Accident Analysis and Prevention 130 (2019) 75–83

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Assessing crash risk considering vehicle interactions with trucks using point detector data

Kyung (Kate) Hyun[a,*], Kyungsoo Jeong[b], Andre Tok[c], Stephen G. Ritchie[d]

[a] Department of Civil Engineering, University of Texas at Arlington, 416 Yates St., 425 Nedderman Hall, Arlington, TX, 76019, United States
[b] Department of Civil and Environmental Engineering, Intelligent Transportation Systems Lab., 77 Massachusetts Avenue, Building 1-180, Massachusetts Institute of Technology, United States
[c] Institute of Transportation Studies, 4000 Anteater Instruction and Research Building (AIRB), University of California, Irvine, Irvine, CA, 92697, United States
[d] Department of Civil and Environmental Engineering, 4014 Anteater Instruction and Research Building (AIRB), Institute of Transportation Studies, University of California, Irvine, Irvine, CA, 92697, United States

ARTICLE INFO

Keywords:
Truck
Crash risk
Vehicle interactions
Inductive loop detector
Conditional logistic regression

ABSTRACT

Trucks have distinct driving characteristics in general traffic streams such as lower speeds and limitations in acceleration and deceleration. As a consequence, vehicles keep longer headways or frequently change lane when they follow a truck, which is expected to increase crash risk. This study introduces several traffic measures at the individual vehicle level to capture vehicle interactions between trucks and non-trucks and analyzed how the measures affect crash risk under different traffic conditions. The traffic measures were developed using headways obtained from Inductive Loop Detectors (ILDs). In addition, a truck detection algorithm using a Gaussian Mixture (GM) model was developed to identify trucks and to estimate truck exposure from ILD data. Using the identified vehicle types from the GM model, vehicle interaction metrics were categorized into three groups based on the combination of leading and following vehicle types. The effects of the proposed traffic measures on crash risk were modeled in two different cases of prior- and non-crash using a case-control approach utilizing a conditional logistic regression. Results showed that the vehicle interactions between the leading and following vehicle types were highly associated with crash risk, and further showed different impacts on crash risk by traffic conditions. Specifically, crashes were more likely to occur when a truck following a non-truck had shorter average headway but greater headway variance in heavy traffic while a non-truck following a truck had greater headway variance in light traffic. This study obtained meaningful conclusions that vehicle interactions involved with trucks were significantly related to the crash likelihood rather than the measures that estimate average traffic condition such as total volume or average headway of the traffic stream.

1. Introduction

Trucks in general traffic streams are distinguished by large physical dimensions, low speeds and limited capabilities of acceleration and deceleration. Consequently, non-trucks tend to show different vehicle-following behavior when they follow a truck and when they follow a non-truck (Green and Yoo, 1999; Stuster, 1999; Kostyniuk et al. 2002). In particular, non-trucks following a truck keep longer headways than those following a non-truck (Ye and Zhang, 2009). In addition, non-truck drivers tend to show frequent lane changing behavior to by-pass a truck even though the speed of truck is not lower than their tolerance level since they experience discomfort behind a truck (Peeta et al., 2005). The influence of trucks on non-truck drivers or even other truck drivers often raises traffic safety issues. Simply, if a leading truck has slower speed than the following vehicle and the following vehicle does not have sufficient time to respond to the slower speed of the leading truck, a likelihood of crash would increase (Lee et al., 2002). Kostyniuk et al (2002) also reported that crashes between cars and trucks were highly related to unanticipated driving maneuvers of vehicle that were possibly influenced by adjacent trucks.

This study explored the causal relationships between crash risk and vehicle interactions involving trucks (e.g., cases of a truck following a non-truck, a non-truck following a truck, and a truck following a truck). It was assumed that traffic flow becomes unstable if vehicles have varying speeds and headways in a traffic stream, which would increase crash risk. Hence, this study aimed to investigate how traffic instability caused by vehicle interactions involved with trucks relates to crash risk in different traffic conditions (e.g. heavy and light traffic). Various

* Corresponding author.
E-mail addresses: kate.hyun@uta.edu (K.K. Hyun), kjeong@mit.edu (K. Jeong), ytok@uci.edu (A. Tok), sritchie@uci.edu (S.G. Ritchie).

https://doi.org/10.1016/j.aap.2018.03.002
Received 31 July 2016; Received in revised form 10 February 2018; Accepted 1 March 2018
Available online 12 March 2018
0001-4575/

traffic measures were developed to estimate the instability of a traffic stream using temporal headways of individual vehicles. This study utilized data at an individual vehicle level collected from Inductive Loop Detectors (ILDs) to capture the vehicle interactions.

Even though many researchers have utilized ILD data for crash analysis, traffic measures provided at an aggregate level, such as occupancy and speed during each 30 s interval, were mainly utilized due to data availability (Golob et al., 2004; Oh et al., 2005; Zheng et al., 2010; Xie et al., 2016). However, these 30-s or longer aggregated measures are not sufficient to capture vehicle interactions and their influences on crash risk. For example, even though multiple vehicles experience different degrees of interaction with a leading vehicle in a short time period, the variation in speeds and headways, which are assumed to be critical determinants for crash occurrence, might be smoothed in aggregate measures. In this vein, finer resolution data such as individual vehicle speeds or headways would provide better measures in crashes. This study utilized ILDs that provided individual vehicle occupancies and time-stamped records when vehicles passed over detectors. Temporal headway was estimated and utilized to represent vehicle interaction measures using time-stamped data of individual vehicles. Moreover, a new method was developed using individual vehicle data to distinguish trucks from non-trucks in this study. The truck detection algorithm developed in this study was based on a Gaussian Mixture (GM) model solely using vehicle duration data from ILDs.

Relationships between the developed traffic measures and crash risk were estimated using a case-control approach based on a conditional logistic regression. For a case-control approach, the proposed traffic measures corresponding to prior-crash and non-crash conditions were used. The models were separately employed in heavy and light traffic to examine if the proposed traffic measures differently affected crash risk in different traffic conditions.

The remainder of this paper is organized as follows. Section 2 presents a review of previous studies, and Section 3 introduces the traffic and crash data used in this study. Section 4 describes the modeling approach of the case-control design, conditional logistic model, and truck detection algorithm. The developed metrics are also introduced in Section 4. An evaluation and discussion of the model results is presented in Section 5, followed by conclusions in Section 6.

2. Literature review

Vehicle interactions between trucks and non-trucks have been studied using headway measures in previous studies. Aghabayk et al. (2012) found that a truck following another truck had the largest headway while a car following another car yielded the smallest headway. Green and Yoo (1999) also showed that headways of cars that follow another car were 10 percent smaller than headways of cars that follow a pickup, bus or truck. Ye and Zhang (2009) analyzed time headways for four different combinations of leading and following vehicles – car-truck, truck-car, truck-truck and car-car – and observed longer headways in descending orders of truck-truck, truck-car, car-truck, and car-car.

These vehicle interactions by vehicle types likely lead to different driving behaviors. Peeta et al. (2005) studied driving behavior of non-truck drivers in the vicinity of trucks. Discomfort level for non-truck drivers in car-truck interactions was found, and non-truck drivers were more likely to change lanes if they follow a truck less than a 2-s gap. The finding showed that the interactions of cars with trucks were distinct from interactions with other cars. Studies have shown that these driving behaviors often lead to risky conditions on roads. Stuster (1999) conducted an interview survey with collision investigators, trucking experts and truck drivers and found that abrupt maneuvers of vehicles in the vicinity of trucks and driving too close to a truck were the most significant factors in collisions involving large trucks. Kostyniuk (1998) presented similar results showing that improper vehicle following and lane-changing in the vicinity of trucks were important factors in car-truck crashes.

Despite its significant impact on crash risk, proportion of trucks and vehicle interactions with trucks have been given less attention in safety analysis due to lack of individual vehicle data. Golob and Regan (2004) found that the proportion of trucks, particularly large (five axle or more) trucks, had a positive relationship with accident likelihood; however annual average truck volume was used to represent the truck exposure. Such annual traffic measures only represent general traffic conditions and have significant limitations in representing traffic when the crash occurred (Sullivan, 1990).

With the advent of advanced traffic surveillance systems, traffic measures that are observed in shorter time intervals, such as every 30 s, have been recently utilized in safety analysis, such as traffic volume, speed, and density. However, these measures still lack consideration of the influence of trucks as crash determinants. Golob et al. (2004) showed that smaller variations in speed and flow obtained from 30 s aggregate traffic data reduced crash rates.

As a critical precursor of crashes, flow disruptions have been described using speed and density variations (Lee et al., 2002; Oh et al., 2005; Zheng et al., 2010). Lee et al. (2002) identified that high variations in speed and high density traffic condition represented the riskiest traffic conditions, and Zheng et al. (2010) concluded that standard deviation of speed was the most significant variable affecting crash likelihood. Xu et al. (2013) showed that impacts of traffic flow characteristics on severity of crashes were significant. In severe crashes, high speed and large speed variance were critical factors. Oh et al. (2005) identified the standard deviation of speed in 5-min intervals as the best indicator for accident prediction. Similar findings by Abdel-Aty and Abdelwahab (2004) pointed out that 5 min average occupancy at an upstream segment during the preceding 5–10 min and 5 min speed variation at a downstream segment were significant factors that affect crash risk. Ahmed et al. (2012) utilized Automatic Vehicle Identification (AVI) systems to collect traffic data at adjacent segments where crashes occurred. The standard deviation of speed at the segment before the crash was associated with increased crash risk, while average speed of the downstream segment showed negative impact on crash risk. Hourdos et al. (2006) captured live crashes using a video camera and found that crash likelihood was affected by speed variations and environmental factors such as lighting.

The aforementioned studies are in consensus that speed-related conditions are one of the key factors for crash occurrence. However, there is disagreement in the association of speed with the type or severity of crashes. Some have argued that high speed increases the likelihood of severe crashes (Davis et al., 2006; Yan and Radwan, 2006; Xu et al., 2013) while others showed speed has a negative relationship with crash risk, especially in rear-end crashes involving two vehicle crashes (Christoforou et al., 2011). Similarly, effects of congestion on crash risk were investigated in several studies. Shefer and Rietveld (1997) found that congestion increased the risk of crashes because of the increased interactions between vehicles. However, Noland and Quddus (2005), and Quddus et al. (2009) concluded that there is little evidence of association between congestion and crash frequencies.

Previous studies showed that different vehicle interaction was observed by leading and following vehicle types. In addition, crash analysis literature investigated several traffic measures as crash determinants. However, a gap still exists in the study of traffic measures that consider vehicle interactions by vehicle types and the impact of these traffic measures on crash risk. To the best of our knowledge, this is the first study to address this void by investigating the detailed interactions of vehicles with trucks at the vehicular level and associating them with crash occurrences.

3. Data

3.1. Traffic data

ILD systems are predominant in the U.S., and are one of the most common data sources in crash analysis (Golob et al., 2004; Oh et al., 2005; Zheng et al., 2010). Data from ILDs are typically aggregated in 30 s – or longer – intervals and produce measures of volume, occupancy and sometimes average speed. These measures are publically available in California through the Performance Measurement System (Caltrans Performance Measurement System (PeMS, 2016). However, the available aggregate measures from this system are inadequate for capturing vehicle interactions because detection data are not available at the individual vehicle level.

This study is based on individual vehicle-level data records obtained directly from freeway ILDs. Temporal headways were primarily used to investigate vehicle interactions. ILDs are well-suited for crash analysis because of their prevalence in California and the United States. Furthermore, traffic measures required for this analysis can be obtained without any in-pave installation or retrofit of existing detection systems. The impact area of the ILD sites was defined as a one-mile segment centered by the inductive loop sensors. If the impact area contained lane drops or highway intersections, those sites were not included since traffic volumes could significantly change within the area. This process ensures that the data obtained from ILDs can represent traffic conditions of the impact area. As shown in Fig. 1, a total of 38 ILD sites (bi-directional) located on highways were selected throughout the State of California.

It is a commonly reported issue that ILDs can sometimes provide erroneous traffic measures due to loop calibration or data transmission problems (Zheng et al., 2010). Therefore prior to developing a model, a data pre-filtering step was implemented to remove erroneous records such as vehicles with abnormal duration (e.g., greater than 10 s) or speed (e.g., greater than 120 mph). In addition, daily total volume was evaluated and days with traffic that deviated significantly from the expected volume were excluded from the dataset using an interquartile range (Washington et al., 2010). It should be noted that the data pre-filtering step using a duration and a speed threshold removes abnormal individual vehicle records typically caused by data communication issues or abnormal vehicle lateral positioning on the loop during lane changing. However, the volume inspection filters abnormal daily data caused by a temporary malfunction of the ILD. The dataset used in this study contained a total of 827,633 individual vehicle data records obtained from February 2015 through April 2016. From the data pre-filtering step by speed and duration thresholds, 591 individual data records were removed, which consisted of approximately 0.0007% of the total data records.

Of note, this study did not consider an impact from motorcycles. Since motorcycles in California are allowed line sharing, inductive loops are not able to capture them effectively. Because of this characteristic, motorcycles would not also be expected to have a car-following situation.

3.2. Crash data

Crash data were obtained from the Statewide Integrated Traffic Records Systems (SWITRS, 2016). The SWITRS compiles crash records from the California Highway Patrol (CHP) and its allied agencies throughout California. Crashes in the impact area were selected and linked to the traffic data using QGIS software. A total of 375 reported accidents from February 2015 to April 2016 were analyzed in this study. Prior to the modeling, the crash data were cleaned by removing possible erroneous data such as a record that includes inconsistency in the same items asked multiple times in the report or that does not contain information on general questions (e.g., weather or location). Although there is a possibility that a reported crash time differs from an actual crash time, SWITRS records crash time with a higher resolution (minute), which would maintain quality and reliability of the data source (Zheng et al., 2010). Therefore, the filtering step applied to the high resolution data likely removes invalid records that contain a large time lag between an actual and a reported crash time. A minor lag is negligible in a modeling process because 15-min time window was applied to aggregate traffic measures, as explained in a later section (see Section 4.1).

4. Modelling approach

4.1. Case control design

Prior- and non-crash traffic conditions were initially analyzed in preparation for modeling the relationships between vehicle interactions and crash risk. A prior-crash condition is defined as a predetermined time window before a crash occurs. The ideal time windows prior to crashes should be determined by considering traffic states because excessively short or long windows cannot adequately capture flow disruptions or traffic instability that increases crash risk. Tests of time window lengths ranging from 5 min to 60 min in 5-min increments were performed, and resulted in the recommended length of 15 min.

Two main approaches for selecting non-crash cases were adopted in previous studies: random extraction (Xu et al., 2013), and case-control design (Híjar et al., 2000; Zheng et al., 2010; Ahmed et al., 2012; Gross, 2013). In the random extraction, non-crash cases are randomly chosen from any time period among non-crash days for the given length of time window. Therefore, even if a crash occurs during a non-peak time, non-crash data may include traffic measures during a peak time period. As a consequence, these comparative datasets (non-crash datasets) may not capture effects of traffic conditions that are relevant to crash cases because exogenous factors such as dissimilar traffic volume or lighting conditions can be included in crash and non-crash cases and adversely affect the model (Davis et al., 2006).

In this regard, the case-control design is more appropriate for understanding traffic exposure (i.e., traffic measures) and its associated outcomes (i.e., accidents) (Lewallen and Courtright, 1998; Ahmed et al., 2012). In the case control design, traffic data in the control cases (i.e., non-crash cases) were chosen from the same time period when a crash case was collected but from randomly selected days that a crash did not occur. For example, if a crash occurred on March 10th at 3PM, traffic data between 2:45PM–3:00PM would be chosen as a crash case based on the determined length of time window. Control cases (i.e., non-crashes cases) then would be chosen for the identical time window of 2:45PM–3:00PM from randomly chosen multiple days when crashes did not occur. Since the control data were collected from the same time window but for randomly chosen days, the influence of traffic measures on crashes can be better captured using this approach.

Another important issue in the control-case design is to determine the optimal control-to-case ratio, as this may significantly affect complexity and estimates of the models. Previous studies (Rothman and Greenland, 1986; Hennekens et al., 1987; Grimes and Schulz, 2005) recommended applying a 4:1 control-to-case ratio by considering cost-effectiveness. In addition, it was observed that statistical power was not improved when one case had more than 4 controls. This study implemented the 4:1 ratio based on the literature recommendations.

4.2. Conditional logistic regression modeling

A case-control study widely adopts a conditional logistic model because each case (prior-crash) and its corresponding controls (non-crashes) depend entirely on each other within the model, thereby considering the paired effect of the case-controls (Breslow and Day, 1993). Moreover, if a large number of cases exist, a conditional logistic model is much more advantageous than an unconditional one. This is because the unconditional method will require far too many dummy

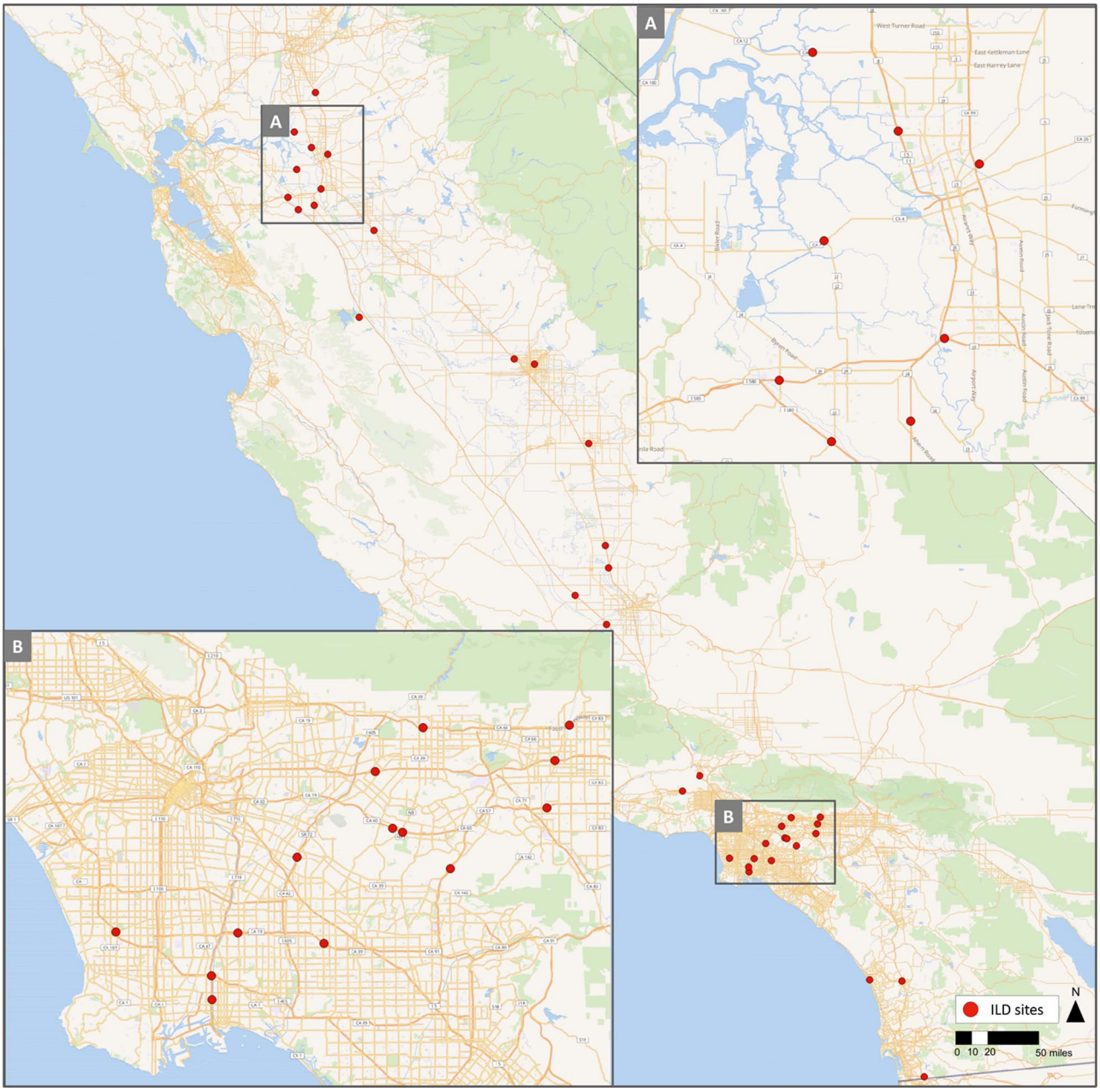


Fig. 1. Study sites.

variables – one for each matching pair. Because of its conceptual advantages in a case-control study, the conditional logistic regression model provides consistent and unbiased estimates.

Suppose that each set i contains a case and corresponding M control cases randomly extracted from the location where the case occurs, the conditional logistic regression can be written as follows.

$$P(Y = crash) = \prod_{i=1}^{I} \frac{\exp(\sum_{k=1}^{K} \beta_k X_{i0k})}{\sum_{j=0}^{M} \exp(\sum_{k=1}^{K} \beta_k X_{ijk})} = \prod_{i=1}^{I} \frac{1}{1 + \sum_{j=1}^{M} exp(\sum_{k=1}^{K} (\beta_k X_{ijk} - \beta_k X_{i0k}))} \quad (1)$$

where i represents a matched set including one case and M control cases, I represents the number of matched sets, j represents a case or control case (i.e., j = 0 is a case, j = 1,…,M is a control case), M represents the number of control cases, X_{ijk} represents k^{th} explanatory variable for i and j, and β_k represents a parameter of the k^{th} variable.

4.3. A truck detection algorithm using Gaussian Mixture (GM) model

A binary truck detection algorithm was developed to distinguish trucks and non-trucks from ILD data. The definition of trucks in this study refers to tractor-trailer trucks which correspond to FHWA axle based class 8–13. Non-trucks refer to passenger cars, pickups, bus and single unit trucks, which represent FHWA class 2 through 7 (FHWA, 2001). Of note, FHWA class is a vehicle classification scheme determined by the number of axles, axle spacing, and weight, which include 13 vehicle classes in the US.

Several previous studies have investigated the classification of vehicles based on ILD data from single loop sensors. Initial attempts provided percentage of long vehicles using aggregated flow and occupancy measures (Kwon et al., 2003; Wang and Nihan, 2004). Coifman and Kim (2009) developed an individual length-based classification scheme. Individual vehicle speed and length estimated by ILD data were utilized for vehicle classification. However, this approach requires

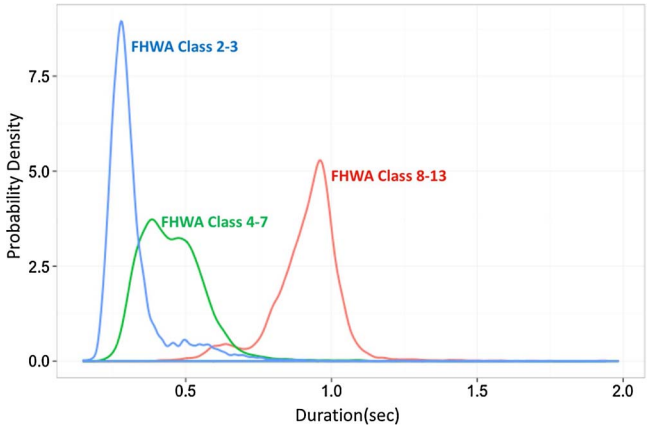


Fig. 2. Duration distributions by vehicle type.

several pre-determined traffic values to estimate individual vehicle length and speed such as defining speeds higher than 45 mph as free flow. These assumptions may result in inaccurate individual speed estimates under different traffic conditions, which eventually yield misclassification outcomes. Their results also showed that their model performance was not reliable during congested periods.

This algorithm focuses on the difference in duration by vehicle type to identify trucks. Duration data of three different vehicle types which correspond to FHWA class 2–3, FHWA class 4-7, and FHWA class 8-13 were collected from the four ILD sites in California. For more details on data collection sites and data processing methods, refer to Hernandez et al. (2016). As illustrated in Fig. 2, the duration of FHWA class 8–13 vehicles is noticeably longer than the others, which indicates that duration can be used to distinguish trucks from non-trucks.

However, the range of duration measures associated with trucks and non-trucks vary across traffic conditions. For example, the ranges are typically longer in congested traffic compared with uncongested traffic. Therefore, in order to use the duration range as an identifier of vehicle type, the range should be updated at short time intervals to effectively capture the changes in traffic state. This study developed an approach using the Gaussian Mixture (GM) model to determine and update the duration ranges by vehicle type over short time periods, deemed as 15 min in this study to obtain sufficient volumes to construct GM model. It should be noted that shorter or longer time period is easily adapted in the proposed GM model.

A GM model is a linear composition of Gaussian distributions, $\mathcal{N}(\mu_m, \Sigma_m)$ with a mixing proportion of p_m (Friedman et al., 2001). Since this study estimated two duration distributions – for non-trucks and trucks – bimodal GM distributions were estimated by applying a mixing proportion of the two distributions.

$$\mathbf{f(x)} = \sum_{\mathbf{m}=1}^{\mathbf{M}} \mathbf{p_m} \bullet \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{m}, \boldsymbol{\Sigma}_\mathbf{m}) \tag{2}$$

where m is number of mixture components, $\mathcal{N}(\mu_m, \Sigma_m)$ is a Gaussian distribution with mean μ and covariance matrix Σ, and p_m is the mixing proportion.

In this study, since a one dimensional GM model is applied, Σ_m can be replaced by σ_m^2, where σ is the standard deviation. The proposed algorithm has several practical advantages. Implementation of the algorithm is not restricted by temporal or spatial conditions since the algorithm can reflect the change of traffic state in real-time. Additionally, any labor extensive data collection and process are not required to implement the algorithm at different ILD sites since the proposed algorithm does not need a training step or any assumptions for model development.

4.4. Metric development

In this study, various metrics were developed to capture the interactions between vehicles based on the individual traffic measures obtained from ILDs. Timestamp and duration was obtained when a vehicle passes over the loop sensor. Headways were then directly calculated from the timestamp differences of successive vehicles. The headway measures were subsequently categorized into three categories defined by combination of a leading and following vehicle type determined by the GM model – (i) a non-truck following a truck (NT), (ii) a truck following a non-truck (TN), and (iii) a truck following a truck (TT). The average and standard deviation of headway measures for the determined length of time window by following vehicle type were used to represent vehicle interactions in this study.

Measures representing general traffic conditions such as total volume, truck proportion, headway average and variation, and speed were also included and compared with the measure that captured the individual level of vehicle interactions in the model. Truck proportion was estimated using the vehicle types identified by the GM model. Vehicle speed was estimated by vehicle duration as follows.

$$\textbf{speed (ft/s)} = \frac{3600}{5200} * \frac{(\textbf{\textit{vehicle length}} + \textbf{\textit{detector length}})}{\textbf{\textit{occupancy}}} \tag{3}$$

However, individual vehicle length varies by vehicle type and cannot be obtained from ILDs. Therefore, there have been several studies to estimate effective vehicle length, which is approximately the sum of the length of the loop detector and vehicle length, by vehicle type. Kwon et al. (2003) observed a bimodal distribution of effective vehicle lengths where non-trucks and trucks showed average lengths of 18.6 feet and 61.2 feet, respectively. Coifman and Kim (2009) estimated 20 feet and 70 feet for non-trucks and trucks, respectively. This paper adopted the average effective vehicle lengths of these two studies. Thus, a non-truck was represented as 19 feet long and a truck as 65 feet long to estimate the average speed of the traffic stream. Table 1 presents

Table 1
Descriptions of explanatory variables.

Variables[a]	Description
Total volume (Vol)	Total number of vehicles
Truck proportion (TruckPr)	Percentage of trucks over the total number of vehicles
Headway average (AvgHw)	Average temporal headway of all vehicles
Headway variation (DevHw)	Standard deviation temporal headway of all vehicles
Speed average (Speed)	Average speed of all vehicles
Headway average in NT (NTAvgHw)	Average temporal headway of vehicles in NT (non-truck following a truck)
Headway variation in NT (NTDevHw)	Standard deviation temporal headway of vehicles in NT (non-truck following a truck)
Headway average in TN(TNAvgHw)	Average temporal headway of vehicles in TN (truck following a non-truck)
Headway variation in TN(TNDevHw)	Standard deviation temporal headway of vehicles in TN (truck following a non-truck)
Headway average in TT(TTAvgHw)	Average temporal headway of vehicles in TT (truck following a truck)
Headway variation in TT(TTDevHw)	Standard deviation temporal headway of vehicles in TT (truck following a truck)

[a] All variables were estimated for the determined length of time window (This study used 15 min window from the time of the crash).

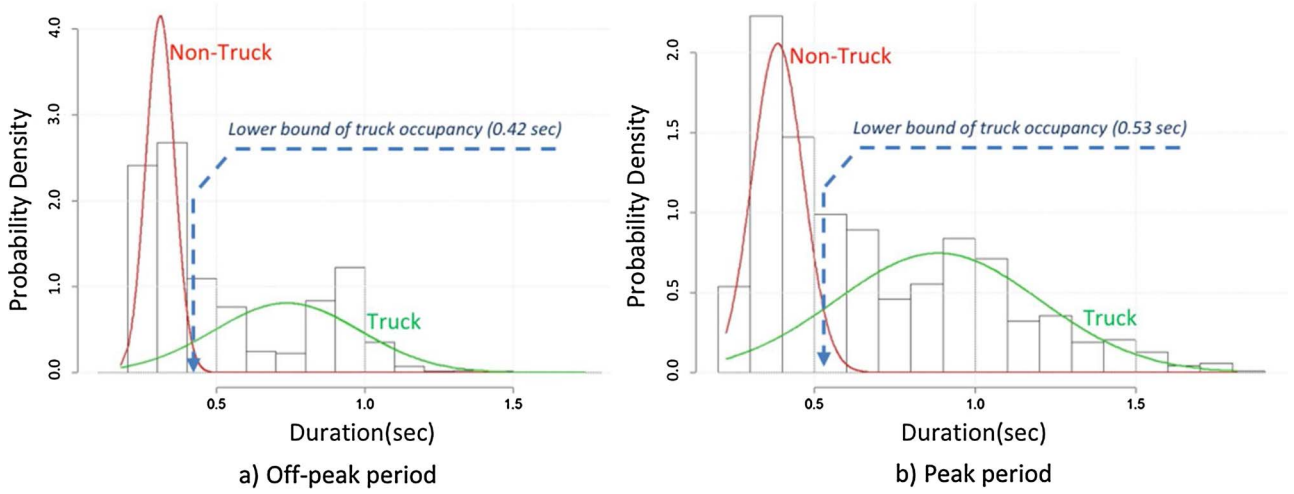


Fig. 3. Duration distribution for non-trucks (red) and trucks (green) in off-peak (a) and peak (b) period. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

brief descriptions of the variables used in this study.

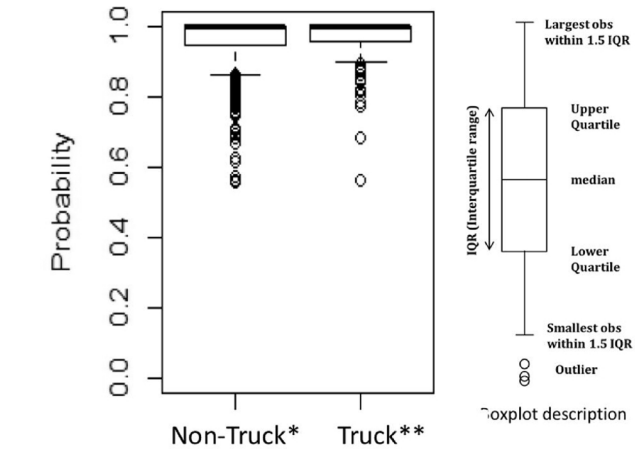
5. Result

5.1. Validation of a truck detection algorithm

The proposed truck detecting GM algorithm was tested with 28,328 vehicles collected at four ILD sites in California from the previous study (Hernandez et al., 2016). Along with the individual vehicle duration and timestamp, side-fire images for each passing vehicle were stored together in a database to identify vehicle types (i.e. non-truck or truck). In the algorithm, every individual vehicle was classified into two types, truck or non-truck, based on its duration. In every 15 min, duration ranges for trucks and non-trucks were updated in the GM model. Fig. 3 shows results of two sample duration densities from an off-peak (a) and peak period (b), respectively. In these examples, the lower bound of duration for trucks in an off-peak period was 0.42 s while a peak period showed a lower bound of 0.53 s, which was 0.11 s longer than that of the off-peak time period.

To validate the proposed model, the estimated vehicle types were compared to the actual vehicle types. A total of 323 15-min periods were collected from the dataset. Since each time period provided classification results (i.e. probability of correct classification), summary statistics of mean, 25 percentiles, and 75 percentiles for correct classification cases are illustrated using a box-plot approach as shown in Fig. 4. The average correct classification rates were 95% for non-trucks and 97% for trucks. In other words, only 5% of non-trucks were identified as trucks, and 3% of trucks were classified as non-trucks.

The proposed algorithm was applied to a set of sample data obtained from 8 different ILD sites among 38 study sites in California observed for 48 h from March 29th (Tuesday) to March 30th (Wednesday), 2016. As shown in Fig. 5, the proportion of trucks varied temporally and spatially for the 48-h period. Although repetitive time of day patterns were observed at some locations, the times with higher truck traffic varied by site. For example, site 1 and 3 showed high proportions of truck traffic during the night time while high proportions of truck traffic were observed during the daytime in site 2 and 4. These findings confirmed that aggregated truck proportion measure using an annual or daily basis cannot capture truck travel pattern accurately, and this may further cause misinterpretation of general traffic conditions, vehicle interactions and their influence on crash risk.



	Non-Truck*	Truck**
Mean	0.951	0.969
Median	1.0	1.0
Lower Quartile	0.945	0.956
Upper Quartile	1.0	1.0

* Non-Truck: A non-truck classified as a non-truck

** Truck: A truck classified as a truck

Fig. 4. Truck detection algorithm results.
* Non-Truck: A non-truck classified as a non-truck.
** Truck: A truck classified as a truck.

5.2. Results of conditional logistic regression model

This study employed the conditional logistic regression models for different traffic conditions. The traffic conditions were classified using average lane occupancy in the traffic stream. Average lane occupancy rate has been used to evaluate traffic conditions in real-time transportation information systems due to its simplicity in quantifying traffic congestion (Lomax et al., 1997). This paper identified three traffic conditions; heavy traffic (average occupancy greater than 25 percent), medium (average occupancy between 15 percent and 25 percent), and light traffic (average occupancy less than 15 percent). Table 2 presents the basic statistics of median and standard deviation for the explanatory variables categorized by accident (prior-crash cases) and non-accident conditions (non-crash cases). The statistics were calculated for 15 min

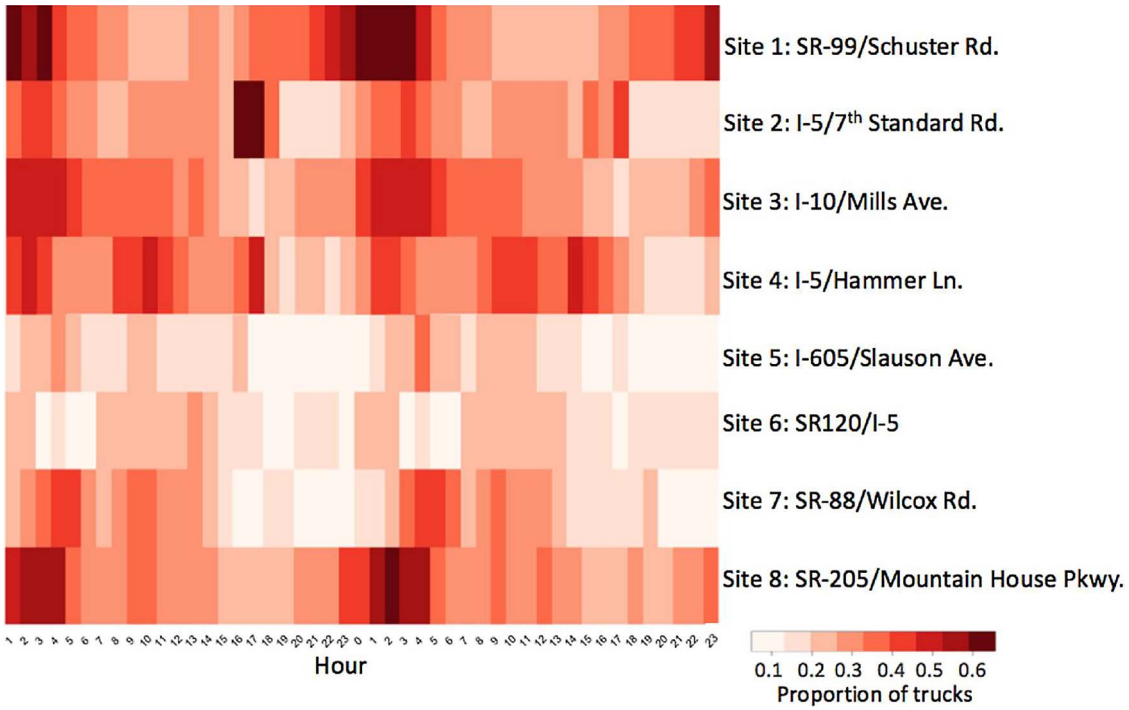


Fig. 5. Proportion of trucks in different segments of freeway in California.

windows before a crash occurred. Proportions of trucks (TruckPr) ranged from 15% to 25% in all traffic conditions. Individual level of headway measures showed varying results by different vehicle-following types. In accident conditions, truck following truck (TTAvgHw) showed the largest average headway in heavy traffic while truck following non-truck (TNAvgHw) showed the largest average headway in light traffic. In non-accident conditions, truck following non-truck (TNAvgHw) showed the largest headway average in medium and light traffic conditions while both truck following non-truck (TNAvgHw) and truck following truck (TTAvgHw) showed large headway average in heavy traffic. In individual headway variation measure, truck following non-truck (TNDevHw) showed the largest standard deviation in medium and light traffic condition for both accident and non-accident cases whereas truck following truck (TTDevHw) presented the largest standard deviation in heavy traffic condition.

Table 3 shows the model coefficients and goodness of fit results. The performance of the model was evaluated by two types of goodness of fit test corresponding to likelihood ratio and Lagrangian multiplier test. All three models show acceptable fit in the overall goodness of fit tests. In addition, multicollinearity between variables was tested before estimating the model, and there is no correlation found between the variables.

Headway measures by vehicle-following types were highly associated with crash risk, however their impacts on crash risk were different by traffic condition. In heavy traffic conditions, a one-unit increase in truck following non-truck headway standard deviation (TNDevHw) significantly increases the likelihood of crash occurrence by about 223% (odds ratio of 2.23). This indicates that crash risk in heavy traffic conditions increases when headway variation between a leading non-truck and a following truck is increased. However, a one-unit increase in average headway of truck following non-truck (TNAvgHw) reduced crash risk by 74% (Odds ratio of 0.26). In medium traffic conditions, longer average headway in the non-truck following truck case (NTAvgHw) likely reduces crash risk. In light traffic conditions, both headway average and standard deviation of non-truck following truck cases (NTAvgHw, NTDevHw) were highly associated with crash risk. Specifically, a one-unit increase in standard deviation of headway of non-truck following truck (NTDevHw) leads to a 15% increase in likelihood of crash occurrence (Odds ratio of 1.15), however a one-unit increase in its average headway (NTAvgHw) reduces the crash

Table 2
Summary Statistics for explanatory variables in 15 min time window.

	Heavy Traffic				Medium Traffic				Light Traffic			
	Accident		Non-Accident		Accident		Non-Accident		Accident		Non-Accident	
	median	std.dev	median	std.dev	median	std.dev	median	std.dev	median	std.dev	median	std.dev
TruckPr	20.56	10.02	16.62	9.12	24.74	9.15	19.54	9.81	15.63	10.92	15.69	11.21
AvgHw	3.13	0.91	3.04	1.26	2.74	0.74	2.87	0.75	3.99	4.07	4.19	4.45
DevHw	1.84	2.82	1.96	1.67	1.86	2.57	2.09	1.59	3.69	4	3.82	4.73
Speed	21.79	7.72	28.13	14.66	43.36	7.89	48.53	9.02	52.83	4.62	53.08	6.27
NTAvgHw	1.32	3.08	1.5	2.62	1.36	1.52	1.69	1.44	3.46	4.33	3.55	4.96
NTDevHw	3.77	2.68	3.52	2.4	2.97	0.79	3.02	0.87	4.31	4.09	4.45	4.63
TNAvgHw	1.91	8.3	2.28	1.84	2.43	4.64	2.62	2.46	4.29	3.85	4.54	5.03
TNDevHw	3.77	2.43	3.96	1.78	3.88	1.56	4.04	1.02	5.81	4.31	6.02	5.04
TTAvgHw	2.6	5.51	2.29	4.22	1.87	6.4	1.97	1.29	3.52	4.34	3.52	4.45
TTDevHw	5.22	3.75	4.35	3.4	3.7	3.67	3.7	0.96	5.06	4.22	5.35	4.43

Table 3
Model coefficients and goodness of fit results.

Variable	Heavy Traffic			Medium Traffic			Light Traffic		
	Coeff.	Odds Ratio	Std.Err	Coeff.	Odds Ratio	Std.Err	Coeff.	Odds Ratio	Std.Err
TruckPr	0.15***	1.16	0.05	0.06*	1.07	0.03	−0.04**	0.96	0.01
AvgHw	−1.09	0.34	1.53	−0.96	0.38	1.29	0.29	1.33	0.19
DevHw	−1.44	0.24	0.95	−0.08	0.92	0.34	−0.32	0.72	0.13
Speed	−0.11**	0.90	0.05	−0.08**	0.92	0.04	0.04*	1.04	0.02
NTAvgHw	1.02	2.77	0.65	−2.11**	1.37	0.51	−0.17*	0.85	0.09
NTDevHw	−0.84	0.43	0.53	0.37	1.07	0.19	0.14**	1.15	0.06
TNAvgHw	−1.36*	0.26	0.75	0.32	0.12	0.96	−0.03	0.97	0.07
TNDevHw	0.80**	2.23	0.37	0.07	1.45	0.43	0.04	1.04	0.06
TTAvgHw	0.34	1.40	0.35	0.14	1.15	0.36	0.03	1.03	0.07
TTDevHw	0.49	1.63	0.31	0.06	1.06	0.20	0.00	1.00	0.06
Goodness of Fit									
Likelihood ratio test		64.9***			36.1***			23.5***	
Lagrangian multiplier test		22.4***			21.4***			18.8**	

*** p > |z| 0.01.
** p > |z| 0.05.
* p > |z| 0.1.

likelihood by 15% (Odds ratio of 0.85).

Notably, speed was consistently chosen as a significant traffic measure contributing to crash risk in all traffic conditions. However, the impact of speed on crash risk was either positive or negative depending on traffic conditions. As speed increased, crash risk is more likely decreased in heavy and medium traffic conditions whereas it more likely increased in light traffic condition. Truck proportion (TruckPr) also affects crash risk in all traffic conditions. However, the impact of truck proportion on crash risk was either positive or negative depending on traffic conditions. As truck proportion is increased, crash risk is more likely increased in heavy and medium traffic condition, whereas corresponding crash risk is more likely decreased during light traffic.

Interestingly, the average and standard deviation headway of the general traffic stream (AvgHw and DevHw) without considering vehicle-following types were not statistically significant factors for crash occurrence in any traffic condition. These results suggest that the proposed traffic measures reflecting vehicle interactions were able to better explain crash risk compared to general traffic measures, which have been commonly used in previous studies. Particularly, when vehicles interacting with trucks have shorter and varying headway, crash risk is higher. Moreover, their impacts on crash risk by different type of vehicle-following cases were significantly different depending on traffic conditions. This could be because multiple factors such as visibility, speed, and acceleration and deceleration capability of vehicle types were associated with vehicle-following behavior. However, in general, maintaining longer and constant headways with trucks tends to reduce the crash risk in all traffic conditions.

6. Conclusions

This study investigated relationships between vehicle interactions influenced by trucks and crash risk. ILD data were utilized in this study to estimate vehicle interactions at an individual vehicle level. Using time-stamps and duration records from the selected ILDs, this study defined various traffic metrics. Specifically, headway metrics in different vehicle-following types, such as a non-truck following a truck and a truck following a non-truck, were introduced to measure vehicle interactions in the traffic stream. Individual vehicle speed and the proportion of trucks were also estimated using individual vehicle duration data. To identify trucks from the ILD data, a truck detection algorithm was developed using a Gaussian Mixture (GM) model. The algorithm was tested with over 28,000 vehicles collected from California and showed 97 percent correct matching rate to the truck detection.

To investigate the effects of the proposed metrics on crash risk, prior- and non-crash cases were applied in a case-control approach with a conditional logistic regression. The model was estimated with 375 cases of crash data in California and 827,633 associated traffic measures of individual vehicle in prior- and non-accident cases. Results highlighted that the proposed traffic measures estimating vehicle interactions were significantly associated with crash risk, however their impacts varied by traffic conditions. Specifically, crashes more likely occurred when a truck following a non-truck had a shorter average headway and greater headway variance in heavy traffic while a non-truck following a truck had shorter average headway but greater headway variance in light traffic. Speed was also a significant variable of crash likelihood in all traffic conditions. However, the average and standard deviation of headway without considering vehicle-following types did not affect the crash likelihood in this study. From these results, we obtained meaningful conclusions that vehicle interactions involved with trucks significantly relate to the crash likelihood rather than total volume or average headway of the traffic stream.

The findings of this study can be used to identify high risk truck corridors for safety improvement strategies. Some possible solutions include instrumenting truck-only lanes to separate truck and non-truck traffic, developing a real-time traffic detection system using ILDs to identify unsafe traffic conditions, and implementation of a driver warning system using intelligent transportation systems technologies such as variable message signs or vehicle on-board devices. Advanced technologies facilitating truck platooning could be also considered an active crash prevention strategy since truck and non-truck interactions would be minimized in an overall traffic stream. As this paper focused on associations between vehicle interactions and crash, other factors that possibly affect crash risk such as weather, light, and road geometry were not considered in the model. A future study will explore more contributing attributes including road, environmental, and driver factors as well as the vehicle interactions and analyze their impacts on crash frequencies and severities.

Acknowledgements

The authors gratefully acknowledge the assistance provided by the California Air Resources Board and California Department of Transportation (Caltrans) with the data collection. The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of California Air

Resources Board and California Department of Transportation. This paper does not constitute a standard, specification, or regulation.

References

Abdel-Aty, M., Abdelwahab, H., 2004. Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. Accid. Anal. Prev. 36 (3), 447–456.

Aghabayk, K., Sarvi, M., Young, W., 2012. Understanding the dynamics of heavy vehicle interactions in car-following. J. Transp. Eng. 138 (12), 1468–1475.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012. Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. Transp. Res. Rec.: J. Transp. Res. Board 2280, 60–67.

Breslow, N.E., Day, N.E., 1993. Statistical methods in cancer research. The Analysis of Case-Control Studies, vol. 1 International Agency for Research on Cancer, 1980, Lyon, France.

Caltrans Performance Measurement System (PeMS), Access June 2016. http://pems.dot.ca.gov/.

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2011. Identifying crash type propensity using real-time traffic data on freeways. J. Saf. Res. 42 (1), 43–50.

Coifman, B., Kim, S., 2009. Speed estimation and length based vehicle classification from freeway single-loop detectors. Transp. Res. Part C: Emerg. Technol. 17 (4), 349–364.

Davis, G.A., Davuluri, S.U.J.A.Y., Pei, J., 2006. Speed as a risk factor in serious run-off-road crashes: Bayesian case-control analysis with case speed uncertainty. J. Transp. Stat. 9 (1), 17.

Federal Highway Administration, 2001. Traffic Monitoring Guide. FHWA-PL-01-021. U.S. Department of Transportation. http://www.fhwa.dot.gov/ohim/tmguide/index.htm.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, vol. 1. Springer series in statistics, New York, pp. 337–387.

Golob, T.F., Regan, A.C., 2004. Traffic Conditions and Truck Accidents on Urban Freeways. Institute of Transportation Studies.

Golob, T.F., Recker, W.W., Alvarez, V.M., 2004. Freeway safety as a function of traffic flow. Accid. Anal. Prev. 36 (6), 933–946.

Green, P., Yoo, H., 1999. Driver Behavior While Following Cars, Trucks, and Buses.

Grimes, D.A., Schulz, K.F., 2005. Compared to what? Finding controls for case-control studies. Lancet 365 (9468), 1429–1433.

Gross, F., 2013. Case–control analysis in highway safety: accounting for sites with multiple crashes. Accid. Anal. Prev. 61, 87–96.

Hennekens, C.H., Buring, J.E., Mayrent, S.L., 1987. Epidemiology in Medicine. Little, Brown and Company, Boston.

Hernandez, S., Tok, A., Ritchie, S.G., 2016. Multiple-Classifier Systems for Truck Body Classification at WIM Sites with Inductive Signature Data. Transp. Res. Part C Emerging Technol. 68, 1–21.

Híjar, M., Carrillo, C., Flores, M., Anaya, R., Lopez, V., 2000. Risk factors in highway traffic accidents: a case control study. Accid. Anal. Prev. 32 (5), 703–709.

Hourdos, J., Garg, V., Michalopoulos, P., Davis, G., 2006. Real-time detection of crash-prone conditions at freeway high-crash locations. Transp. Res. Rec.: J. Transp. Res. Board 1968, 83–91.

Kostyniuk, L.P., 1998. Exploring Rear-End Roadway Crashes from the Driver's Perspective. No. UMTRI-98-52. .

Kostyniuk, L.P., Streff, F.M., Zakrajsek, J., 2002. Identifying Unsafe Driver Actions That Lead to Fatal Car-Truck Crashes, vol. 9 AAA Foundation for Traffic Safety No. 10. www.aaafoundation.org/sites/default/files/CarTruck.pdf.

Kwon, J., Varaiya, P., Skabardonis, A., 2003. Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation. Transp. Res. Rec.: J. Transp. Res. Board 1856, 106–117.

Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. Transp. Res. Rec.: J. Transp. Res. Board 1784, 1–8.

Lewallen, S., Courtright, P., 1998. Epidemiology in practice: case-control studies. Commun. Eye Health 11 (28), 57.

Lomax, T.J., Turner, S.M., Shunk, G., Levinson, H.S., Pratt, R.H., Bay, P.N., Douglas, G.B., 1997. NCHRP Report 398: Quantifying Congestion, vol. 1. Final Report. TRB, National Research Council, Washington, DC.

Noland, R.B., Quddus, M.A., 2005. Congestion and safety: a spatial analysis of London. Transp. Res. Part A: Policy Pract. 39 (7–9), 737–754.

Oh, J.S., Oh, C., Ritchie, S.G., Chang, M., 2005. Real-time estimation of accident likelihood for safety enhancement. J. Transp. Eng. 131 (5), 358–363.

Peeta, S., Zhang, P., Zhou, W., 2005. Behavior-based analysis of freeway car–truck interactions and related mitigation strategies. Transp. Res. Part B: Methodol. 39 (5), 417–451.

Quddus, M.A., Wang, C., Ison, S.G., 2009. Road traffic congestion and crash severity: econometric analysis using ordered response models. J. Transp. Eng. 136 (5), 424–435.

Rothman, K.J., Greenland, S., 1986. Modem Epidemiology 2. Little, Brown, Boston, pp. 23–31.

Shefer, D., Rietveld, P., 1997. Congestion and safety on highways: towards an analytical model. Urban Stud. 34 (4), 679–692.

Statewide Integraded Traffic Records System (SWITRS), Accessed June 2016. http://iswitrs.chp.ca.gov/Reports/jsp/userLogin.jsp.

Stuster, J., 1999. The Unsafe Driving Acts of Motorists in the Vicinity of Large Trucks (No. Final Report). Anacapa Sciences.

Sullivan, E.C., 1990. Estimating accident benefits of reduced freeway congestion. J. Transp. Eng. 116 (2), 167–180.

Wang, Y., Nihan, N.L., 2004. Dynamic estimation of freeway large-truck volumes based on single-loop measurements. July. Intelligent Transportation Systems, vol. 8. Taylor & Francis Group, pp. 133–141 No. 3.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis. CRC Press.

Xie, W., Wang, J., Ragland, D.R., 2016. Utilizing the eigenvectors of freeway loop data spatiotemporal schematic for real time crash prediction. Accid. Anal. Prev. 94, 59–64.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39.

Yan, X., Radwan, E., 2006. Analyses of rear-end crashes based on classification tree models. Traffic Inj. Prev. 7 (3), 276–282.

Ye, F., Zhang, Y., 2009. Vehicle type-specific headway analysis using freeway traffic data. Transp. Res. Rec.: J. Transp. Res. Board 2124, 222–230.

Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. Accid. Anal. Prev. 42 (2), 626–636.