Accident Analysis and Prevention 125 (2019) 311–319

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Use of real-world connected vehicle data in identifying high-risk locations based on a new surrogate safety measure

Kun Xie[a], Di Yang[b,*], Kaan Ozbay[b], Hong Yang[c]

[a] Department of Civil and Natural Resources Engineering, University of Canterbury, 20 Kirkwood Ave, Christchurch 8041, New Zealand
[b] Department of Civil and Urban Engineering, Connected Cities for Smart Mobility towards Accessible and Resilient Transportation (C2SMART) Center, and Center for Urban Science and Progress (CUSP) New York University, 370 Jay Street, Brooklyn, NY 11201, United States
[c] Department of Modeling, Simulation & Visualization Engineering, Old Dominion University, 4700 Elkhorn Ave, Norfolk, VA 23529, United States

ARTICLE INFO

Keywords:
Connected vehicles
Identify high-risk locations
Surrogate safety measure
Proactive safety solutions

ABSTRACT

Traditional methods for the identification of high-risk locations rely heavily on historical crash data. Rich information generated from connected vehicles could be used to obtain surrogate safety measures (SSMs) for risk identification. Conventional SSMs such as time to collision (TTC) neglect the potential risk of car-following scenarios in which the following vehicle's speed is slightly less than or equal to the leading vehicle's but the spacing between two vehicles is relatively small that a slight disturbance would yield collision risk. To address this limitation, this study proposes time to collision with disturbance (TTCD) for risk identification. By imposing a hypothetical disturbance, TTCD can capture rear-end conflict risks in various car following scenarios, even when the leading vehicle has a higher speed. Real-world connected vehicle pilot test data collected in Ann Arbor, Michigan is used in this study. A detailed procedure of cleaning and processing the connected vehicle data is presented. Results show that risk rate identified by TTCD can achieve a higher Pearson's correlation coefficient with rear-end crash rate than other traditional SSMs. We show that high-risk locations identified by connected vehicle data from a relatively shorter time period are similar to the ones identified by using the historical crash data. The proposed method can substantially reduce the data collection time, compared with traditional safety analysis that generally requires more than three years to get sufficient crash data. The connected vehicle data has thus shown the potential to be used to develop proactive safety solutions and the risk factors can be eliminated in a timely manner.

1. Introduction

Accurate identification of high-risk locations can result in efficient allocation of government resources for safety treatments. Traditional methods for the identification of high-risk locations rely heavily on historical crash data. High-risk locations cannot be detected until a sufficient amount of crashes happen, which often requires a significant amount of time. A more proactive and in some conditions time-efficient approach is based on surrogate safety measures (SSMs), which can assess traffic risks by capturing the more frequent "near-crash" situations. The idea behind this approach is that SSMs are causally related to accidents or injuries and they can indicate safety related performance and help understand the process that leads to accidents (Davis et al., 2011; Laureshyn et al., 2016). Since its first introduction in the 1950-60 s (Forbes, 1957; Perkins and Harris, 1967), many studies have been undertaken over the years to use SSMs as alternatives to evaluate the traffic safety performance (Gao et al., 2018; Kurkcu, 2018; Xu et al., 2012; Yang, 2012). A detailed scoping review of the development of SSMs and current study methods were summarized in a newly published report funded by the European Union (Laureshyn et al., 2016). Besides summarizing different types of trends of SSM studies and brief discussion of each typical SSM, it pointed out that because of the great improvement of sensor technologies in the last decade, data collected from these sensors opens new opportunities but also poses new questions and challenges for SSM studies.

One of the most commonly used tools is video recording. Several recent studies have been focused on using computer vision technics to detect and track vehicles as well as other road users in videos and studying conflicts within a complex traffic environment (Ismail et al., 2009; Laureshyn and Ardö, 2006; Xie et al., 2016). These studies have shown a good correlation between the risks identified by SSMs and crash data or manually extracted sample events by human observers according to certain conflict standards. It is well accepted in literature that SSM-based approach can help identify high-risk locations before

[*] Corresponding author.
E-mail addresses: kun.xie@canterbury.ac.nz (K. Xie), dy855@nyu.edu (D. Yang), kaan.ozbay@nyu.edu (K. Ozbay), hyang@odu.edu (H. Yang).

https://doi.org/10.1016/j.aap.2018.07.002
Received 14 February 2018; Received in revised form 11 May 2018; Accepted 1 July 2018
Available online 06 July 2018

Nomenclature

d	The deceleration rate of the leading vehicle
d^*	The deceleration rate of the leading vehicle that causes the collision to occur exactly when the leading vehicle stops
v_1	Initial speed of the leading vehicle
v_2	Initial speed of the following vehicle (remains constant in the car following scenario)
t_0	The time when a disturbance is given
t^*	The time interval between the given of the disturbance and the full stop of the leading vehicle
l_v	The length of the vehicle
l_0	Initial relative distance between the leading vehicle and the following vehicle
l_1	Distance travelled by the leading vehicle after being given a disturbance
l_2	Distance travelled by the following vehicle before colliding with the leading vehicle after a disturbance has been given to the leading vehicle
N	Number of Monte Carlo simulations
TTCD	The time interval between the given of the disturbance and the collision of the two vehicles

the occurrence of actual crashes.

Further, the emerging connected vehicle (CV) technologies are envisioned to transform the future transportation systems. CVs are generally equipped with devices such as GPS loggers and cameras and can act as mobile sensors in the road network. Rich information generated by these vehicles then can be used to detect traffic risks in real time. For example, if hard breakings were observed frequently at one road segment, we can infer that it could have design deficiencies. Combining with surrogate measures, high-risk locations can be identified and predicted in a timely manner even without crash data. For example, Yang et al. (2017) used the CV data to examine surrogate measures in analysing the risk of secondary crashes on highways. Proactive countermeasures for mitigating the risks and impacts of traffic crashes can be developed in advance.

Under the supervision of U.S. Department of Transportation (USDOT), Connected Vehicle Safety Pilot is a research initiative that features real-world implementation of connected vehicle safety technologies, applications, and systems (Henclewood et al., 2014). As a part of this research initiative, the Safety Pilot Model Deployment (SPMD) in Ann Arbor, Michigan hosted approximately 3000 vehicles equipped with vehicle-to-vehicle (V2V) communication devices (Henclewood et al., 2014). These vehicles generated a large volume of high-resolution data from examining their operational and safety performance.

This study aims to exploit the potential of using data generated by CVs to develop more proactive safety solutions. Specifically, the real-world CV pilot test data from SPMD is used in this study. A detailed procedure of data cleaning and processing is presented. A novel SSM is proposed to overcome the shortcomings of existing ones. The proposed SSM is justified by verifying its correlation with historical crash data. High risk locations identified by the SSM are also inspected.

2. Literature review

2.1. Surrogate safety measures

Time to collision (TTC), which was proposed by Hayward (1972), is one of the most widely used SSMs. TTC is defined as the time required for two vehicles to collide if they continue at their present speeds and on the same path. Mathematically, it is given by:

$$TTC = \begin{cases} \frac{l_0 - l_v}{v_2 - v_1} & v_2 > v_1 \\ \infty, & \text{otherwise} \end{cases} \tag{1}$$

where l_0 is the initial relative distance between the leading and following vehicles, l_v is the length of the vehicle, v_1 is the initial speed of the leading vehicle, and v_2 is the initial speed of the following vehicle. A car-following scenario is considered as unsafe if its TTC value is less than a certain threshold value. Due to its simplicity, TTC has been widely used to evaluate the risk of car-following scenarios (van der Horst and Hogema, 1993; Yang, 2012). As derivatives of TTC, time exposed TTC (TET) and time integrated TTC (TIT) were introduced by Minderhoud and Bovy (2001) to assess risk for one driver during a certain period of time for a specific trip. They were shown to be useful surrogate measures in microsimulation studies that focus on safety impacts (Ozbay et al., 2008). Ozbay et al. (2008) modified TTC by including accelerations/decelerations rate of the leading and following vehicles into consideration and derived a new surrogate measure, namely MTTC. It has been confirmed that MTTC has the capability to highlight the high-risk locations based on simulation studies (Yang et al., 2010).

Deceleration rate to avoid crash (DRAC) is another widely used surrogate measure. It is defined as the minimum deceleration rate required by the following vehicle to come to a timely stop (or match the leading vehicle's speed) and hence to avoid a crash (Cooper and Ferguson, 1976). Mathematically, DRAC is defined as:

$$DRAC = \begin{cases} \frac{(v_2 - v_1)^2}{l_0 - l_v} & v_2 > v_1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

A vehicle is at risk if its DRAC value exceeds a certain threshold value. It is recognized as an effective measure of safety performance and seems to be closer to crash mechanism by considering the evasive action taken by the following vehicle (Kuang and Qu, 2014).

Based on DRAC, crash potential index (CPI) is introduced by Cunto and Saccomanno (2007) to estimate the probability that a given vehicle's DRAC exceeds its maximum available deceleration rate (MADR). Different from aforementioned surrogate measures, CPI gives the probability of a vehicle involving in conflict risk, and it does not require the selection of a certain threshold value. Based on the assumption that MADR follows a truncated normal distribution, Souza et al. (2011) calibrated the distributions of MADR for cars, trucks and buses using field tests data, which includes initial speeds of vehicles ranging from 80 to 100 km/h and the distances travelled by these vehicles to come to a full stop.

Kuang et al. (2015) summarized three major shortcomings of the SSMs mentioned above: (a) any scenarios in which the following vehicle's speed is lower than or equal to the leading vehicle's speed are regarded as safe even when the spacing between the two vehicles is very small; (b) driver's reaction time is not taken into consideration; (c) the arbitrary selection of threshold might result in inaccurate outcomes.

Kuang et al. (2015) proposed to use aggregated crash index (ACI) for risk evaluation, which overcomes these three shortcomings. By imposing a hypothetical disturbance to the leading vehicle, which makes it to decelerate at a constant deceleration rate, the potential collision risk can be captured for the car-following scenarios when the following vehicle's speed is less than or equal to the leading vehicle's speed. The reaction time and MADR of the following vehicle, which follows two different distributions, are considered in evaluating evasive actions taken by the following vehicle after imposing the disturbance. ACI captures the probability of collision thus a threshold for risk identification is not needed. ACI was calculated using the Monte Carlo simulation method and its performance was validated in a micro traffic simulation model. The results showed that ACI outperformed traditional surrogate measures in achieving higher correlation with rear-end crashes. However, the ACI is not very intuitive and can be difficult to compute. Moreover, the value of ACI is subject to the selection of distributions of reaction time and MADR.

2.2. Previous studies using SPMD data

The United States Department of Transportation's Volpe National Transportation Systems Center (Volpe) conducted an independent evaluation of crash warning applications based on V2V communications for the National Highway Traffic Safety Administration (NHTSA), using data collected from the SPMD field test (Nodine et al., 2015). The capability of the V2V applications and drivers' reaction to warnings were exploited. Data measurement and collection errors were reported in this evaluation. A GPS positioning error was found to cause a missed blind spot/lane change warning (BSW/LCW) alert and it is presumed that more missed alerts might potentially occur. A programming error was found within the safety applications that on rare occasions caused a larger velocity value to be transmitted over-the-air (OTA) than the actual velocity reading. Clearly, it is necessary to perform data quality check before further investigation.

A portion of the SPMD CV test data is available to the public and was used in some recent studies. One Day Sample data collected from participating vehicles equipped with Data Acquisition System (DAS) on April 11, 2013 was used by Liu and Khattak (2016) to capture extreme driving events. Basic Safety Message (BSM) dataset collected in October 2012 and April 2013 was used by Kamrani et al. (2017), Kamrani et al. (2018) and Liu and Khattak (2018) to study location based volatility in the Ann Arbor area and by Zhang and Khattak (2018) to identify extreme lane change events. One month BSM data collected in April 2013 was used by Mousa and Ishak (2017) to estimate travel time on the selected highways and by Machiani et al. (2017) to develop a driver behavior model to predict risky behaviors at horizontal curves. Data collected by road-side units (RSU) in two months was used by Zhao and Zhang (2017) to observe queueing dynamics at signalized intersections and by Zheng and Liu (2017) to estimate traffic volumes for signalized intersections along with signal status data from Signal Phase and Timing (SPaT) messages. Despite of the necessity of data quality check, only Liu and Khattak (2016) and Kamrani et al. (2017) mentioned that they checked the data quality based on the descriptive statistics of key variables. No detailed procedure of data cleaning and processing was presented in the aforementioned studies.

3. Time to collision with disturbance (TTCD)

To address the shortcomings of SSMs employed in previous studies, we propose a new SSM namely, time to collision with disturbance (TTCD) to capture the risks of rear-end collisions. Similar to ACI, a hypothetical disturbance is imposed on the leading vehicle, so TTCD can capture the risk of car-following scenarios when the speed of the following vehicle is not higher than that of the leading vehicle. Intuitively, TTCD can be regarded as the time to collision with a disturbance imposed on the leading vehicle. Compared with ACI, TTCD is more intuitive since it is based on the well-known and widely used concept of TTC. Also, it is much easier to compute than ACI without including reaction time and MADR as random variables.

Let's assume that a car following scenario shown in Fig. 1. A disturbance is imposed on the leading vehicle at time t_0. It is assumed that, after being given the disturbance, the leading vehicle will decelerate at a constant rate d until it is fully stopped. d is the deceleration rate of the leading vehicle, which follows a certain probability distribution. If the speed of the following vehicle remains the same throughout the entire car following process, a collision will eventually occur after a certain time. This time duration is defined as TTCD.

Fig. 1 depicts the vehicle movements after imposing a disturbance on the leading vehicle. The relationship described by Eq. (3) holds for all possible collision outcomes.

$$l_0 + l_1 = l_2 + l_v \tag{3}$$

There are two possible collision outcomes, depending on the status of the leading vehicle: (a) collision outcome 1 – the leading vehicle is still decelerating when collision occurs; and (b) collision outcome 2 – the leading vehicle is fully stopped when collision occurs. The critical scenario to distinguish between these two collision outcomes is that the following vehicle collides with the leading vehicle exactly at the time when the leading vehicle stops. The distances travelled by the leading vehicle l_1, the distance travelled by the following vehicle l_2, and the time it takes for the leading vehicle to stop t^* in this critical scenario are given by Eqs. (4)–(6), respectively:

$$l_1 = \frac{v_1^2}{2d^*} \tag{4}$$

$$l_2 = v_2 t^* \tag{5}$$

$$t^* = \frac{v_1}{d^*} \tag{6}$$

where d^* is the deceleration rate of the leading vehicle that satisfies the critical condition mentioned above. Substituting Eqs. (4)–(6) into (3), we can get:

$$d^* = \frac{2v_1v_2 - v_1^2}{2(l^0 - l_v)} \tag{7}$$

Based on the deceleration rate d imposed by the disturbance and the critical deceleration rate d^*, there are two possible collision outcomes.

Collision outcome 1: $d \le d^* = \dfrac{2v_1v_2 - v_1^2}{2(l_0 - l_v)}$

Collision outcome 1 represents the scenario where the following vehicle collides with the leading vehicle before/when the leading vehicle fully stops. The distances travelled by the leading and the following vehicle in this scenario are given by Eqs. (8) and (9):

$$l_1 = v_1 t_1 - \frac{1}{2} d t_1^2 \tag{8}$$

$$l_2 = v_2 t_1 \tag{9}$$

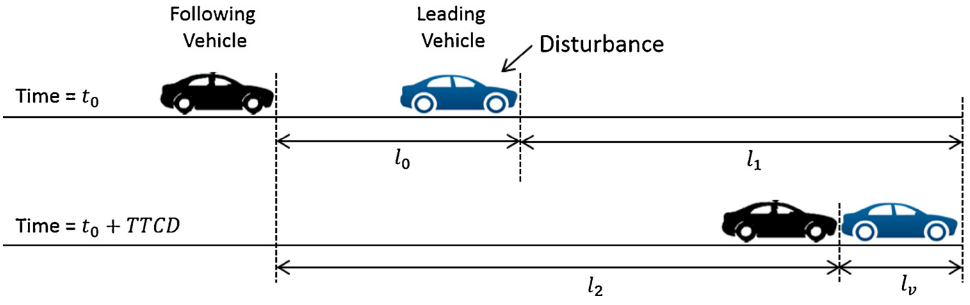

Fig. 1. A car following scenario after imposing a disturbance on the leading vehicle.

By substituting Eqs. (8) and (9) into Eq. (3), we get:

$$TTCD = \frac{(v_1 - v_2) + \sqrt{(v_1 - v_2)^2 + 2d(l_0 - l_v)}}{d} \tag{10}$$

Collision outcome 2:$d > d^* = \dfrac{2v_1v_2 - v_1^2}{2(l_0 - l_v)}$

Collision outcome 2 represents the scenario in which the following vehicle collides with the leading vehicle after the leading vehicle fully stops. The distances travelled by the leading and the following vehicles in this scenario are shown by Eqs. (11) and (12):

$$l_1 = \frac{v_1^2}{2d} \tag{11}$$

$$l_2 = v_2 t_1 \tag{12}$$

By substituting Eqs. (11) and (12) into Eq. (3), we get:

$$TTCD = \frac{2d(l_0 - l_v) + v_1^2}{2dv_2} \tag{13}$$

Eq. (14) summarizes the computation of TTCD for the two above scenarios:

$$TTCD = \begin{cases} \dfrac{(v_1 - v_2) + \sqrt{(v_1 - v_2)^2 + 2d(l_0 - l_v)}}{d}, & d \le \dfrac{2v_1v_2 - v_1^2}{2(l_0 - l_v)} \\ \dfrac{2d(l_0 - l_v) + v_1^2}{2dv_2}, & d > \dfrac{2v_1v_2 - v_1^2}{2(l_0 - l_v)} \end{cases} \tag{14}$$

To differentiate risky encounters from situations where the driver remains safely in control, an appropriate threshold TTCD* must be defined. TTCD* is related to the perception and reaction times as well as driving conditions. Similar to the approach of identifying conflicts with TTC, vehicle pairs with TTCD lower than TTCD* are involved with conflicts. Clearly, TTCD is not only related to the relative speed and distance between the two vehicles, but also the speeds of the leading and following vehicles and deceleration rate of the leading vehicle. In this case, car following scenarios with the initial speed of the following vehicle less than or equal to the initial speed of the leading vehicle could still yield risks rather than always regarded as safe by methods based on TTC and DRAC. Note that when the speed of the following vehicle v_2 is zero, TTCD will be infinity, which will not lead to conflict.

The disturbance imposed has a direct impact on TTCD. Disturbances are random in nature and it will lead to random deceleration rates for the leading vehicle. Therefore, it is necessary to include all possible deceleration rates into consideration for each car following scenario. In this study, we assume that the deceleration rate of the leading vehicle follows a shifted gamma distribution (17.315, 0.128, 0.657), which is calibrated by Kuang and Qu (2014) by analyzing the normal deceleration rate taken due to the lane-changing maneuver on freeways using the Next Generation Simulation (NGSIM) data (FHWA 2005).

A Monte Carlo simulation method was applied to calculate TTCD for each deceleration rate sample for i^{th} car following scenario. Setting a threshold value of TTCD*, we obtained the proportion of samples with TTCD below the TTCD* out of all the samples generated by Monte Carlo method. Conflict Risk with Disturbance (CRD), which indicates the probability of being involved with conflict under hypothetical disturbance, is defined in Eq. (15):

$$CRD_i = \frac{N_i(TTCD < TTCD^*)}{N} \tag{15}$$

where CRD_i indicates the probability of involving into a conflict for i^{th} car following scenario, $N_i(TTCD < TTCD^*)$ is the number of samples with TTCD below the TTCD* for i^{th} car following scenario, and N is the total number of samples generated by the Monte Carlo simulation method. We choose $N = 10,000$, which is sufficient to depict the distribution of the deceleration rate. It is worth to mention that CRD is a continuous variable, ranging from 0 to 1, and thus has better flexibility to quantify the risk. On the contrary, the TTC- and DRAC-based methods use the presence of conflict to quantify the risk, which is a dummy variable.

4. Data preparation

4.1. SPMD data

SPMD is a comprehensive data collection effort, under real-world conditions. Two sample data, namely SPMD One Day Sample (April 11, 2013) and SPMD Two Months Sample (October 2012 and April 2013), are open to the public on the ITS Data Hub of US Department of Transportation (https://www.its.dot.gov/data/). The safety pilot environment contains eight datasets, namely Data Acquisition System 1 (DAS1), Data Acquisition System 2 (DAS2), Basic Safety Message (BSM), Roadside Equipment, Network, Weather, Schedule, and Road-Work Activity. The DAS1 dataset, collected during April 2013 by the University of Michigan Transportation Research Institute (UMTRI), was used in this study. A total of 62,589,725 messages, collected by 90 vehicles equipped with DAS from April 1 to April 30, 2013 were investigated. Within the DAS1 dataset, the *DataFrontTargets* file can be used to detect traffic risks and *DataWsu* file provides information to

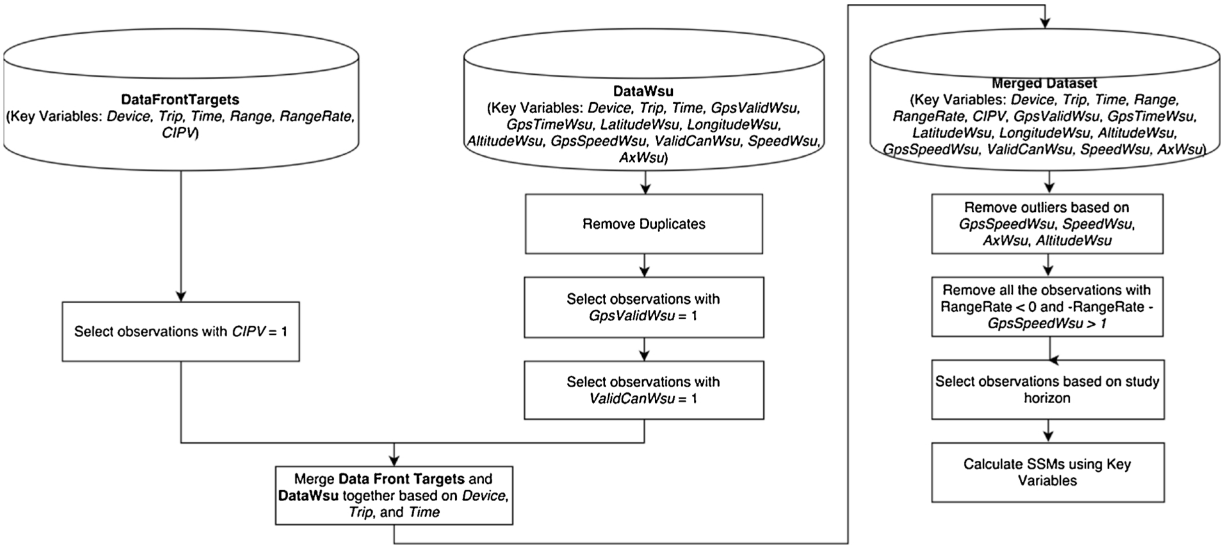
DataFrontTargets
(Key Variables: Device, Trip, Time, Range, RangeRate, CIPV)
Select observations with CIPV = 1
DataWsu
(Key Variables: Device, Trip, Time, GpsValidWsu, GpsTimeWsu, LatitudeWsu, LongitudeWsu, AltitudeWsu, GpsSpeedWsu, ValidCanWsu, SpeedWsu, AxWsu)
Remove Duplicates
Select observations with GpsValidWsu = 1
Select observations with ValidCanWsu = 1
Merge Data Front Targets and DataWsu together based on Device, Trip, and Time
Merged Dataset
(Key Variables: Device, Trip, Time, Range, RangeRate, CIPV, GpsValidWsu, GpsTimeWsu, LatitudeWsu, LongitudeWsu, AltitudeWsu, GpsSpeedWsu, ValidCanWsu, SpeedWsu, AxWsu)
Remove outliers based on GpsSpeedWsu, SpeedWsu, AxWsu, AltitudeWsu
Remove all the observations with RangeRate < 0 and -RangeRate - GpsSpeedWsu > 1
Select observations based on study horizon
Calculate SSMs using Key Variables

Fig. 2. Data processing procedure.

associate the detected risks with specific locations. As discussed in Literature Review, data quality check should be performed before using this data for further analysis. Thus, a detailed procedure of cleaning and integrating those two files is illustrated in Fig. 2. R Programming Language (Team, 2013) was used for data manipulation and ArcGIS software (ESRI, 2011) was used for spatial processing.

DataFrontTargets file contains the data collected by Mobileye sensors (Mobileye 2017) at a sample rate of 10 Hz. Mobileye sensors track all the objects in front of the host vehicles using vision-based technologies. The relative distance (*Range*) and relative speed (R*angeRate*) between each connected vehicle and its leading vehicles were recorded in *DataFrontTargets* file. The field *CIPV* (Closest In Path Vehicle) indicates whether a target is the closest in a vehicle's path. The criterion "*CIPV* = 1" was used to ensure that we were investigating the risk between each host vehicle and its closest leading vehicle in the same path.

DataWsu file logs GPS and Controller Area Network (CAN) Bus via the onboard Wireless Safety Units (WSUs) at a sample rate of 10 Hz. It contains the movement information of host vehicles, including position (*LatitudeWsu* and *LongitudeWsu*), epoch time (*GpsTimeWsu*), speed (*GpsSpeedWsu* and *SpeedWsu*) and longitudinal acceleration (*AxWsu*). It should be noted that *GpsSpeedWsu* was obtained from WSU GPS receiver and *SpeedWsu* was obtained from vehicle CAN Bus. Generally, *GpsSpeedWsu* provides better measurements of speed. As the first step of data cleaning procedure, duplicated records were removed. Then we used the criterion "GpsValidWsu = 1 and ValidCanWsu = 1" to filter out records with invalid WSU or CAN Bus messages. The *DataFrontTargets* file was merged with the *DataWsu* file based on the fields *Device*, *Trip* and *Time.* The field description of the merged dataset is reported in Table 1.

Erroneous data points were removed based on the values of *GpsSpeedWsu*, *SpeedWsu*, *AxWsu*, and *AltitudeWsu.* According to the information regarding latest automative news and technologies (Car and Driver, 2014; Production Car Speed Record, 2017), the fastest car in the world can travel at a speed of 415 miles/hour (115 m/s) and the maximum acceleration rate of vehicles is around 10 m/s^2. Measurements of speeds higher than 115 m/s and the absolute values of acceleration rate greater than 10 m/s^2 were removed as outliers. The average elevation in Ann Arbor is 267 m (LatLong.net, 2017). The values of *AltitudeWsu* were within a reasonable range from 128.51 m to 305.65 m and thus no outliers were detected.

Additionally, the accuracy of *Rangerate*, which was defined as the speed of the leading vehicle minus the speed of the following vehicle, was examined. Therefore, if *Rangerate* is negative and the absolute value of *Rangerate* is greater than the host vehicle's speed (*GpsSpeedWsu*), this indicates that the leading vehicle is backing up at a speed of *Rangerate* - *GpsSpeedWsu*. According to the SPMD Data Handbook (Henclewood and Rajiwade, 2015), a Mobileye camera is able to cover three or more lanes in front of host vehicles, so it is likely that it can detect vehicles from the opposite directions if the road is narrow and does not have a median. Moreover, in the cases that both the host vehicle and leading vehicle are running at the same speed, it is possible that the *Rangerate* is negative and *Rangerate – GpsSpeedWsu* is greater than zero due to the measurement errors. The criterion "*Rangerate – GpsSpeedWsu > 1*" (i.e., backing up at a speed higher than 1 m/s) was used to rule out records which actually described the vehicle movements in the opposite direction.

There are totally 15,721,962 GPS points after the data cleaning process. The heat map of all the GPS data is shown in Fig. 3 (histogram equalization is used to transform density). Most vehicle GPS points are located around the city of Ann Arbor. The roads with bright colour are mostly interstate, U.S. and state highways.

4.2. Traffic and crash data

A total of 75 highway segments that have complete traffic data were selected for analysis, as shown in Fig. 6. Annual average daily traffic (AADT) of those highways in 2013 were obtained from the Michigan Department of Transportation (MDOT) Open Data (http://gis-mdot.opendata.arcgis.com/). Crash dataset was obtained from the Michigan's Open Data Portal (https://data.michigan.gov/). Based on the available information, the issue of underreporting is unknown at this time. Crash occurrence time, location, three levels of severity namely, fatal, injury, and property damage only (PDO), and other characteristics were included in the crash dataset. A total of 2323 crashes of all three severity levels occurred on the selected highways in Ann Arbor in 2013 and 1027 of them were rear-end crashes.

5. Analysis and results

5.1. Determination of optimal SSM thresholds

As discussed in the literature review section, the arbitrary setting of thresholds is another shortcoming when applying SSMs to identify risks. This section proposes a method to identify the optimal thresholds by maximizing the correlation coefficients between risk and rear-end crash data.

Rear end crashes and GPS points collected by CVs were assigned to the nearest road segments, if the distances to the nearest road segments

Table 1
Description of the Fields of the Merged Dataset (based on Henclewood and Rajiwade (2015)).

Origin File	Field Name	Description
DataFrontTargets & DataWsu	Device	A unique numeric ID assigned to each DAS. This ID also doubles as a vehicle's ID
	Trip	Count of ignition cycles—each ignition cycle commences when the ignition is in the on position and ends when it is in the off position
	Time	Time in centiseconds since DAS started, which (generally) starts when the ignition is in the on position (centiseconds)
DataFrontTargets	Range	Longitudinal position of an object, typically the closest object, relative to a reference point on the host vehicle, according to the Mobileye sensor (m)
	RangeRate	Longitudinal velocity of an object, typically the closest object, relative to the host vehicle, according to the Mobileye sensor (m/sec)
	CIPV	Field communicating whether a target is the closest in a vehicle's path 1: Identified target is the closest in a vehicle's path; 0: otherwise
DataWsu	GpsValidWsu	1: GPS data point is valid; 0: otherwise
	GpsTimeWsu	Epoch GPS time received from the remote vehicle that has been targeted by the host vehicle's WSU (millisecond)
	LatitudeWsu	Latitude from WSU receiver (degree)
	LongitudeWsu	Longitude from WSU receiver (degree)
	AltitudeWsu	Altitude from WSU receiver (m)
	GpsSpeedWsu	Speed from WSU GPS receiver (m/sec)
	ValidCanWsu	1: Valid Vehicle CAN Bus message to WSU; 0: otherwise
	SpeedWsu	Speed from vehicle CAN Bus via WSU (kph)
	AxWsu	Longitudinal acceleration from vehicle CAN Bus via WSU (m/sec2)

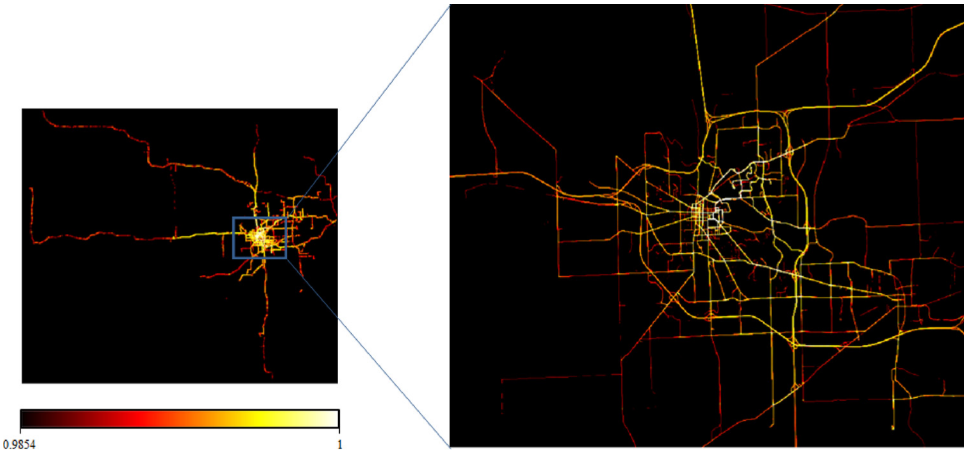

Fig. 3. Heat map of trajectories in the cleaned connected vehicle dataset.

are less than 10 m. To obtain the optimal SSM thresholds, every possible threshold value incremented by 0.1 within a reasonable range was tested. Each GPS point will be given a risk value based on the current threshold used. If we use TTC and DRAC for risk identification, the risk value associated with each GPS point is the presence of conflict (dummy variable: 0 for no conflict, 1 for conflict); whereas if we use the proposed TTCD for risk identification, the risk value associated with each GPS point is CRD (continuous variable, ranging from 0 to 1). When aggregating these risk values by road segment, we obtained the aggregated risk for the road segment i, $AggRisk_i$. The rear-end crash count for the road segment i, $CrashCount_i$ was also obtained. To account for the effect of exposure indicators, $RiskRate_i$ and $CrashRate_i$ are calculated for the road segment i using Eqs. (16) and (17), respectively:

$$RiskRate_i = \frac{AggRisk_i}{NumGPS_i} \tag{16}$$

$$CrashRate_i = \frac{CrashCounts_i}{AADT_i} \tag{17}$$

where i is the road segment index, $NumGPS_i$ is the number of GPS points assigned to the road segment i, and $AADT_i$ is the annual average daily traffic volume for the road segment i.

For each possible threshold value, we computed the Pearson's correlation coefficient (Pearson, 1895) between $RiskRate_i$ and $CrashRate_i$ for three SSMs, TTC, DRAC and TTCD, as shown in Fig. 4. Looking at each subplot separately, the optimal thresholds to achieve the highest correlation coefficients are 2.3 s for TTC, 3.0 m/s^2 for DRAC and 1.7 s for TTCD, respectively. Correlation significance tests were performed for all three SSMs when applying optimal thresholds. The p-values of the Pearson's correlation coefficient between SSMs and rear-end crashes were calculated. As shown in Table 2, all three p-values are less than 0.05, which indicates statistically significant correlation between SSMs and rear-end crashes. It is found that risk identified by TTCD can result in the highest correlation coefficient 0.45 among all the three SSMs.

To further validate our method of finding the optimal method, we tested its performance by using 30% of the segments that have not been used in training process. All the segments in this study were split into training and testing sets. To reduce the effect of randomness, the training and testing split was repeated 30 times. For each time, the optimal threshold for TTCD obtained from the training set was applied to the testing set to calculate correlation within the testing set only. The average correlation between observed crashes and calculated SSMs from the training set is found to be 0.58 while the average correlation

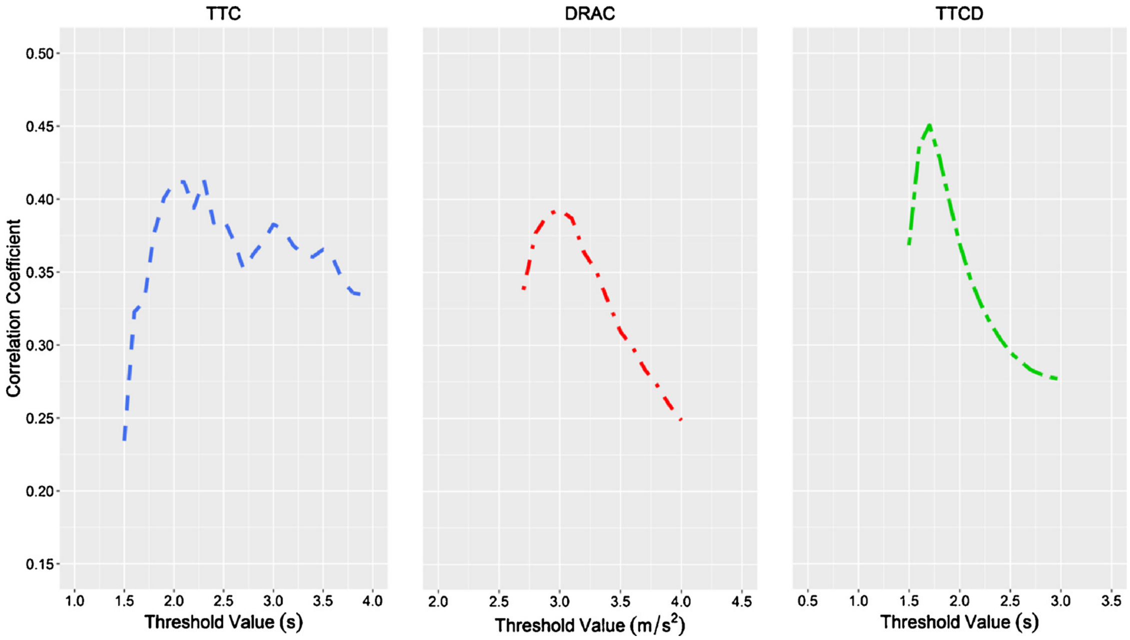

Fig. 4. Correlation coefficient between crash rate and risk rate using different threshold values.

Table 2
Correlation Significance Test Summary.

	Optimal Threshold	Pearson's correlation coefficient	P-value
TTC	2.3 s	0.41	0.0002
DRAC	3.0 m/s^2	0.39	0.0005
TTCD	1.7 s	0.45	0.0000

using the data in the testing set is found to be 0.46. The decrease in average correlation when using the testing sets is expected but this decrease that is relatively small indicates a stable results.

5.2. Risk identification using SPMD data

To better illustrate the risk identification approach proposed in this paper, TTC, DRAC and TTCD were used to identify risks for one trip (Device = 10181, Trip = 119). Optimal threshold value was applied for each SSM (i.e. $TTC^* = 2.3\,s$, $DRAC^* = 3.0\,m/s^2$, $TTCD^* = 1.7\,s$). Risk identified for each 0.1 s time interval is shown in Fig. 5. Clearly, TTCD can capture risks that cannot be identified by TTC and DRAC. For example, both the TTC and DRAC ignored the potential risk from 3762th to 3895 th time intervals, when the speed of the following vehicle is slightly less than that of the leading vehicle and the two vehicles are very close to each other. A portion of vehicle movement data during that period is reported in Table 3. Intuitively, conflict risks could exist, because the distance between the leading and following vehicles is less than 3 m and the vehicles are moving at a speed higher than 13 m/s. When using TTC and DRAC for risk identification, if the following vehicle is slower than the leading vehicle (i.e., $v_2 < v_1$), no conflict is detected, according to Eqs. (1) and (2). By contrast, The CRD identified by TTCD for the same period is greater than zero. TTCD could capture risks in car following scenarios, even when the leading vehicle has a higher speed.

Fig. 6 shows the spatial distribution of rear-end crashes in 2013, risks in April, 2013 identified by TTCD on the selected highways, and AADT. R package ggmap (Kahle and Wickham, 2013) was used for data visualization. Darker color indicates high density of crashes (in Fig. 6a), high aggregated risks (in Fig. 6b), and higher value of AADT (in Fig. 6c). High-risk locations identified by connected vehicle data was found to be similar to the ones identified by historical rear-end crash data for one year, even though risk data is extracted from only one month of CV data. It shows the potential of using CV data to detect risk and thus supports more proactive safety management. In this paper, in contrast with the distribution of AADT values, the distribution of conflicts is observed to be more similar to the distribution of crashes. This indicates that in this study conflicts identified by TTCD have greater potential to infer crash risks than AADT values. Further, this observation gives new insights to the discussion on whether traffic conflicts contribute sufficiently valuable additional information that justifies the costs incurred for their collection. From the findings of this study, it seems that it is not enough to model crashes using AADT values only, which is also the point made by Guttinger (1982).

6. Conclusions

This study aims to explore the potential of using the CV data to identify high-risk locations in a more proactive manner, without relying on the historical crash data that often takes long time to collect. Real-world CV pilot test data collected in Ann Arbor, Michigan was used to generate SSMs for risk identification. A detailed procedure of cleaning and processing the SPMD data is presented in this study, which can serve as a guidance for other researchers who want to use this dataset. Traditional SSMs like TTC and DRAC consider any scenarios when following vehicle's speed is lower than the leading vehicle's as safe, which ignores the potential risks of scenarios when following vehicle's speed is lower but the distance between vehicles is small and speeds of vehicles are high. By imposing a hypothetical disturbance, TTCD is able to detect risks in various car following scenarios, even when the leading vehicle has a higher speed. Optimal threshold values for TTC, DRAC and TTCD were determined by maximizing the correlation coefficient between risk identified by those SSMs and rear-end crash rate. Results showed that risk data captured by TTCD could achieve the highest level of correlation with historical rear-end crash data compared with other traditional SSMs. In addition, high-risk locations identified by connected vehicle data was found to be similar to the ones identified by

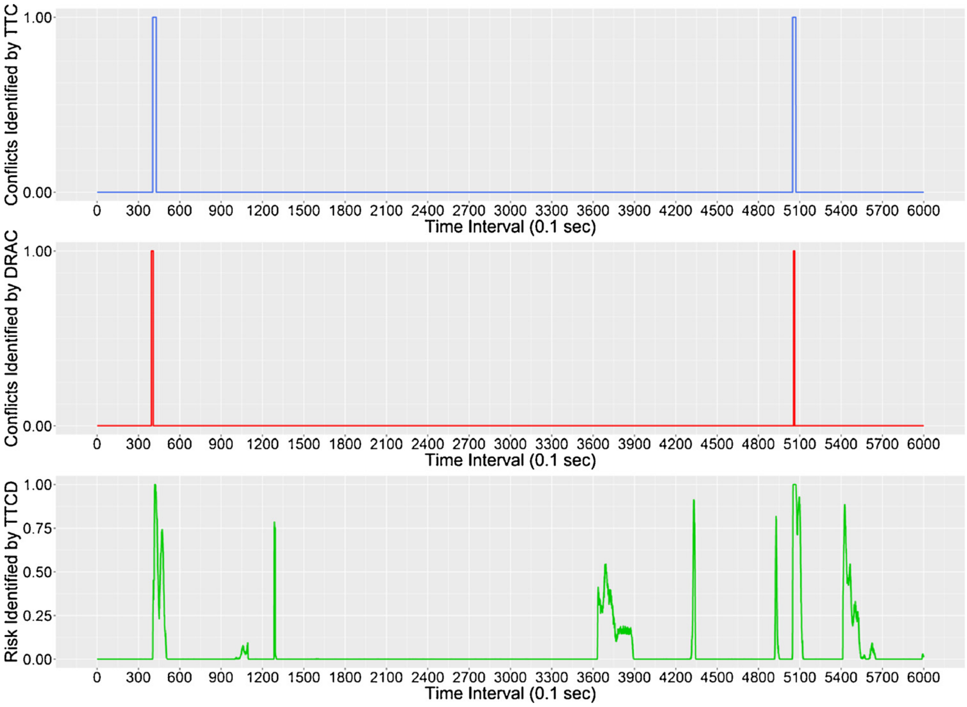

Fig. 5. Risks measured by TTC, DRAC and TTCD for one trip.

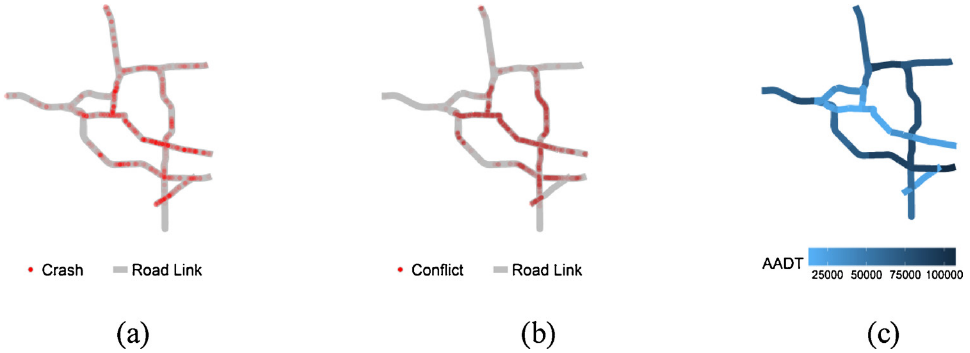

Fig. 6. High-risk locations identified by historical rear-end crash data (a) and by TTCD proposed (b).

Table 3
Vehicle Movement Data and Identified Risks from 3762th to 3768th Time Intervals.

Time Interval	Relative Distance (m)[a]	Relative Speed (m/s)[b]	Following Vehicle Speed (m/s)	Conflict Presence Identified by TTC	Conflict Presence Identified by DRAC	CRD Identified by TTCD
3762	2.77	0.21	13.05	0	0	0.14
3763	2.79	0.21	13.09	0	0	0.07
3764	2.81	0.21	13.07	0	0	0.14
3765	2.83	0.21	13.11	0	0	0.11
3766	2.85	0.21	13.17	0	0	0.17
3767	2.88	0.19	13.21	0	0	0.10
3768	2.90	0.17	13.31	0	0	0.08
…	…	…	…	…	…	…

[a] Relative distance is the distance between the leading vehicle's rear bumper and the following vehicle's front bumper. It is the same as "Range" in Table 1.

[b] Relative speed is the speed of the leading vehicle minus the speed of the following vehicle.

historical rear-end crash data. One major limitation of traditional safety analysis is that the time period required to collect sufficient crash data is generally more than three years. In this study, we tested the hypothesis that SSMs collected from connected vehicles during a much shorter time period can reduce the safety data collection time requirements substantially. The connected vehicle data has thus the potential to be used to advance real-time road safety management by developing more proactive safety solutions before crashes occur.

Regarding the limitations of this study, the extent of measurement errors in the vehicle trajectory data was unknown, although the findings in this study were reasonable. Crash data obtained from MDOT could be subject to underreporting issue. All three SSMs used in this study, namely TTC, DRAC, TTCD, aim to capture the rear-end crash risk. SSMs accounting for lateral movements will be studied in future as additional work. Additionally, the number of crashes and conflicts are normalized by AADT and the number of data points (exposure), respectively. However, it should be pointed out that the relationship between AADT and crashes may not be linear. Thus, a careful investigation of the relationship between AADT and crashes as well as the relationship between number of conflicts and exposure are proposed as an important future study. In such a future study, after accounting for the exposure effect properly, the correlations between each SSM and crashes need to be recalculated to further verify the improvement of TTCD compared with TTC and DRAC. Additional data is needed to further verify the type of correlation of each SSM with observed crashes.

Besides, another future research direction is to predict future high-risk locations dynamically. Additional CV data is needed to quantify the risk of the road network continuously. The potential of using CV data to evaluate and reduce the risk of secondary crashes can also be studied Yang et al. (2018). Time series model can be trained to associate the future risk with the past. Traffic simulation might be implemented to isolate confounding factors. Prediction of future high-risk locations can also guide the development of short-term patrolling plans of police vehicles or safety measures in terms of modifying existing signs, traffic lights, etc.

Acknowledgments

This study was partially supported by the Connected Cities for Smart Mobility towards Accessible and Resilient Transportation (C^{2}SMART) a Tier 1 USDOT University Transportation Center (http://c2smart.engineering.nyu.edu/) at New York University. The contents of this paper reflect views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the paper do not necessarily reflect the official views or policies of the funding agencies.

References

Car and Driver (2014). "Car and Driver Track Sheet." http://media.caranddriver.com/files/2015-porsche-918-spyder-feature-car-and-driver2015-porsche-918-spyder.pdf. (March 25, 2017).

Cooper, D.F., Ferguson, N., 1976. Traffic studies at T-junctions. 2. A conflict simulation record. Traffic Eng. Control 17 (Analytic).

Cunto, F.J.C., Saccomanno, F.F., 2007. Microlevel traffic simulation method for assessing crash potential at intersections. Proc., Transportation Research Board 86th Annual Meeting of the Transportation Research Board.

Davis, G.A., Hourdos, J., Xiong, H., Chatterjee, I., 2011. Outline for a causal model of traffic conflicts and crashes. Accid. Anal. Prev. 43 (6), 1907–1919.

ESRI, R, 2011. ArcGIS Desktop: Release 10. Environmental Systems Research Institute, CA.

Forbes, T.W., 1957. Analysis of" Near Accident" Reports. Highway Res. Board Bull. 152, 23–37.

Gao, J., Ozbay, K., Zuo, F., Kurkcu, A., 2018. A Life-Cycle Cost-Analysis Approach for Emerging Intelligent Transportation Systems with Connected and Autonomous Vehicles.

Guttinger, V., 1982. From accidents to conflicts: alternative safety measurement. Proc., Proceedings of the Third International Workshop on Traffic Conflict Techniques.

Hayward, J.C., 1972. Near miss determination through use of a scale of danger. Transp. Res. Rec.: J. Transp. Res. Board 384, 24–34.

Henclewood, D., Rajiwade, S.S., 2015. Safety Pilot Model Deployment –Sample Data Environment Data Handbook. Research and Technology Innovation Administration, US Department of Transportation, Mclean, VA.

Henclewood, D., Abramovich, M., Yelchuru, B., 2014. Safety pilot model deployment-one day sample data environment data handbook. USDOT Res. Technol. Innov. Adm. 1.

Ismail, K., Sayed, T., Saunier, N., Lim, C., 2009. Automated analysis of pedestrian-vehicle conflicts using video data. Transp. Res. Rec.: J. Transp. Res. Board (2140), 44–54.

Kahle, D., Wickham, H., 2013. ggmap: spatial visualization with ggplot2. R J. 5 (1).

Kamrani, M., Khattak, A.J., Wali, B., 2017. Can data generated by connected vehicles enhance safety? Proactive approach to intersection safety management. Proc., 96th Annual Meeting of the Transportation Research Board.

Kamrani, M., Arvin, R., Khattak, A.J., 2018. Extracting useful information from connected vehicle data: an empirical study of driving volatility measures and crash frequency at intersections. Proc., 97th Annual Meeting of theTransportation Research Board.

Kuang, Y., Qu, X., 2014. A review of crash surrogate events. Vulnerability Uncertainty Risk: Quantif. Mitig. Manage. 2254–2264.

Kuang, Y., Qu, X., Wang, S., 2015. A tree-structured crash surrogate measure for freeways. Accid. Anal. Prev. 77, 137–148.

Kurkcu, A., 2018. Connected Transportation System: Next Generation Traffic Simulation and Data Collection Tools and Techniques. New York University.

LatLong.net, 2017. Ann Arbor, Michigan, the United States. Retrieved from http://www.latlong.net/place/ann-arbor-michigan-the-united-states-610.html.

Laureshyn, A., Ardö, H., 2006. Automated video analysis as a tool for analysing road user behaviour. In: Proc., ITS World Congress. London, UK. pp. 8.

Laureshyn, A., Johnsson, C., De Ceunynck, T., de Svensson, Å., Goede, M., Saunier, N., Włodarek, P., van der Horst, R., and Daniels, S. (2016). Review of current study methods for VRU safety. Appendix 6–Scoping review: surrogate measures of safety in site-based road traffic observations: Deliverable 2.1–part 4.

Liu, J., Khattak, A.J., 2016. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. Transp. Res. Part C: Emerg. Technol. 68, 83–100.

Liu, J., Khattak, A.J., 2018. Mapping location-based driving volatility for connected and automated vehicles. Proc., 97th Annual Meeting of theTransportation Research Board.

Machiani, S.G., Jahangiri, A., Balali, V., Belt, C., 2017. Predicting driver risky behavior for curve speed warning systems using Real Field connected-vehicle data. Proc., 96th Annual Meeting of theTransportation Research Board.

Minderhoud, M.M., Bovy, P.H., 2001. Extended time-to-collision measures for road traffic safety assessment. Accid. Anal. Prev. 33 (1), 89–97.

Mousa, S.R., Ishak, S., 2017. An extreme gradient boosting algorithm for freeway short-term travel time prediction using basic safety messages of connected vehicles. Proc., 96th Annual Meeting of theTransportation Research Board.

Nodine, E., Stevens, S., Lam, A., Jackson, C., Najm, W.G., 2015. Independent Evaluation of Light-Vehicle Safety Applications Based on Vehicle-to-Vehicle Communications Used in the 2012-2013 Safety Pilot Model Deployment. National Highway Traffic Safety Administration, United States.

Ozbay, K., Yang, H., Bartin, B., Mudigonda, S., 2008. Derivation and validation of new simulation-based surrogate safety measure. Transp. Res. Rec.: J. Transp. Res. Board 105–113.

Pearson, K., 1895. Note on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. 58, 240–242.

Perkins, S.R., Harris, J.I., 1967. Criteria for Traffic Conflict Characteristics, Signalized Intersections. Research Laboratories, General Motors Corporation.

Production Car Speed Record, 2017. Production Car Speed Record. (March 27). https://en.wikipedia.org/w/index.php?title=Production_car_speed_record&oldid=772434722.

Souza, J.Q., Sasaki, M.W., Cunto, F.J.C., 2011. Comparing simulated Road safety performance to observed crash frequency at signalized intersections. Proc., Artigo Aceito Para Apresentação in 3rd International Conference on Road Safety and Simulation.

Team, R.C., 2013. R: A Language and Environment for Statistical Computing.

van der Horst, R., Hogema, J., 1993. Time-to-collision and collision avoidance systems. Proc., 6th ICTCT Workshop Salzburg.

Xie, K., Li, C., Ozbay, K., Dobler, G., Yang, H., Chiang, A.-T., Ghandehari, M., 2016. Development of a comprehensive framework for video-based safety assessment. In: Proc., Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE. pp. 2638–2643.

Xu, C., Wang, X., Chen, X., 2012. Urban Expressway Speed Spatial Inconsistency and Its Effect on Safety.

Yang, H., 2012. Simulation-Based Evaluation of Traffic Safety Performance Using Surrogate Safety Measures. Rutgers University-Graduate School-New Brunswick.

Yang, H., Ozbay, K., Bartin, B., 2010. Application of simulation-based traffic conflict analysis for highway safety evaluation. In: Proceedings of the 12th WCTR. Lisbon, Portugal.

Yang, H., Wang, Z., Xie, K., 2017. Impact of connected vehicles on mitigating secondary crash risk. Int. J. Transp. Sci. Technol. 6 (3), 196–207.

Yang, H., Wang, Z., Xie, K., Ozbay, K., Imprialou, M., 2018. Methodological evolution and frontiers of identifying, modeling and preventing secondary crashes on highways. Accid. Anal. Prev. 117, 40–54.

Zhang, M., Khattak, A.J., 2018. Identifying and analyzing extreme lane change events using basic safety messages in a connected vehicle environment. Proc., 97th Annual Meeting of the Transportation Research Board.

Zhao, S., Zhang, K., 2017. Observing space-time queueing dynamics at a signalized intersection using connected vehicles as Mobile sensors. Proc., 96th Annual Meeting of theTransportation Research Board.

Zheng, J., Liu, H.X., 2017. Estimating traffic volumes for signalized intersections using connected vehicle data. Transp. Res. Part C: Emerg. Technol. 79, 347–362.