khattak. Can Data Generated by Connected Vehicles Enhance Safety?: Proactive Approach to Intersection Safety Management. *Transportation Research Record*. 2017. 2659(1): 80-90.
26. Kamrani M, R. Arvin, and A.J. Khattak. Extracting Useful Information From Basic Safety Message Data: an Empirical Study of Driving Volatility Measures and Crash Frequency at Intersections. *Transportation Research Record*. 2018. 2672 (38): 290-301
27. Wali B, A.J. Khattak, H. Bozdogan, and M. Kamrani. How is driving volatility related to intersection safety? A Bayesian heterogeneity-based analysis of instrumented vehicles data. *Transportation Research Part C: Emerging Technologies*. 2018.92:504-24.
28. Feng, Y., K.L. Head, S. Khoshmagham, and M. Zamanipour. A Real-Time Adaptive Signal Control in a Connected Vehicle Environment. *Transportation Research Part C: Emerging Technologies*. 2015. 55:460-73.
29. Guler, S.I., M. Menendez, and L. Meier. Using Connected Vehicle Technology to Improve the Efficiency of Intersections. *Transportation Research Part C: Emerging Technologies*. 2014. 46:121-31.
30. Lee, J., B.B. Park, K. Malakorn, and J.J. So. Sustainability Assessments of Cooperative Vehicle Intersection Control at an Urban Corridor. *Transportation Research Part C: Emerging Technologies*. 2013. 32:193-206.
31. Day, C.M., and D.M. Bullock. Detector-Free Signal Offset Optimization with Limited Connected Vehicle Market Penetration: Proof-of-Concept Study. *Transportation Research Record*. 2016. 2558(1):54-65.
32. Zheng, J., and H.X. Liu. Estimating Traffic Volumes for Signalized Intersections Using Connected Vehicle Data. *Transportation Research Part C: Emerging Technologies*. 2017. 79:347-62.

33. Newell, G.F. Approximation Methods for Queues with Application to the Fixed-Cycle 39 Traffic Light. *Siam Review*, 1965. 7(2): 223-240.
34. Mirchandani, P.B., and N. Zou. Queuing Models for Analysis of Traffic Adaptive Signal Control. *IEEE Transactions on Intelligent Transportation Systems*, 2007. 8(1): 50-59.
35. Chang, T.H. and J.T. Lin. Optimal Signal Timing for an Oversaturated Intersection. *Transportation Research Part B: Methodological*, 2000. 34(6): 471-491.
36. Zhao S, and K. Zhang. Observing Space-time Queueing Dynamics at a Signalized Intersection using Connected Vehicles as Mobile Sensors. InProc., 96th Annual Meeting of the Transportation Research Board. 2017.
37. Nodine, E., S. Stevens, A. Lam, C. Jackson, and W.G. Najm. Independent Evaluation of Light-Vehicle Safety Applications Based on Vehicle-To-Vehicle Communications Used in the 2012-2013 Safety Pilot Model Deployment. No. DOT HS 812 222. United States. National Highway Traffic Safety Administration, 2015.
38. Bridgelall, Raj, P. Lu, D.D. Tolliver, and T. Xu. Mining Connected Vehicle Data for Beneficial Patterns in Dubai Taxi Operations. *Journal of Advanced Transportation*, 2018 (2018).
39. Han, J, M. Kamber, and J. Pei. In Data mining: concepts and techniques. Elsevier, 2011.
40. Wang, W., and J. Xi. A rapid pattern-recognition method for driving styles using clustering-based support vector machines. In 2016 American Control Conference (ACC), IEEE. 2016:5270-5275.,
41. Jahangiri, A., and H. A. Rakha. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE transactions on intelligent transportation systems.* 2015.16(5): 2406-2417.