TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Article

Transportation Research Record
2018, Vol. 2672(38) 1–11
© National Academy of Sciences:
Transportation Research Board 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0361198118790861
journals.sagepub.com/home/trr
SAGE

Assessing Surrogate Safety Measures using a Safety Pilot Model Deployment Dataset

Zhaoxiang He[1], Xiao Qin[1], Pan Liu[2], and Md Abu Sayed[1]

Abstract

Emerging data sources such as Safety Pilot Model Deployment (SPMD) provide a great opportunity to gain a better understanding of collision mechanisms and to develop novel safety metrics. The SPMD program was a comprehensive data collection effort under real-world conditions in Ann Arbor, Michigan, covering over 73 lane-miles and including approximately 3,000 pieces of onboard vehicle equipment and 30 pieces of roadside equipment. In-vehicle data (e.g., speed, location) collected by the SPMD program can potentially be an important supplement to traditional crash data-oriented safety analysis. The goal of this study was to assess roadway link-level surrogate safety measures using the vehicle trajectory data from SPMD. The study's objectives included: 1) developing a framework to process the SPMD dataset using big-data analytics; 2) converting raw vehicle motion data from SPMD to surrogate safety measures; and 3) analyzing the statistical relationship between crash records and the calculated safety index. The statistical models showed that modified time to collision (MTTC) outperforms time to collision (TTC) and deceleration rate to avoid collision (DRAC) with respect to its goodness of fit. The findings are promising in that augmenting safety analysis with surrogate measures and vehicle performance (e.g., speed and brake duration from connected vehicles) improves the overall model performance. Such information is vital for safety analysis, especially in the absence of detailed roadway and traffic data.

Traffic deaths in 2015 set the record for the largest annual increase in the United States since 1966 (*1*). The preliminary information from the National Highway Traffic Safety Administration (NHTSA) shows that traffic fatalities increased by 10.4% from 2015 to 2016 (*1*). On average, more than 6 million crashes are reported annually, resulting in more than 30 thousand fatalities and 2 million injuries every year on the U.S. highways and streets (*1*). Because of the tremendous loss caused by traffic accidents, there is a keen interest in developing better safety metrics and seeking solutions to reduce crashes.

Crash history has long been considered by researchers and safety professionals as a reliable performance measure for road safety, and possibly the most straightforward measure. However, the shortcoming of using crash history is that sites without crashes cannot be properly evaluated. Moreover, collecting crash data for a valid statistical evaluation usually takes a long time, and crash data often are biased because of issues like underreporting and reporting thresholds. The predictive methods in the Highway Safety Manual (HSM) provide scientific approaches to calculating the expected annual average crashes, given a series of contributing variables, but the requirement for detailed roadway and traffic data makes the application impractical in many local agencies.

The traffic conflict technique (TCT), which supports surrogate safety measures for road safety, provides an alternative to evaluate road safety. Traffic conflict is defined as "an observable situation in which two or more road users approach each other in time and space for such an extent that there is a risk of collision if their movements remain unchanged" (*2*). This type of measurement is usually derived from the vehicle kinematic characteristics before possible conflicts using the collision theory. TCT not only offers an objective view of collision likelihood based on a vehicle's motion, but also considers

[1] Department of Civil and Environmental Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI
[2] Jiangusu Key Laboratory of Urban ITS, Southeast University, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Nanjing, China

Corresponding Author:
Address correspondence to Xiao Qin: qinx@uwm.edu

driver behaviors such as speed, acceleration/deceleration, and following distance. Traditionally, traffic microsimulation models and field observations, either from observers or by fixed videos, can help to measure conflicts. Nowadays, emerging data from initiatives such as Safety Pilot Model Deployment (SPMD) present new opportunities to test, validate, and evaluate surrogate safety measures with high-resolution vehicle trajectory data.

The main use of the surrogate measures is to evaluate road safety in the absence of crash data or when crash data is limited. Surrogate measures can be used to measure safety performance of a highway facility or to estimate effectiveness of safety treatment. These measures can provide rapid evaluation for innovative intersection designs or new traffic control strategies that usually require a longer period of time to accumulate an adequate number of sites and crash history. They can measure the safety improvements for rare crash types such as pedestrian or bicycle crashes. Moreover, surrogate safety measures can be used to evaluate "what if" scenarios in microscopic traffic simulation models.

SPMD is a research initiative that demonstrates connected vehicle safety technologies in real-world implementation. Data are collected from vehicles equipped with vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication devices in Ann Arbor, Michigan (*3*). Approximately 3,000 vehicles are instrumented with data acquisition systems (DAS) and V2V communication devices (*3*). In-vehicle data such as speed, location, and direction are collected and communicated through basic safety messages (BSMs) by dedicated short range communications (DSRC) 10 times per second (*3*). The goal of this study is to retrieve proper information from vehicle trajectory data in SPMD, construct several surrogate safety measures, and assess their performance in a safety evaluation.

Literature Review

Surrogate safety measures originate from quantifying the kinematics of a conflict in traffic microsimulation models (*2*). Broadly, surrogate safety measures can be categorized as either temporal or nontemporal (*2*). Time to collision (TTC), "the time to collide if two vehicles continue at their present speed and along the same path," is a commonly used temporal surrogate measure that has been widely used to evaluate road safety in different traffic conditions (*4–7*). TTC is usually measured for each time stamp, and a threshold is used to determine whether a collision will happen if the current speed and direction are maintained (*5*). Dijkstra used TTC to measure traffic conflicts with various levels of risk, identifying the relationship between conflicts and observed crashes (*4*). Laureshyn used TTC to analyze conflicts between turning movements and through movements near the intersection (*7*). TTC threshold values have been proposed for different conditions (*8*, *9*), and a threshold of 4 s has been used to differentiate between safe and uncomfortable situations (*8*). Hydén and Linderholm use 1.5 s as the TTC threshold to detect a severe conflict (*9*). Note that the severity of a conflict or the value of a surrogate measure is to estimate the crash risk rather than the crash severity.

Time Exposed TTC (TET) and Time Integrated TTC (TIT) are two modified TTC indicators that can be used when TTC is below the threshold value (*10*). TET is the summation of all times that a driver approaches the front vehicle with a TTC below the threshold, and TIT is the integration of the TTC profile when TTC is below the threshold (*10*). TTC assumes a constant speed and ignores possible conflicts caused by the change of speed; therefore, modified time to collision (MTTC) is suggested because it considers the vehicle's acceleration or deceleration (*11*). Charly and Mathew used MTTC to identify mid-block conflicts under mixed traffic conditions, and evaluated the temporal and spatial correlation between conflicts and observed mid-block crashes (*12*). Other temporal indicators include post-encroachment time (PET), where PET is the time between the first vehicle leaving a common spatial area and the second arriving at the area (*13*).

Distance-related and deceleration-related are two types of nontemporal safety indicators. The proportion of stopping distance (PSD), "the ratio of the remaining distance to the point of collision to the minimum acceptable stopping distance," is commonly used with the assumption of the maximum available deceleration rate (MADR) (*14*). An unsafe situation is detected if PSD is less than 1, when a collision cannot be avoided even under MADR. The conventional deceleration-based index is deceleration rate to avoid collision (DRAC), which is the deceleration required to avoid a crash (*15*). The Crash Potential Index (CPI) is defined as the probability that a given vehicle's DRAC exceeds its MADR during a given time interval (*16*). Evaluation should be conducted for candidate surrogate measures, as there are no current benchmarks (*17*).

Surrogate safety measures can be calculated from traffic microsimulation models which assume that drivers do not practice "unsafe" behaviors; however, driver error contributes tremendously to observed conflicts (e.g., crash or near crash) (*17*). Field observations or video data can also provide non-simulation–based safety surrogate measures. The latest automated video-based techniques make video data processing more efficient (*18*, *19*). Saunier and Sayed proposed an automated traffic conflict detection method to assess road safety based on video data with a feature-based vehicle tracking algorithm for intersection (*18*). Zangenehpour developed a new video-based methodology to extract different road users in traffic videos automatically, and then evaluated the relationship between calculated PET and historical crash data (*19*). Vehicle trajectory information can be derived from video data, which help calculate surrogate safety measures for different situations (*20–22*). Astarita used observed vehicle tracking data from a fixed video to calculate TTC and DRAC in an attempt to measure the behavior of drivers approaching an intersection (*20*). Meng and Weng explored the risk of rear-end crashes in work zones by using the trajectory data to calculate DRAC (*21*). Oh and Kim also used the vehicle trajectory to calculate TTC for rear-end crashes (*22*).

Emerging data sources now offer new opportunities and insights for traffic safety evaluation. SHRP2 Naturalistic Driving Study (NDS) provides a detailed examination of the role of driver performance and behavior in safety. The study focuses on 1) analyzing the statistical relationship between surrogate safety measures of collisions (e.g., conflicts, critical incidents, near-collisions) and actual collisions, and 2) using surrogate measures to formulate exposure-based risk measures (*23–26*). By using the in-vehicle NDS data supplemented by driver characteristics (e.g., gender, age, driver experience, speed selection), researchers captured surrogate measures to understand driver behavior during traffic conflicts (*25*). Montgomery et al. found a statistically significant difference in TTC at braking for different gender and age groups by using the 100-car NDS dataset (*27*). Although these studies show promising results, more work is required to refine methodologies and validate surrogate measures (*24*, *25*, *27*).

Another emerging data source is SPMD, which is collected from vehicles equipped with vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication devices under the real-world conditions in Ann Arbor, Michigan (*3*). The deployment included approximately 3,000 pieces of onboard vehicle equipment and 30 pieces of roadside equipment in an area of over 73 lane-miles. SPMD data includes driving data (DAS data), BSM data, roadside equipment data, weather data, and network traffic volume data (*3*). Driving data provide vehicles' kinematic and geographic information, and BSM data contain vehicles' position and motion with the status of components such as brakes, turning lights, wipers, and so on (*3*). Compared with video image based data in the NDS, the data such as DAS and BSM from SPMD are relatively easy to process. Moreover, SPMD data can provide real-time safety evaluation in a connected vehicle environment.

SPMD data have been applied in studies on transportation planning, traffic volume/congestion estimation, and extreme events identification (*28–32*). Deering processed spatial aggregation of trips into origin and destination zones for transportation planning by organizing SPMD data (basic safety message and driving data) into a trip-level dataset (*28*). Vasudevan et al. predicted congestion states from BSMs by using big-data graph analytics (*31*), and Zheng et al. used SPMD data to estimate traffic volumes for signalized intersections (*32*). The authors in Zheng's study developed an approach to estimate traffic volume using GPS trajectory data from connected vehicle (CV) devices under low market penetration rates (*32*). Liu et al. used data analytics to extract critical information (e.g., extreme event) embedded in BSMs, providing drivers instantaneous feedback about dangers in surrounding roadway environments (*30*). Extreme events were identified if instantaneous acceleration (the sum of motion vectors of longitudinal and lateral accelerations) exceeded the 95th percentile thresholds which change with speed (*30*). The identified extreme events were then connected to the vehicle maneuvering status (e.g., brake, turn signal) and driving context (e.g., number of objects, distance to the closest objects) to explore their relationship (*30*). Another study predicted the risky behavior of drivers (speed) at the point of curvature (PC) for different monitoring periods using the BSMs (*29*).

The challenge of using SPMD data is its size and complexity. The raw data tables are very large; even the 2-month BSM files are more than 400 GB. Deering found that during data processing, the implementation framework affected computation (*28*). A distributed computing framework like Hadoop had significantly reduced computation time when processing the CV data (*28*). Other challenges include understanding the data format and dealing with null values and outliers (*28*).

Data Collection

Because of the complexity of crash events and associated surrogate measures, only rear-end crashes in the mid-block were considered for this study. Road network data and crash data for Washtenaw County were collected from the Southeast Michigan Council of Governments Open Data Portal (http://maps-semcog.opendata.arcgis.com/). There were 52,386 crashes in the County from 2011 to 2015, including 17,103 rear-end crashes. According to Michigan's definition of intersection-related crashes, the 75-ft radius was used to remove rear-end crashes at the intersection (*33*), and the rest were kept as mid-block crashes. Figure 1 shows the road network and the crash points in Washtenaw County.

Complete speed limit information is not available for the road network, and an estimation was therefore made based on the road's functional class. A primary road has a value of 55 mph, a secondary road has a value of

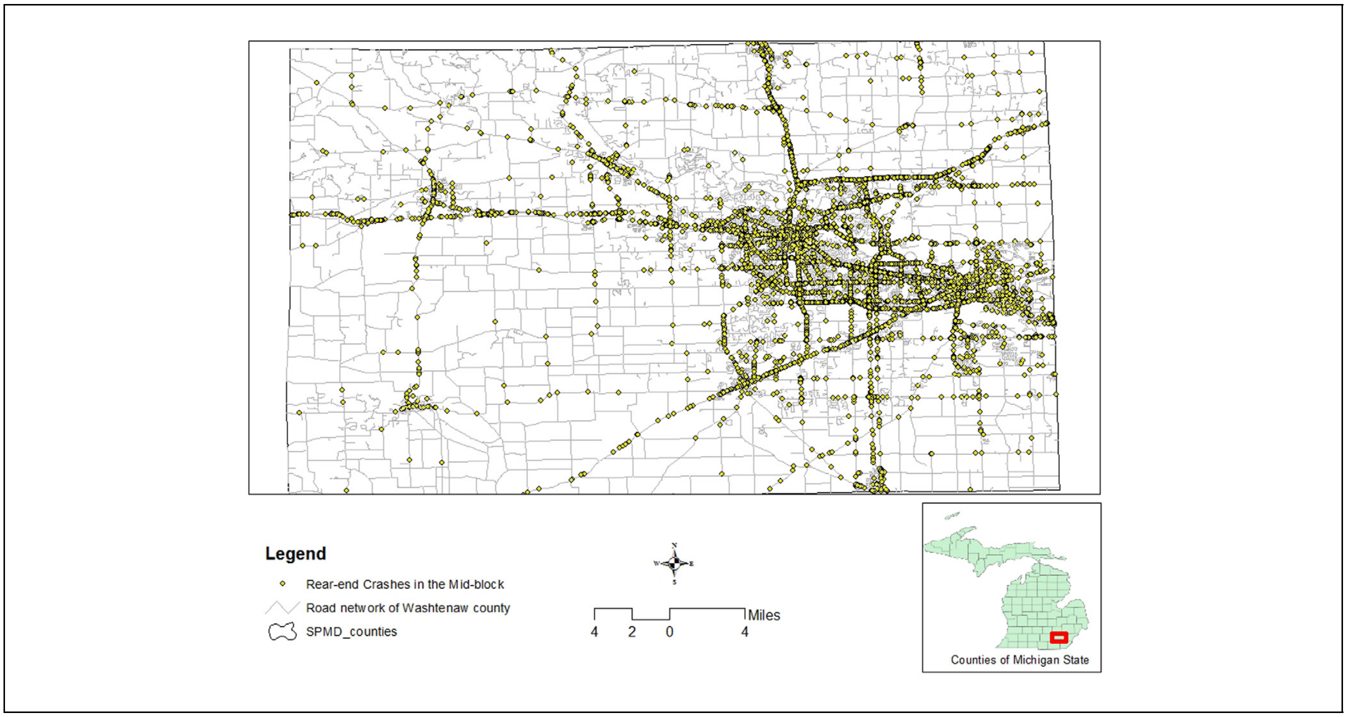
Legend
Rear-end Crashes in the Mid-block
Road network of Washtenaw county
SPMD_counties
N
W
E
S
Miles
4 2 0 4
Counties of Michigan State

Figure 1. Road network and crash points.

35 mph, and a local road or city road has a value of 25 mph. Road functional class information can be retrieved from "TIGER/Line Shapefiles" (https://www.census.gov/geo/maps-data/data/tiger-line.html). Annual average daily traffic (AADT) can be obtained from the same Governments Open Data Portal (http://maps-semcog.opendata.arcgis.com/), but only state highways are included.

Around 2,800 vehicles participated in the SPMD program, including 2,450 equipped with a vehicle awareness device (VAD), 300 with aftermarket safety devices (ASD), 19 with retrofit safety devices (RSD), and 67 with integrated safety devices (ISD). Among the vehicles with VAD, 2,350 are cars, 60 are trucks and 85 are transit buses. Vehicles with VAD can only transmit BSM. ASD can transmit and receive BSM but is only installed on cars. Similar to ASD but designed for freight and transit, RSD is installed on 16 trucks and three transit buses. Compared with ASD or RSD, ISD can not only send and receive BSM but also connect to the vehicle data processing system. Generally, drivers who reported the most driving within the SPMD area were selected to the SPMD program. For the 64 cars with ISD, participants were selected based on age and gender to ensure a similar number of drivers in each group.

Two months of the SPMD dataset (October 2012 and April 2013) are free to download in the transportation data sharing platform, or Research Data Exchange (RDE) (https://www.its-rde.net/index.php/rdedataenvironment/10018#). This study used the DataWsu file (12 GB) and the DataFrontTargets file (4.34 GB) in the driving dataset DAS1 collected from around 100 equipped vehicles. DataWsu, from the wireless safety unit (WSU), includes mainly GPS-based data elements (e.g., speed, longitudinal acceleration) and the state of some components such as brake status and headlamp status. DataFrontTargets, which was populated mainly with the aid of Mobileye's vision-based Advanced Driver Assistance Systems (http://www.mobileye.com/), collects front target information such as distance to the front target and relative speed for the front target. In this study, common data fields "Device," "Trip," and "Time" were used to link the two datasets. Table 1 contains the primary data elements in the DataWsu file and the DataFrontTargets file, along with a brief description of each.

Methodology

TTC, MTTC, and DRAC were chosen as the three safety surrogate measures because of their popularity. Methods for calculating link-based surrogate safety measures were developed using information extracted from the real-world SPMD data set. Vehicle kinematics from the SPMD data set were used to calculate vehicle-level safety surrogate measures at each timestamp, which were aggregated into trip-level safety surrogate measures for each link. The link-level indexes combine trip-level indexes for

Table 1. Major Data Elements in the DataWsu File and DataFrontTargets File

DataWsu			
Field name	Type	Units	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 ignition cycle commences when the ignition is in the on position and ends when it is in the off position
Time	Integer	centiseconds	Time in centiseconds since DAS started, which (generally) starts when the ignition is in the on position
GpsValidWsu	Integer	none	Communicates whether a GPS data point is valid or not
GpsTimeWsu	Integer	ms	Epoch GPS time received from the remote vehicle that has been targeted by the host vehicle's WSU
LatitudeWsu	Float	degrees	Latitude from WSU receiver
LongitudeWsu	Float	degrees	Longitude from WSU receiver
AltitudeWsu	Real	m	Altitude from WSU receiver
GpsHeadingWsu	Real	degrees	Heading from WSU GPS receiver
GpsSpeedWsu	Real	m/s	Speed from WSU GPS receiver
SpeedWsu	Real	km/h	Speed from vehicle CAN Bus via WSU
TurnSngRWsu	Integer	none	Right turn signal from vehicle CAN Bus via WSU
TurnSngLWsu	Integer	none	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/s^2	Longitudinal acceleration from vehicle CAN Bus via WSU
PrndlWsu	Integer	none	Current transmission state (Park, Reverse, Neutral, Drive, Low) 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	degrees	Steering angle/position from vehicle CAN Bus via WSU

DataFrontTargets			
Field name	Type	Units	Description
TargetId	Integer	none	Numeric ID assigned by the Mobileye sensor to distinguish between the different objects being tracked; the closest obstacle is given a TargetId value of 1
ObstacleId	Integer	none	ID of new obstacle, as assigned by the Mobileye sensor, and its value will be the last used free ID
Range	Integer	m	Longitudinal position of an object, typically the closest object, relative to a reference point on the host vehicle, according to the Mobileye sensor
RangeRate	Real	m/s	Longitudinal velocity of an object, typically the closest object, relative to the host vehicle, according to the Mobileye sensor
Transversal	Real	m	The lateral position of the obstacle, as determined by the Mobileye sensor
TargetType	Integer	none	Classification of an identified obstacle/target as a car, truck, pedestrian, and so on
Status	Integer	none	Classification of the motion (kinematic state) of an identified obstacle/target as stopped, moving, and so on
CIPV	Integer	none	Field communicating whether an obstacle is the closest in a vehicle's path

each trip on the link. The statistical relationships between surrogate safety measures and rear-end crashes were developed. Additional safety information such as vehicle maneuvering decisions (e.g., excess speed and brake time duration) and traffic volume were also included in the regression models.

Vehicle-Level Safety Surrogate Measures

Safety surrogate measures are usually measured at each timestamp between vehicles that are interacting with each other. In Figure 2, it is assumed that if the distance to the front vehicle is larger than 250 m, or the lateral difference is greater than 3 m, there will be no conflict at this time-spatial point (*14*). Let V_F be the front vehicle speed, and V_S be the subject vehicle speed; the relative speed equals ΔV and is $(V_S - V_F)$. Let a_F be the front vehicle acceleration and a_S be the subject vehicle acceleration; the relative acceleration Δa equals $(a_S - a_F)$. D represents the distance between the two vehicles. The SPMD dataset provides all of this information except for a_F, which can be calculated from the front vehicle speed information between two consecutive time stamps. Equations 1–3 detail surrogate safety measures TTC, MTTC, and DRAC, and the assumptions.

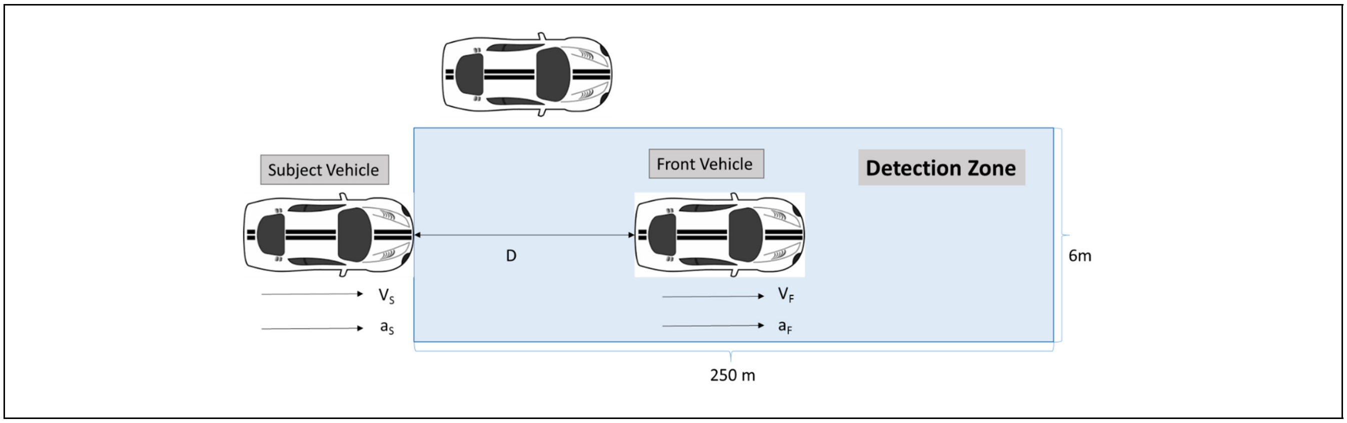

Figure 2. Illustration of vehicle motion in SPMD dataset.

- TTC is defined as "the time that remains until a collision between two vehicles would have occurred if the collision course and speed difference are maintained" (*2*):

$$\text{TTC} = \frac{D}{\Delta V} \tag{1}$$

When $\Delta V \leq 0$, there is no risk of collision at that moment. Let L be the link length. The upper limit of TTC is set to be $\frac{L}{V_s}$, ensuring that if there is a conflict, it is within the roadway segment link.

- MTTC considers the trajectory parameters of the two consecutive vehicles, including their relative distance, speed, and acceleration (*11*). The threshold of MTTC is also assumed to be $\frac{L}{V_s}$:

$$\text{MTTC} = \begin{cases} \max(t_1, t_2), & \textit{if } \Delta a > 0 \\ \min(t_1, t_2), & \textit{if } \Delta a < 0 \textit{ and } \Delta V > 0 \\ t_3, & \textit{if } \Delta a = 0 \textit{ and } \Delta V > 0 \\ \text{na}, & \textit{if } \Delta a \leq 0 \textit{ and } \Delta V \leq 0 \end{cases} \tag{2}$$

where
$$t_1 = \frac{-\Delta V + \sqrt{\Delta V^2 + 2\Delta a D}}{\Delta a}$$
$$t_2 = \frac{-\Delta V - \sqrt{\Delta V^2 + 2\Delta a D}}{\Delta a}$$
$$t_3 = \frac{D}{\Delta V}$$

- DRAC is "the minimum deceleration rate required by the following vehicle to avoid a crash with the leading vehicle if the speed of leading vehicle is unchanged during the process" (*15*):

$$\text{DRAC} = \frac{{V_S}^2 - {V_F}^2}{2(D - V_S{*}\text{PRT})} \tag{3}$$

where PRT is the perception–reaction time. The default PRT value used is 0.92 s, which is adopted from Triggs and Harris's study for rear-end collision (*34*). When $V_S - V_F \leq 0$, there is no risk of collision at that moment. When $D \leq V_S{*}\text{PRT}$, the index can be treated as positive infinity, meaning the collision is certain to happen.

Trip-Level Safety Surrogate Measures

A trip is defined as the time a vehicle traverses a roadway segment link. After the vehicle-level safety surrogate measures are available, they will be aggregated into safety surrogate measures for each vehicle trip. During a nonstop long trip, a vehicle may traverse the same link multiple times. Thus vehicle-level safety surrogate measures with the same "Trip" field value can be aggregated into multiple trip-level safety indexes. For the same vehicle, if there are multiple front targets at the same timestamp, the closest one ("CIPV = 1") is kept.

It is assumed that the safety surrogate measure for each time interval equals the value for the end timestamp of that interval. For trip i in a link, the four safety indexes (SI) are formulated in Equations 4–7 to compute the trip-level safety surrogate measures on a link. For the sake of generality, TTC, MTTC, DRAC, and their aggregations are called SI.

- Time Duration (SI_{1i}) is the total time when the surrogate measures exist. Note that the time duration should be less than or equal to the link travel time $\frac{L}{V_s}$:

$$\text{SI}_{1i} = \sum_P (t_{i,j} - t_{i,j-1}) \tag{4}$$

- Average Index (SI_{2i}) is the weighted average of surrogate measures over time:

$$SI_{2i} = \frac{\sum_{P}(t_{i,j} - t_{i,j-1})*Index_{i,j}}{\sum_{P}(t_{i,j} - t_{i,j-1})} \quad (5)$$

- Median Index (SI_{3i}) is the median value of surrogate measures over time:

$$SI_{3i} = \underset{P}{median}\{Index_{i,j}\} \quad (6)$$

- Extreme Index (SI_{4i}) is either the minimum TTC or MTTC, or the maximum DRAC, representing the most dangerous situation:

$$SI_{4i} = \min_{P}\{Index_{i,j}\} or \max_{P}\{Index_{i,j}\} \quad (7)$$

where P = {time intervals when safety surrogate measures are available}, and j is the index of time stamp.

Link-Level Safety Surrogate Measures

One link could include multiple trips: different vehicles can travel on the same link or one vehicle can travel the same link multiple times. Trip-level safety surrogate measures need to be aggregated to a link-level safety surrogate measure. The proposed link-level safety surrogate measures for each link are presented in Equations 8–11.

- Time Duration (SI_1) is the average length of time of all trips in a link:

$$SI_1 = \frac{\sum_1^N SI_{1i}}{N} \quad (8)$$

- Average Index (SI_2) is the average index of all trips in a link:

$$SI_2 = \frac{\sum_1^N (SI_{2i}*SI_{1i})}{\sum_1^N SI_{1i}} \quad (9)$$

- Median Index (SI_3) is the median index of all trips in a link:

$$SI_3 = \underset{1 \le i \le N}{median}\{SI_{3i}\} \quad (10)$$

- Extreme Index (SI_4) is the minimum or maximum index of all trips in a link:

$$SI_4 = \min_{1 \le i \le N}\{SI_{4i}\} or \max_{1 \le i \le N}\{SI_{4i}\} \quad (11)$$

Data Processing

The DataWsu data and DataFrontTargets data were used in this study to calculate SI. Figure 3 shows the framework of data processing and safety index calculation in this study, along with the size and volume of the database after each process.

Initially, each dataset was examined using Python programming language to check the data type and data organization. The datasets were then imported into Hadoop to conduct a query using Apache Hive, a query language that is built on top of Apache Hadoop. Unfortunately, because of the complicated relationship among columns, the aggregation of two big datasets in Hadoop was extremely slow. Moreover, current Hadoop-GIS does not support advanced spatial analysis such as spatial join, so the two datasets were exported into small files to be joined in the PostgreSQL database. The combined data were then imported into ArcGIS to integrate link and intersection information. The combined data were imported back into the PostgreSQL database, and 75-ft buffer zones were created to remove points around the intersections. Finally, a combined dataset was generated for the target type of "car" or "truck."

Vehicle-level SI were calculated within the "range" based on the combined dataset with driving information and front vehicle information (i.e., distance to the front vehicle is less than 250 m and the lateral distance is less than 3 m). Then, safety surrogate measures were calculated for each link. Rear-end crashes in the mid-block and the links with safety surrogate measures, are shown in Figure 4.

Results

The statistical relationship between the link safety surrogate measures and mid-block rear-end crashes was developed using the negative binomial (NB) model, of which the mean is estimated as a log-linear function of the explanatory variables. Only links with observed surrogate measures were selected. Of the 2,772 selected links, the average link length is 352 m, the median length is 213 m, the minimum length is 35 m, and the maximum length is 3,635 m. The surrogate safety measures and other independent variables include vehicle maneuvering actions such as average speed and brake duration and segment link length. The link length is treated as an independent variable rather than as an exposure variable to a crash because the length is used to set the upper limit for a safety index. Variance inflation factors (VIFs) were calculated for each independent variable to examine the possibility of multicollinearity. In this study, all VIFs are less than 10 except for the average index and median index. Small VIF values mean no high correlations exist among the independent variables after excluding average index and median index. Traffic volume such as AADT is typically included as the traffic exposure to crashes. However, during the variable selection process, AADT

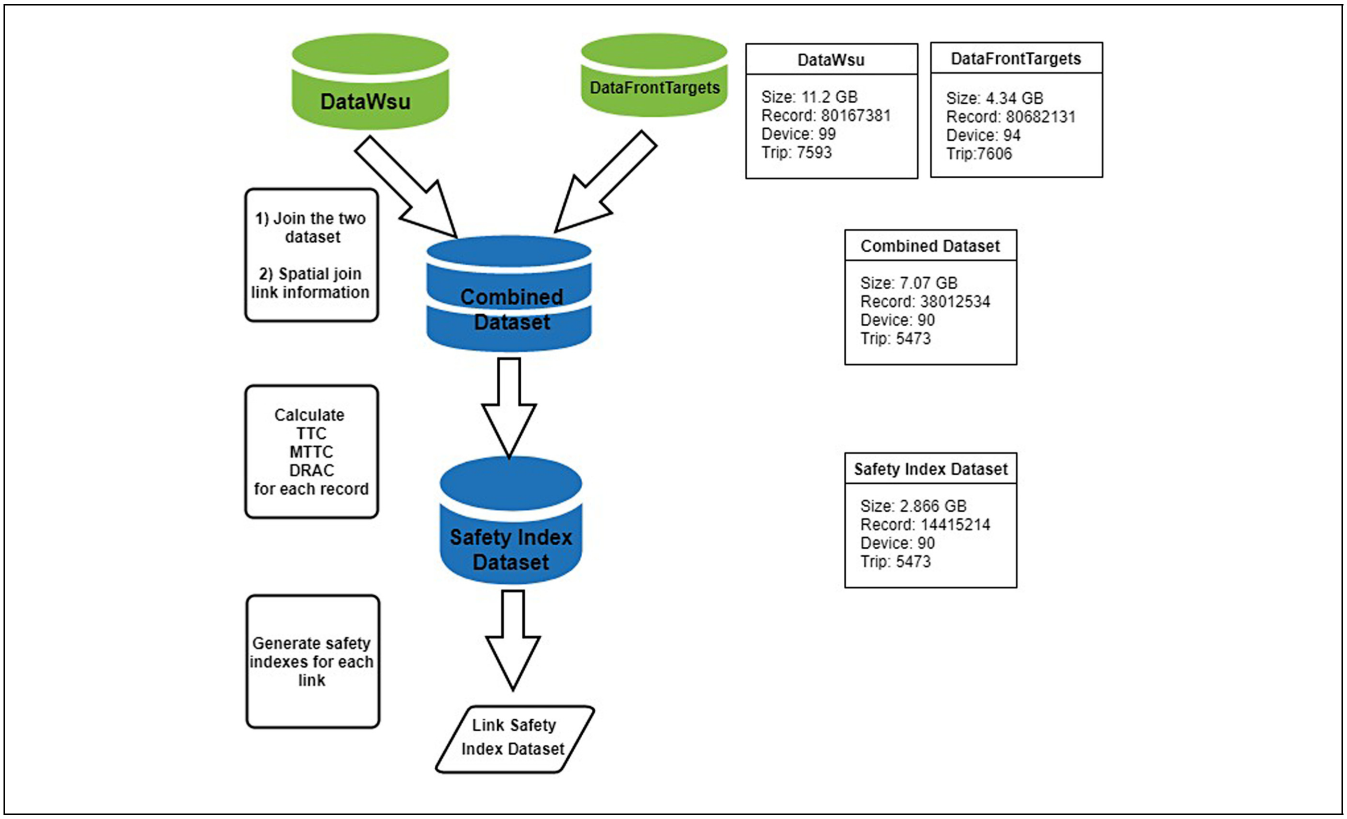

Figure 3. Data processing framework.

was not statistically significant for the selected links and, therefore, AADT was excluded from the final NB models. The results of the models are shown in Table 2.

For TTC, the variables of time duration, extreme index, average speed, brake duration, and link length are statistically significant at the 1% level or lower. Time duration and extreme index are two SI that affect crash frequency at a statistically significant level. The positive sign of time duration means that the longer the dangerous situation lasts, the more frequently a crash will occur. The negative sign of extreme index suggests a larger minimum TTC is associated with a smaller number of crashes. The positive impact of average speed on crashes indicates that a faster speed results in more crashes. The longer brake time within the link means a worse link safety condition (higher crash frequency). The positive effect of link length on crashes simply suggests that more crashes may happen if the link is longer.

When compared with TTC, the differences in parameter estimates for MTTC include the lower time duration value and larger extreme index value. It is found that the mean value of time duration for MTTC is larger than for TTC (1.22 s versus 0.56 s) and the mean value of extreme index for MTTC is lower than for TTC (1.86 s versus 3.43 s), suggesting an MTTC-based conflict is more easily detected compared with a TTC-based conflict. As an acceleration-based surrogate measure, all SI are statistically significant for DRAC. The positive sign of extreme index for DRAC means the increase of the maximum DRAC leads to more crashes. Average speed and brake duration are always statistically significant with the similar parameter estimates, irrespective of the SI. Excess speed, reflecting the driver's aggressiveness in choosing speed, was calculated and included in the model. The excess speed equals zero when the average speed is less than the speed limit and equals the difference between the two if the average speed is greater than the speed limit. Model results show insignificant SI when replacing average speed with excess speed.

The dispersion parameter was estimated in the NB model to measure the data overdispersion. The respective values of 0.399, 0.419, and 0.372 in the models for TTC, MTTC, and DRAC justify the choice of the NB model. Goodness-of-fit measures such as Akaike Information Criterion (AIC) and pseudo *R*-squared were used to compare the model performance. AIC uses the maximum log-likelihood function with a penalizing term related to the number of variables. A lower AIC value indicates a better fit. McFadden pseudo *R*-squared, analogous to the *R*-squared value for linear regression models, equals

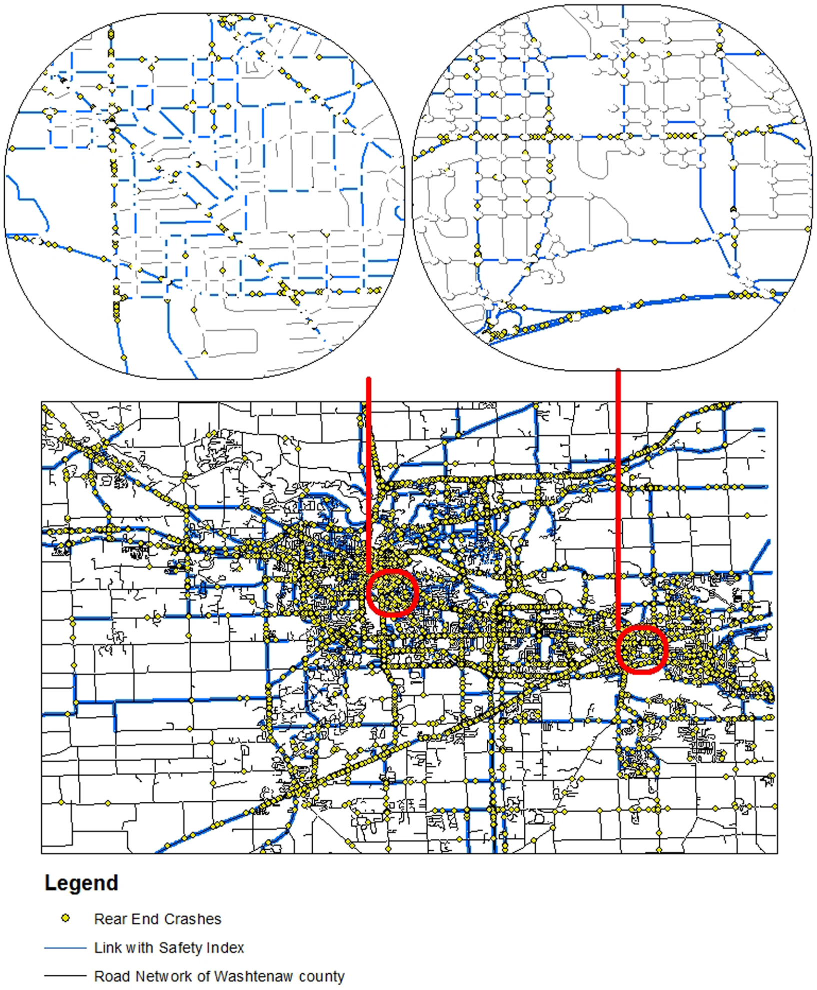

Figure 4. Maps with observed safety index and crash points.

Table 2. NB Models for Mid-Block Rear-End Crashes

	TTC		MTTC		DRAC	
	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value
(Intercept)	–0.990***	$<2e^{-16}$	–0.624***	$<2e^{-16}$	–1.237***	$<2e^{-16}$
Time duration	0.222**	0.002	0.147***	$<2e^{-16}$	0.063***	$<2e^{-16}$
Extreme index	–0.100***	$<2e^{-16}$	–0.518***	$<2e^{-16}$	4×10^{-6}**	0.007
Average speed	0.082***	$<2e^{-16}$	0.083***	$<2e^{-16}$	0.077***	$<2e^{-16}$
Brake duration	1.136***	$<2e^{-16}$	1.040***	$<2e^{-16}$	1.390***	$<2e^{-16}$
Link length	0.001***	$<2e^{-16}$	0.001***	$<2e^{-16}$	0.001***	$<2e^{-16}$
Number of observations	2,772		2,772		2,772	
Dispersion parameter	0.399		0.419		0.372	
2 log-likelihood	–9615.862		–9542.735		–9701.151	
AIC	9629.900		9556.700		9715.200	
McFadden pseudo *R*-squared	0.066		0.075		0.054	

Note: AIC = Akaike Information Criterion.
****p* = .001. ***p* = .01. **p* = .05.

1 minus the ratio of the log-likelihood of the full model to the log-likelihood of the intercept-only model. The pseudo R-squared takes a value between 0 and 1, and a higher value indicates better model performance. The relatively low R-squared values mean that if crash data are considered as ground truth, it cannot be concluded with confidence which surrogate measure is the best safety metric. Among the three indexes, MTTC is considered as the most statistically significant surrogate safety measure for crashes because it has the highest pseudo R-squared value and the lowest AIC value. In summary, the proposed safety information can be useful for evaluating segment link safety when roadway geometric and traffic information is limited.

Conclusion

An SPMD dataset was used to evaluate rear-end crashes in the mid-block. Three main objectives were accomplished: 1) develop a framework to process the SPMD big data; 2) construct surrogate safety measures from SPMD data; and 3) analyze the statistical relationship between crash records and a calculated safety index.

Unlike other studies that adopted surrogate safety measures to identify traffic conflicts, this study attempted to evaluate segment safety using surrogate safety measures. Aggregated surrogate measures were developed for a trip-level and a link-level safety index, as surrogate measures are usually taken at the vehicle level by measuring the time and space between a pair of vehicles. Surrogate measures are treated as SI, which quantify the severity of a potential traffic conflict. For example, a higher value of TTC means a higher safety index. In the real world, however, a TTC of 3 s may result in a crash whereas a TTC of 1 s may not; this is dependent on the driver. Thus, arbitrary threshold values were not used in the study to preserve the integrity of the information. The logic of this study is to keep the calculated surrogate safety measures for all the time points and then provide indicators (e.g., time duration, average index, median, or minimum/maximum index) to summarize these SI on the same trip for each link.

The NB models for mid-block rear-end crashes show the expected impact of explanatory variables on crashes. Among the three models, MTTC has a better goodness of fit when compared with TTC and DRAC. The findings show that augmenting safety analysis with surrogate measures and vehicle performance (i.e., speed and brake duration from CVs) improves the overall model performance. Such information can be vital when detailed roadway and traffic data are absent.

Some abnormal numbers were produced and removed, and the measurement errors in the dataset were unknown. The study is also less comprehensive because there is no record of the dataset in some columns (e.g. right turn or left turn signal). The complexity of the dataset means that some of the assumptions or data processing approaches used in this study may not be optimal in all situations; thus, future studies should search for other effective approaches. Future studies can be expanded into the comparison of other safety surrogate measures such as PET, Delta-V, and extended Delta-V. New and emerging safety surrogate measures can be developed with the rich information provided through CV safety technologies. In addition, future studies can also take advantage of the high-resolution vehicle kinematics and Signal Phasing and Timing (SPaT) data in the SPMD program to study risky driver behaviors such as red light running.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Qin, He; data collection: He, Sayed; analysis and interpretation of results: He, Qin, Sayed; draft manuscript preparation: He, Qin, Liu, Sayed. All authors reviewed the results and approved the final version of the manuscript.

References

1. NHTSA. *Crash Stats: Early Estimate of Motor Vehicle Traffic Fatalities for the First Half, January to June of 2016*. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812332.
2. Gettman, D., and L. Head. Surrogate Safety Measures from Traffic Simulation Models. *Transportation Research Record: Journal of the Transportation Research Board*, 2003. 1840: 104–115.
3. Henclewood, D., M. Abramovich, and B. Yelchuru. *Safety Pilot Model Deployment—One Day Sample Data Environment Data Handbook, V. 1.2*. USDOT Research and Technology Innovation Administrations, Washington, D.C., 2014.
4. Dijkstra, A., P. Marchesini, F. Bijleveld, V. Kars, H. Drolenga, and M. van Maarseveen. Do Calculated Conflicts in Microsimulation Model Predict Number of Crashes? *Transportation Research Record: Journal of the Transportation Research Board*, 2010. 2147: 105–112.
5. Hayward, J. C. Near Miss Determination through Use of a Scale of Danger. *Highway Research Record*, Vol. 384, 1972, pp. 24–34.
6. Kuang, Y., and X. Qu. A Review of Crash Surrogate Events. In M. Beer, S.-K. Au, & J. W. Hall (eds.), *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*. American Society of Civil Engineers, Reston, VA, 2014. pp. 2254–2264.
7. Laureshyn, A., Å. Svensson, and C. Hydén. Evaluation of Traffic Safety, Based on Micro-Level Behavioural Data: Theoretical Framework and First Implementation. *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1637–1646.

8. Horst, R. Time-to-Collision as a Cue for Decision-Making in Braking. *Vision in Vehicles*, Vol. 3, 1991, pp. 10–26.
9. Hydén, C., and L. Linderholm. The Swedish Traffic-Conflicts Technique. In *International Calibration Study of Traffic Conflict Techniques*, Springer, Berlin Heidelberg, 1984, pp. 133–139.
10. Minderhoud, M. M., and P. H. Bovy. Extended Time-to-Collision Measures for Road Traffic Safety Assessment. *Accident Analysis & Prevention*, Vol. 33, No. 1, 2001, pp. 89–97.
11. Ozbay, K., H. Yang, B. Bartin, and S. Mudigonda. Derivation and Validation of New Simulation-Based Surrogate Safety Measure. *Transportation Research Record: Journal of the Transportation Research Board*, 2008. 2083: 105–113.
12. Charly, A., and T. V. Mathew. Estimation of Modified Time To Collision as Surrogate For Mid-Block Crashes under Mixed Traffic Conditions. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.
13. Gettman, D., L. Pu, T. Sayed, and S. G. Shelby. *Surrogate Safety Assessment Model and Validation: Final Report*. Federal Highway Administration, Washington, D.C., 2008.
14. Allen, B. L., B. T. Shin, and P. J. Cooper. Analysis of Traffic Conflicts and Collisions. *Transportation Research Record: Journal of the Transportation Research Board*, 1978. 667: 67–74.
15. Almqvist, S., C. Hydén, and R. Risser. Use of Speed Limiters in Cars for Increased Safety and a Better Environment. *Transportation Research Record: Journal of the Transportation Research Board*, 1991. 1318: 34–39.
16. Cunto, F. J. C., and F. F. Saccomanno. Microlevel Traffic Simulation Method for Assessing Crash Potential at Intersections. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
17. Young, W., A. Sobhani, M. G. Lenné, and M. Sarvi. Simulation of Safety: A Review of the State of the Art in Road Safety Simulation Modelling. *Accident Analysis & Prevention*, Vol. 66, 2014, pp. 89–103.
18. Saunier, N., and T. Sayed. Automated Analysis of Road Safety with Video Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2007. 2019: 57–64.
19. Zangenehpour, S. *A Video-Based Methodology for Extracting Microscopic Data and Evaluating Safety Countermeasures at Intersections Using Surrogate Safety Indicators*. Doctoral dissertation. McGill University, Montreal, Québec, Canada, 2016.
20. Astarita, V., G. Guido, A. Vitale, and V. Giofré. A New Microsimulation Model for the Evaluation of Traffic Safety Performances. *European Transport\Trasporti Europei*, No. 51, 2012, pp. 1–16.
21. Meng, Q., and J. Weng. Evaluation of Rear-End Crash Risk at Work Zone Using Work Zone Traffic Data. *Accident Analysis & Prevention*, Vol. 43, No. 4, 2011, pp. 1291–1300.
22. Oh, C., and T. Kim. Estimation of Rear-End Crash Potential Using Vehicle Trajectory Data. *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1888–1893.
23. Campbell, K. L. The SHRP 2 Naturalistic Driving Study: Addressing Driver Performance and Behavior in Traffic Safety. *TR News*, No. 282, 2012, pp. 30–35.
24. Tarko, A. P. Use of Crash Surrogates and Exceedance Statistics to Estimate Road Safety. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 230–240.
25. Wu, K. F., and P. P. Jovanis. Crashes and Crash-Surrogate Events: Exploratory Modeling with Naturalistic Driving Data. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 507–516.
26. Markkula, G., J. Engström, J. Lodin, J. Bärgman, and T. Victor. A Farewell to Brake Reaction Times? Kinematics-Dependent Brake Response in Naturalistic Rear-End Emergencies. *Accident Analysis & Prevention*, Vol. 95, 2016, pp. 209–226.
27. Montgomery, J., K. D. Kusano, and H. C. Gabler. Age and Gender Differences in Time to Collision at Braking from the 100-Car Naturalistic Driving Study. *Traffic Injury Prevention*, Vol. 15, No. Suppl. 1, 2014, pp. S15–S20.
28. Deering, A. M. *A Framework for Processing Connected Vehicle Data in Transportation Planning Applications*. Doctoral dissertation. The University of Texas at Austin, Austin, TX, 2017.
29. Ghanipoor Machiani, S., A. Jahangiri, V. Balali, and C. Belt. Predicting Driver Risky Behavior for Curve Speed Warning Systems Using Real Field Connected Vehicle Data. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.
30. 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*, Vol. 68, 2016, pp. 83–100.
31. Vasudevan, M., D. Negron, M. Feltz, J. Mallette, and K. Wunderlich. Predicting Congestion States from Basic Safety Messages by Using Big-Data Graph Analytics. *Transportation Research Record: Journal of the Transportation Research Board*, 2015. 2500: 59–66.
32. Zheng, J., and H. X. Liu. Estimating Traffic Volumes for Signalized Intersections Using Connected Vehicle Data. *Transportation Research Part C: Emerging Technologies*, Vol. 79, 2017, pp. 347–362.
33. Abdel-Aty, M., and X. Wang. Identifying Intersection-Related Traffic Crashes for Accurate Safety Representation. *Institute of Transportation Engineers. ITE Journal*, Vol. 79, No. 12, 2009, p. 38.
34. Triggs, T. J., and W. G. Harris. *Reaction Time of Drivers to Road Stimuli*. Monash University Accident Research Centre, Victoria, Australia, 1982.

The Safety Section (ANB00) peer-reviewed this paper (18-04931).