Transportation Research Part C 24 (2012) 288–298

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

Real-time prediction of visibility related crashes

Mohamed A. Abdel-Aty [a], Hany M. Hassan [a,b,*], Mohamed Ahmed [a], Ali S. Al-Ghamdi [b]

[a] University of Central Florida, Department of Civil, Environmental and Construction Engineering, Orlando, FL 32816-2450, United States
[b] King Saud University, Prince Mohamed Bin Naif Chair for Traffic Safety Research, P.O. Box 800, Riyadh 11421, Saudi Arabia

ARTICLE INFO

Article history:
Received 9 June 2011
Received in revised form 31 March 2012
Accepted 2 April 2012

Keywords:
Real-time crash risk
Loop/radar detectors
Automatic vehicle identification
Reduced visibility
Bayesian matched case-control logistic regression

ABSTRACT

More researchers started using real-time traffic surveillance data, collected from loop/radar detectors (LDs), for proactive crash risk assessment. However, there is a lack of prior studies that investigated the link between real-time traffic data and crash risk of reduced visibility related (VR) crashes. Two issues that have not explicitly been addressed in prior studies are; (1) the possibility of predicting VR crashes using traffic data collected from the Automatic Vehicle Identification (AVI) sensors installed on Expressways and (2) which traffic data are advantageous for predicting VR crashes; LDs or AVIs. Thus, this study attempts to examine the relationships between VR crash risk and real-time traffic data collected from LDs installed on two Freeways in Central Florida (I-4 and I-95) and from AVI sensors installed on two Expressways (SR 408 and SR 417). Also, it investigates which data are better for predicting VR crashes. The approach adopted here involves developing Bayesian matched case-control logistic regression models using the historical crashes, LDs and AVI data. Regarding the model estimated based on LDs data, the average speed observed at the nearest downstream station along with the coefficient of variation in speed observed at the nearest upstream station, all at 5–10 min prior to the crash time, were found to have significant effect on VR crash risk. However, for the model developed based on AVI data, the coefficient of variation in speed observed at the crash segment, at 5–10 min prior to the crash time, affected the likelihood of VR crash occurrence. The results showed that both LDs and AVI systems can be used for safety application (i.e., predicting VR crashes). It was found that up to 73% of VR crashes could be identified correctly. Argument concerning which traffic data (LDs or AVI) are better for predicting VR crashes is also provided and discussed.

1. Introduction

The emphasis in freeway management has been growing toward identifying patterns (i.e., turbulence in the traffic flow) in real-time traffic data that potentially precede traffic crashes on roadways. Additionally, in recent years, there has been a growing emphasis on employing Automatic Vehicle Identification (AVI) data for the provision of real-time travel time information to motorists within Advanced Traveler Information Systems (ATISs) (Dion and Rakha, 2006).

Numerous studies have established statistical links between freeway crash risk and traffic flow characteristics collected from subsurface loop detectors or radar sensors (LDs). These studies include Madanat and Liu (1995), Zhou and Sisiopiku (1997), Oh et al. (2001), Lee et al. (2002, 2003), Golob et al. (2004), Abdel-Aty et al. (2004, 2008), Abdel-Aty and Pande (2005), and Pande and Abdel-Aty (2006).

Obviously, very few studies investigated the relationship between real-time traffic parameters and crash occurrence while controlling for visibility and/or weather conditions. For example, Golob and Recker (2001) examined how the types of freeway accidents are related to both the flow of traffic, weather and ambient lighting conditions. The results indicated

* Corresponding author. Tel.: +1 407 823 5657; fax: +1 407 823 3315.
E-mail address: hhassan@knights.ucf.edu (H.M. Hassan).

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2012.04.001

that median traffic speed and temporal variation in speed in the left and interior lanes are strongly related to the type of collision. Also, when controlling for weather and lighting conditions, the findings suggested that crash severity is influenced more by volume than by speed.

Noticeably, there is a lack of studies that strive to gain a good understanding of the relationship between real-time traffic flow characteristics and crashes occurring under reduced visibility (VR crashes). Although, the percentage of VR crashes is small compared to crashes that occurred at clear visibility conditions (CV crashes), these crashes tend to be more severe and involve multiple vehicles. A recent example of fog-related crashes happened on I-4 in Polk County, Florida in January 2008, resulting in 70 vehicle pile-up. This multi-vehicle crash caused 5 fatalities, many injuries, and shutting down I-4 for extended time.

In this regard, Whiffen et al. (2004) indicated that approximately 700 traffic fatalities occurred in the United States every year due to driving in areas of dense fog. Al-Ghamdi (2007) reported that the injury and death rates of fog related crashes (injuries and deaths per 100 crashes) are 3.75 and 2.25 times the injury and death rates of crashes occurring at normal weather conditions, respectively. The numbers and rates mentioned above show the severity of fog-related crashes. Thus, a real-time assessment of traffic flow characteristics may help in reducing the risk of VR crashes.

The authors had previously explored the occurrence of VR crashes on freeways using real-time traffic surveillance data (speed, volume and occupancy) collected from LDs data (Hassan and Abdel-Aty, 2011). Preliminary results indicated that the average occupancy downstream during 10–15 min prior to the crash coupled with the average speed downstream and upstream 5–10 min before the crash increase the likelihood of VR crash occurrence in between. The present study, however, focuses on developing real-time prediction models of VR crashes on Expressways using traffic data collected from AVI sensors. Although, the AVI system is designed primary for real-time travel time information and tolling purposes, it provides real-time traffic data (Space Mean Speeds) every one minute at stations installed on Expressways.

Thus, the main goal of the current study was to examine the possibility of using AVI data to predict visibility related crash occurrence and to investigate which traffic data (LDs data or AVI data) would achieve better accuracy for predicting visibility related crashes.

It is worth mentioning that there are significant difference in the nature of the collected speed data from LDs and AVI sensors. LDs measure time-mean-speed (TMS), whereas AVIs measure space-mean-speed (SMS). TMS is defined as the arithmetic mean of the speed of vehicles passing a point during a given time interval. On the other hand, SMS is the average speed of all the vehicles traveling a given section of the road over specified time period.

Historical VR crashes and the corresponding traffic surveillance data of LDs were collected from a 75 mile and 137 mile corridors of Intestate-4 and Intestate-95 in Central Florida, respectively, between December 2007 and March 2009. In addition, historical VR crashes and the corresponding AVI traffic data were collected from two Expressways; SR 408 and SR 417 between 2007 and 2009.

Two stratified case-control datasets consisting of traffic data corresponding to every VR crash (case) and five random non-crash cases (controls) were created for both freeways and expressways under investigation. Hence, a binary classification approach may be adopted. Bayesian matched case-control logistic regression models have been estimated to achieve the goals of the present study. The purpose of using this statistical approach was to explore the effects of traffic flow variables on VR crashes while controlling for the effect of other confounding variables such as crash time (e.g., peak or off-peak time, season) and the geometric design elements of highway sections (e.g., horizontal and vertical alignments).

2. Data collection and preparation

2.1. Study area and parameters

Two sets of data were prepared and used in the study; (1) Freeways LDs data and (2) Expressways AVI data. The first dataset was collected from LDs (loop and radar detectors) sensors spaced at approximately 0.5–0.8 mile for about 75 mile and 137 mile corridors of I-4 and I-95 in Central Florida, respectively, between December 2007 and March 2009. These sensors record and archive 30 s aggregation of speed, volume and occupancy.

VR crashes were gathered during the same period and at the same study area. Based on police reports, two criteria for choosing VR crashes from the crash database were considered: weather (fog or rain) and vision obstructed (inclement weather, fog or smoke). According to the crash database maintained by Florida Department of Transportation (FDOT), there were 2984 mainline crashes reported in the same study period and area. All crashes that occurred under the influence of alcohol and drugs were then excluded. Crashes caused by these reasons can occur under any conditions whether the visibility is low or not. Subsequently, a total of 125 VR crashes were extracted. However, due to LDs data availability, only 67 VR crashes that have corresponding traffic flow data, were obtained and used in the analysis.

The second dataset used in this study was collected from AVI sensors spaced at approximately 1–1.5 mile for about 23 and 33 mile of Expressways SR408 and SR417, respectively, for three years 2007–2009. The Orlando-Orange County Expressway Authority (OOCEA) records and archives only 1-min aggregation of space mean speed and the estimated average travel time along the defined road segments.

Again, VR crashes that occurred on these Expressways during the same period were extracted. A total of 1895 mainline crashes occurred in the same study area and period were extracted. Subsequently, a total of 57 VR crashes were obtained. However, only 39 VR crashes that have corresponding traffic flow data were used in the analysis.

2.2. Data preparation

Regarding the first dataset (Freeways LDs data), based on the location of each VR crash, six nearest LDs stations (three stations upstream and three stations downstream) to the crash location were identified using Geographic Information System (GIS) software. As shown in Fig. 1, the first downstream and upstream LDs stations were named DS1 and US1, respectively. The subsequent stations in the downstream direction were labeled DS2 and DS3, respectively. Similarly, the subsequent stations in upstream direction were named US2 and US3, respectively.

Regarding the second dataset (Expressways AVI data), based on the location of each VR crash, the crash segment (the segment in which the VR crash has occurred) in addition to six nearest segments (three segment in the upstream direction and three segment in the downstream direction) to the crash location were identified. Similar to LDs stations, the three upstream segments were named US1, US2 and US3, respectively while; the three downstream segments were named DS1, DS2 and DS3, respectively. The arrangement of LDs stations and AVI segments is shown in Fig. 1.

Traffic data for LDs (specifically time mean speeds) were then extracted for the day of every VR crash as follows; for example, if a VR crash occurred on January, 14, 2008 (Monday) 8:00 am, I-4 eastbound, the traffic data were extracted from three stations upstream and three stations downstream of the crash location from 7:45 am to 7:55 am (10 min window). Subsequently, five random non-crash cases were also determined for the same location and time on different Mondays (in the same season since Central Florida experience two distinct seasons) where no crashes were observed within 1 h of the original crash time. Traffic data was also extracted for these five non-crash cases during the same 10 min window.

The 5-min interval prior to the crash time was disregarded for two main reasons. First, the practical application of the models that have significant variables at 0–5 min prior to the crash time is doubtful. If a crash time is identified correctly there would be no time for the traffic management center to analyze, react or disseminate the relevant warning information to the drivers. The second reason is to avoid any discrepancy about the exact time of crashes which is about ±2 min (Golob and Recker, 2001).

The next step was the aggregation of LDs and AVI data. Regarding LDs data, since the 30-s raw data was noticed to have random noise and are difficult to work with in a modeling framework therefore, the raw data were combined into 5-min level. The extracted raw data were aggregated to different levels of 3 and 5 min to investigate the best level that will achieve better accuracy in crash prediction. The 5-min aggregation level was found to be the best, which is consistent with prior studies (Abdel-Aty et al., 2008; Pande et al., 2005). It is worth noting that 5-min of aggregation of the data are already carried out by most traffic management agencies for the travel time estimation algorithms (Pande et al., 2011). Thus, the 10-min period for which data were collected was then divided into two time slices. The period of 5–10 min before the crash was named as time slice 2 while the period of 10–15 min prior to the crash was labeled as time slice 3. The averages, standard deviations and coefficient of variation in speed (standard deviation/average) were then calculated for each LDs station during time slices 2 and 3.

To sum up, regarding the first dataset, a stratified case-control dataset consisting of LDs traffic data corresponding to every VR crash (case) and five randomly selected matched non-crash cases (controls) was created. Thus, the first dataset includes 402 observations (67 crashes and 335 non-crash cases).

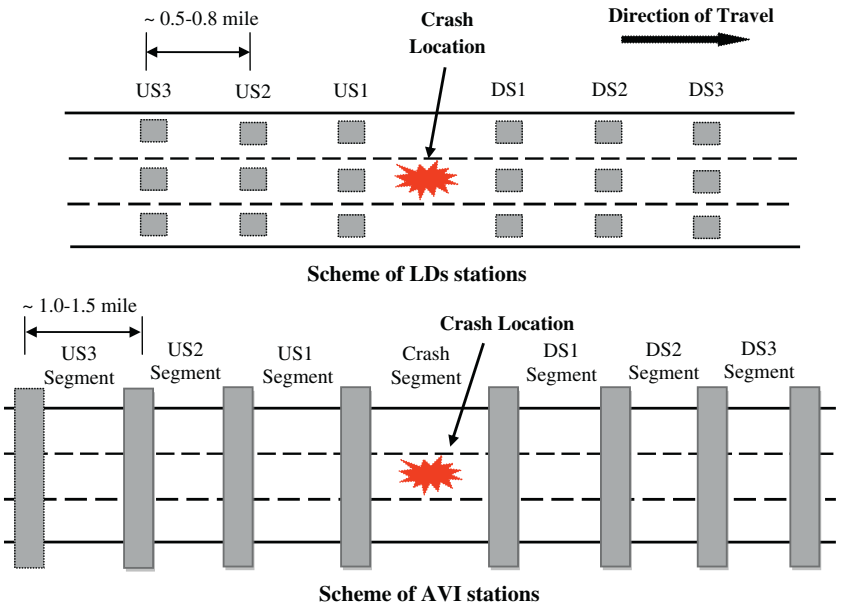

Fig. 1. Arrangement of LDs and AVI stations.

Similarly, traffic data for AVI (space mean speeds data) were extracted for every VR crash that has occurred on Expressways (SR408 and SR417) in addition to five randomly non-crash cases for the same 10 min window mentioned above. These data were extracted for the crash segment and six nearest segments (as shown in Fig. 1). The extracted 1-min space-mean speeds of AVI data were also aggregated into 5-min aggregation level (time slices 2 and 3).

In brief, concerning the second dataset, a stratified case-control dataset consisting of AVI traffic data corresponding to every VR crash (case) and five randomly selected matched non-crash cases (controls) was created. Thus, the second dataset includes 234 observations (39 VR crashes and 195 non-crash cases).

For each of the two datasets, by varying m (No. of controls) from 1 to 5; five datasets have been created which referred to as matched 1:1, 1:2, and 1:5 dataset. Each matched data set (1: m, $m = 1, 2, \ldots, 5$) was analyzed separately. However, no significant differences have been observed when changing m. Therefore, only the detailed description of the analysis of 1:5 matched datasets is presented and discussed.

3. Preliminary analysis of VR crashes

This section presents a preliminary analysis of VR crashes used in this study. Table 1 summarizes the distributions of these crashes for both Freeways (I-4 & I-95) and Expressways (SR417 & SR408) under exploration. Regarding vision obstruction, 4% of the VR crashes have occurred on the freeways under investigation when vision was obstructed by fog while 96% of the VR crashes occurred when vision was obstructed due to heavy rain. In addition, 15% and 85% of the VR crashes extracted for the Expressways have occurred when vision was obstructed by fog and heavy rain, respectively.

Considering lighting conditions, the results revealed that a large percent of the VR crashes on the Freeways and Expressways under study (58.2% and 48.7%, respectively) have occurred during daylight followed by 19.4% and 23.1%, respectively that occurred at night in the absence of street light. Moreover, it was found that about half of the VR crashes, occurred on the Freeways and Expressways under investigation, were rear end crashes (about 48% and 46%, respectively). One possible explanation for this is that at reduced visibility, drivers cannot reduce their speed gradually when they suddenly encounter a relatively higher traffic density, therefore, a crash occurs and most likely rear end. In general, previous studies showed that rear-end crashes represent the highest percent on Freeways and Expressways (Pande et al., 2011; Singh, 2003).

4. Methodology

A flow chart of the overall data analysis process of this study is shown in Fig. 2. The figure shows that LDs data (time-mean speeds data) collected from freeways (I-4 & I-95) was used to predict VR crashes occurrences on Freeways using Bayesian matched case-control logistic regression approach. The final model obtained from this stage was named Model-1. This model was estimated to investigate whether or not one can predict the occurrence of VR crashes using time mean speeds only in the absence of any information regarding volume and occupancy (to be comparable to the case of AVI data).

Subsequently, the freeways LDs data was converted from time-mean speeds into space-mean speeds. This new dataset set was also used to predict VR crash occurrence on Freeways using space-mean speed data. The model was estimated also using Bayesian matched case-control logistic regression approach and labeled Model-2. This dataset is equivalent to AVI data and hence, the results from Model-2 were tested using the AVI expressways data.

It is worth mentioning that Wardrop (1952) derived the relationship between the time-mean speed ($\bar{u}_T$) and space-mean speed ($\bar{u}_S$) as follows:

$$\bar{u}_T = \bar{u}_S + \frac{\sigma_S^2}{\bar{u}_S} \tag{1}$$

Table 1
Distribution of VR crashes.

Factors	Categories	Freeways (I-4&I-95) Percentages	Expressways (SR417&SR408) Percentages
Roadways	I-4/SR417	58.2 (I-4)	43.6 (SR417)
	I-95/SR408	41.8 (I-95)	56.4 (SR408)
Vision obstruction	Fog	6.0	15.0
	Heavy rain	94.0	85.0
Lighting conditions	Daylight	58.2	48.7
	Dusk	4.5	7.7
	Dawn	6.0	5.1
	Dark (street light)	10.5	12.8
	Dark (no street light)	19.4	23.1
	Unknown	1.4	2.6
Crash type	Rear end	47.8	46.1
	Angle	17.9	10.7
	Sideswipe	7.6	30.2
	Others	7.4	13.0

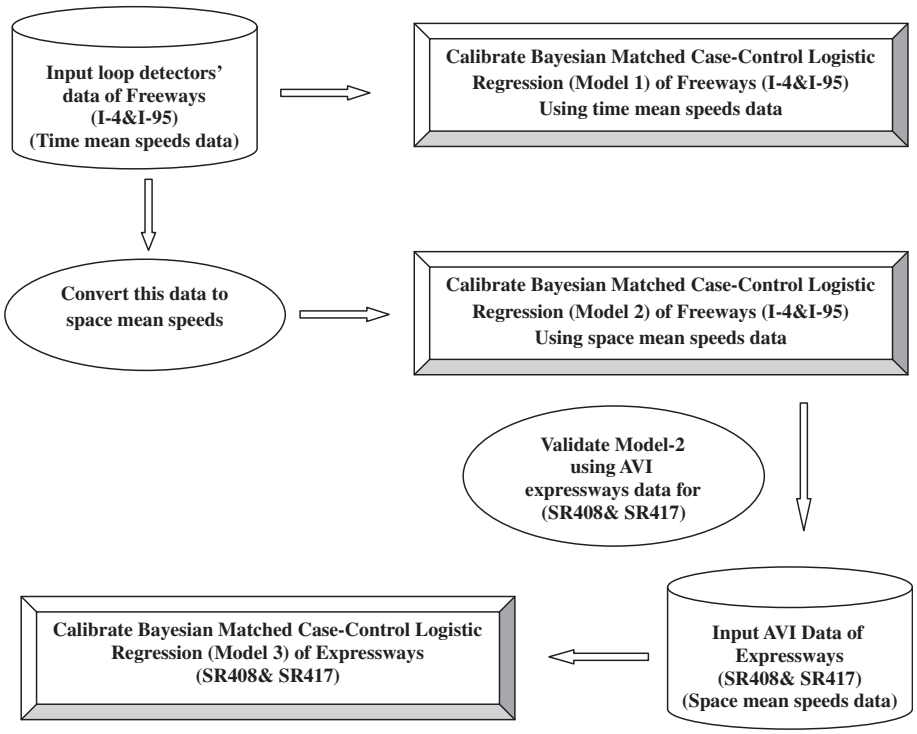

Fig. 2. Flow chart representing the overall data analysis.

where σ_S^2 is the variance in vehicle speeds about the space-mean speed. Rakha and Zhang (2005) indicated that this formulation estimates the time-mean speed from the space-mean speed, which is typically the reverse of what is required (as σ_S^2 is unknown). Therefore, they derived a modified relationship between $\bar{u}_S$ and $\bar{u}_T$ as follows:

$$\bar{u}_T = \bar{u}_S \cdot \left[1 + \frac{\sigma_T^2}{\bar{u}_T \cdot \bar{u}_T}\right] \cong \bar{u}_S + \frac{\sigma_T^2}{\bar{u}_S} \tag{2}$$

where σ_T^2 is the variance in vehicle speeds about the time-mean speed. They also demonstrated that the proposed formulation, which utilizes the variance about the time-mean speed as opposed to the variance about the space-mean speed, produces an estimate error to within 0–1%. Eq. (2) was used in the present study to estimate space-mean speeds from time-mean speeds of LDs data.

Next, AVI data (space-mean speeds data) collected from Expressways (SR408 & SR417) was used to predict the occurrences of VR crashes on Expressways. The developed Bayesian matched case-control from this step was named Model-3. A discussion and comparison between the results of the three developed models in this study is provided in the following sections. In the next section, the statistical approach used in this study is described in detail.

5. Matched crash non-crash analysis

As mentioned earlier, the purpose of the proposed matched crash-non-crash analysis is to explore the effects of traffic flow variables on VR crashes while controlling for the effects of other confounding variables such as crash time (e.g., peak or off-peak hours, season) and the geometric design elements of freeway/expressway sections (e.g., horizontal, vertical alignments, on-ramp and off-ramp vicinity locations, etc.). Matched case-control logistic regression using classical statistic approach has been adopted in epidemiological studies. In addition, it was used in few transportation related studies such as Abdel-Aty et al. (2004) and Hassan and Abdel-Aty (2011). In this study, Bayesian matched case-control logistic regression approach has been adopted. A brief description of this modeling technique is provided here.

In a matched crash non-crash study, crashes are selected first. Then, for each selected crash, some non traffic flow variables associated with each crash are selected as matching factors such as location, day of the week, time of day, and season. Using these matching factors, a total of m non-crash cases are then selected randomly from each subpopulation of non-crash cases. For example, for a given crash, a subpopulation of non-crash cases consist of observations on traffic flow variables obtained from the same LDs/AVI at the same time of the day, same day of the week of crashes but over all other days, are recorded.

The $(m + 1)$ observations of all traffic variables for VR crashes and non-crash cases form one stratum. Within stratum, differences between VR crashes and non-crash cases regarding flow characteristic are utilized in the development of the statistical model. This procedure is conducted under the conditional likelihood of statistical theory.

Assume that there are N strata with 1 crash and m non-crash cases in stratum j, where $j = 1, 2, 3, \ldots, N$. The probability of any observation in a stratum being a crash might be modeled by the following linear logistic regression model:

$$Logit\{P_j(X_{ij})\} = \alpha_j + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij} \tag{3}$$

where $P_j(X_{ij})$ is the probability that the ith observation in the jth stratum being a crash; $X_{ij} = (X_{1ij}, X_{2ij}, \ldots, X_{kij})$ is the vector of k traffic flow variables; $i = 0, 1, 2, \ldots, m$ and $j = 0, 1, 2, \ldots, N$.

It is to be noted that the intercept term α in Eq. (3) summarizes the effect of variables used to form strata on the crash probability and would be different across strata. A conditional likelihood is constructed to take account of the stratification in the analysis. This conditional likelihood function $L(\beta)$ is independent of the intercept terms $\alpha_1, \alpha_2, \ldots, \alpha_N$ and hence, the effects of matching variables cannot be estimated. Therefore, crash probabilities cannot be estimated using Eq. (3). However, the values of β parameters that maximize the conditional likelihood function are also the estimates of β coefficient in Eq. (3). These estimates are log odds ratio and may be used to represent the relative risk of a VR crash.

These relative risks (named as hazard ratio in SAS) are given using SAS procedure *PHREG*. Consider two observation vectors $X_{1j} = (X_{11j}, X_{21j}, X_{31j}, \ldots, X_{k1j})$ and $X_{2j} = (X_{12j,\ X22j}, X_{32j}, \ldots, X_{k2j})$ from the jth strata on the k traffic flow variables. Thus, by substituting the two observation vectors X_{1j} and X_{2j} in Eq. (3), the log odds ratio of VR crash occurrence due to traffic flow vector X_{1j} relative to traffic flow vector X_{2j} will have the following form:

$$\log\left\{\frac{P(X_{1j})/[1 - P(X_{1j})]}{P(X_{2j})/[1 - P(X_{2j})]}\right\} = \sum_{i=1}^{k} \beta_i (X_{i1j} - X_{i2j}) \tag{4}$$

The right hand side of Eq. (4) is independent of α_j and can be calculated using the estimated β coefficients. Thus, the above relative log odds ratio (left hand side of Eq. (4)) may be utilized for predicting VR crashes by replacing X_{2j} with the vector of values of the traffic flow variables in the jth stratum of non-crash cases. One may use simple average of all non-crash observations within the stratum for each variable. Let $\overline{X}_{2j} = (\overline{X}_{12j}, \overline{X}_{22j}, \overline{X}_{32j}, \ldots, \overline{X}_{k2j})$ denote the vector of mean values of non-crash cases of the k variables within the jth stratum. Then the log odds ratio of VR crash relative to non-crash cases may be approximated by:

$$\log\left\{\frac{P(X_{1j})/[1 - P(X_{1j})]}{P(X_{2j})/[1 - P(X_{2j})]}\right\} = \sum_{i=1}^{k} \beta_i (X_{i1j} - \overline{X}_{i2j}) \tag{5}$$

Therefore, log odds ratio in Eq. (5) can be used for predicting VR crashes by establishing a threshold value that achieve the desirable crash classification accuracy.

In this study, Bayesian matched case-control logistic regression approach was adopted using SAS package 9.2, procedure PHREG. This procedure provides Bayesian analysis in addition to the standard (classical) analysis. Procedure PHREG generates a chain of posterior distribution samples by the Gibbs Sampler and provides summary statistics, convergence diagnostics and diagnostic plots for each parameter. It also uses the adaptive rejection sampling (ARS) algorithm to sample parameters sequentially from their univariate full conditional distribution (SAS Institute Inc., 2009).

The advantages of using the Bayesian approach include that (1) it provides a natural and principled way of combining prior information (if it exists) with the data, within a solid theoretical decision framework to yield a posterior belief (when new data become available, the previous posterior distribution can be used as a prior), (2) it presents full distributional profile of parameters rather than single coefficient estimates to fully account for the uncertainty associated with single parameter estimates in classical statistics, and (3) it gives inferences that are exact and conditional on the data, without reliance on asymptotic approximation and hence, small sample inference proceeds in the same manner of a large sample (Rao, 2003; SAS Institute Inc., 2009).

It is worth noting that measurement error is defined as the variation of the observed measurement from the true value. It is composed of two error components; random error and systematic error (Espino-Hernandez et al., 2011). The Bayesian approach provides a rigorous probabilistic framework to account for most sources of errors (i.e., measurement error) that contribute to the uncertainty of the updated parameters). Within a Bayesian framework, the estimation of relative risks linking parameters of interest (X) and dependent variable (Y) takes all the uncertainty of X into consideration as the prediction is achieved by averaging the posterior probability density function over all possible values of the measurement errors (Zhang et al., 2011).

Due to the absence of informative priors, a uniform prior distribution was assumed and used to estimate the first two models developed in this study. The uniform prior is a flat prior which assigns equal likelihood on all possible values of the parameter. However, the third model presented in this study were estimated twice (using uniform prior and using the results of Model-2 as informative priors) as explained in the following sections. The convergence of the generated Markov chains of all developed models was assessed by examining the trace plot, the autocorrelation function plot and the posterior density plot. It was found that, all the models have converged reasonably. The DIC, a Bayesian generalization of AIC, is

used along with the classification accuracy of the three models to measure the models complexity and fit (Spiegelhalter et al., 2003).

6. Predicting VR crashes on freeways using LDs data

6.1. Using time-mean speed data

As indicated earlier, to predict the real-time crash risk of VR crashes on Freeways (I-4 and I-95), the first dataset was used. The first dataset includes 402 observations (67 VR crashes and 335 non-crash cases). Automatic search technique: stepwise, forward and backward were used to identify significant variables. All three search techniques resulted in two significant variables. The estimates of beta coefficients, credible interval, associated summary results; model fit statistics and classification results of actual and predicted VR Crashes obtained from the final model (Model-1) are presented in Table 2.

The results indicated that a decrease in the average speed at the nearest downstream station (ASDS1_2, $\beta = -0.1409$, 95% CI (-0.2010, -0.0898)) coupled with an increase in the logarithm of coefficient of variation in speed (Standard deviation/average speed) at the nearest upstream station (Log. CSUS1_2, $\beta = 0.3979$, 95% CI (0.0671, 0.8536)), all at time slice 2 (5–10 min before the crash time) increase the risk of VR crash occurrence in between. The results from the model may imply that lower average speed at the nearest downstream station (possible due to higher occupancy) coupled with higher standard deviation in speed at the nearest upstream station, all at time slice 2 pointing to potential queue formation under turbulent speed conditions, which could be a cause for high VR crash possibility.

Note that the hazard ratio corresponding to parameter estimates are shown in Table 2. Hazard ratio, equals the exponent of the beta coefficient, is an estimate of the expected change in the risk ratio of having a VR crash versus non-crash cases per unit change in the corresponding factor. For example, hazard ratio of 1.53 corresponding to (Log. CSUS1_2) means that the risk of a VR crash increases about 1.5 times for each unit increase in (Log. CSUS1_2).

As previously explained, the odds ratio in Eq. (5) can be used to classify VR crash and non-crash cases. Therefore, the mean of the two significant variables of all five non-crash cases within each matched set were estimated. The vector X_{2j} in Eq. (5) was then replaced by the vector of non-crash means for the jth matched set. The odds ratio for each observation in the data set was then estimated by substituting the beta coefficient from Table 2 in Eq. (5) where the vector X_{1j} is the actual observation in the data set. A threshold value for these ratios was then set to determine whether the location has to be flagged as a potential "VR crash". After investigating all possible thresholds, it was found that using a threshold of 1.0 for the log odds ratio, over 73% crash identification was achieved (as shown in Table 2). The table shows that the sensitivity, proportion of VR crashes that are correctly identified as VR crashes by the model is 73.13%. Also, the specificity, proportion of non-crashes that are correctly identified as non-crashes by the model is 60.30% (Agresti, 2002).

Table 2
Results of Bayesian matched case-control logistic regression (Model 1) (based on LDs data; time-mean speeds).

Parameter	Mean	Standard deviation	Credible interval	
			2.5%	97.5%
Parameters estimates				
ASDS1_2	−0.1409	0.0283	−0.2010	−0.0898
Log. CSUS1_2	0.3979	0.2350	0.0671	0.8536
Hazard ratios				
ASDS1_2	0.8689	0.0245	0.8179	0.9141
Log. CSUS1_2	1.5304	0.3659	0.9351	2.3481
Model fit statistics				
DIC				143.088
pD (effective number of parameters)				1.989
		Predicted Y		
	Frequency percent Row percent Col percent	0	1	Total
Classification results of actual and predicted VR crashes				
Actual Y	0	202 50.25 60.30 91.82	133 33.08 39.70 73.08	335 83.33
	1	18 4.48 26.87 8.18	49 12.19 73.13 26.92	67 16.67
	Total	220 54.73	182 45.27	402 100.00

It is worth mentioning that this threshold may be changed to achieve desirable classification accuracy for both crashes and non-crash cases. In other words, accuracy can be easily increased by accepting higher false alarm rate and be on the conservative side. If freeway traffic turbulence is identified, even if does not lead to a crash, it would be useful to reduce turbulence and improve flow. This point could be left to implementation and the preferences of the specific traffic agency. To sum up, the predictive power of the model might be evaluated using the rate of crash misclassification or overall misclassification or some combination of the two.

6.2. Using space-mean speeds' data

As discussed previously, the first dataset (freeways LDs data) was converted from time-mean speeds into space-mean speeds. This step was done for two mean reasons. First, to calibrate a prediction model for VR crashes using a dataset that is equivalent to AVI data (named Model-2) and therefore, it might be possible to compare between the results of Model-2 and Model-3 (Expressways' VR crashes prediction model based on AVI data). Second, the results of Model-2 may be tested using the Expressways' AVI data.

Table 3 shows the results of the Bayesian matched case-logistic regression (Model-2) that was estimated based on Freeways' LDs data (space-mean speeds). As expected, similar to the results of Model-1, the results of Model-2 revealed that the average speed at the nearest downstream station (ASDS1_2, $\beta = -0.1573$, 95%CI (-0.2253, -0.0984)) and the logarithm of coefficient of variation in speed at the nearest upstream station (Log. CSUS1_2, $\beta = 0.4434$, 95%CI (0.0926, 0.9775)), all at time slice 2 (5–10 min before the crash time) were found to have significant effect on VR crash risk on Freeways. As shown in Table 3, using a threshold of 1.0 for the log odds ratio, over 71% crash identification was achieved. Considering the results shown in Tables 2 and 3, the results indicate that Model-1 (based on time-mean speeds) is slightly better that model 2 (based on space-mean speeds) as it achieved higher classification accuracy of identifying VR crashes (73.13%) and better fit statistic (DIC = 143.088 compared to DIC = 156.733 of Model-2).

Then, using expressways AVI data (234 observations; 39 VR crashes and 335 non-crash cases) the results of Model-2 were tested. It was found that about 64.6% and 63.1% of VR crashes and non-crash cases, respectively, were correctly identified. It can be noted that this classification accuracy (64.6%) is relatively comparable to the accuracy 71.64% obtained previously by Model-2 which may imply that Model-2 is performing well in correctly predicting the occurrences of VR crashes. One possible explanation for having relatively lower classification accuracy when using the tested dataset is the differences between LDs and AVI arrangements (configurations). As shown in Fig. 1, LDs sensors are spaced at approximately 0.5–0.8 mile compared to AVI sensors that are spaced at approximately 1.0–1.5 mile.

Table 3
Results of Bayesian matched case-control logistic regression (Model 2) (based on LDs data; space-mean speeds).

Parameter	Mean	Standard deviation	Credible interval	
			2.5%	97.5%
Parameters estimates				
ASDS1_2	−0.1573	0.0322	−0.2253	−0.0984
Log. CSUS1_2	0.4434	0.2729	0.0926	0.9775
Hazard ratios				
ASDS1_2	0.8549	0.0274	0.7983	0.9063
Log. CSUS1_2	1.6174	0.4533	0.9116	2.6578
Model fit statistics				
DIC				156.733
pD (effective number of parameters)				1.986

		Predicted *Y*		
	Frequency percent Row percent Col percent	0	1	Total
Classification results of actual and predicted VR crashes				
Actual *Y*	0	177 44.03 52.84 90.31	158 39.30 47.16 76.70	335 83.33
	1	19 4.73 28.36 9.69	48 11.94 71.64 23.30	67 16.67
	Total	196 48.76	206 51.24	402 100.00

7. Predicting VR crashes on expressways using AVI data

An issue that has not been addressed in prior studies is the possibility of predicting the occurrence of VR crashes using traffic data collected from AVI sensors installed on Expressways. Therefore, using space-mean speeds data collected from Expressways SR408 and SR408 for a total of 39 VR crashes and 195 non-crash cases, a Bayesian matched case-control logistic regression model was estimated (Model-3). Table 4 shows the parameter estimate, hazard ratio, goodness of fit indices and classification accuracy of Model-3. The results revealed that the logarithm of coefficient of variation in speed ($\beta = 0.7588$, 95%CI (0.3489, 1.2062)) at the crash segment (see Fig. 1) during time slice 2 (5–10 min prior to crash time) was found to have a significant effect on VR crash risk. These results imply that lower average speed observed at a certain segment coupled with higher standard deviation in speeds at the same segment; all at time slice 2, increase the probability of VR crashes occurrences.

One may wonder if both average speed and standard deviation are significant predictors when used separately instead of combining them into coefficient of variation (standard deviation/average speed) as one variable. To address this issue, we estimated another model using these two variables however; this model showed lower accuracy in identifying VR crashes correctly. Also it showed higher DIC than the model that has Log. CSC_2 and thus we concluded that the best model is the one that has only (Log. CSC_2). No variables from the upstream or downstream segments were found significant. This should not be surprising since reduced visibility due to fog/smoke or heavy rain is most likely localized. As indicated earlier, the lengths of AVI segment vary from about 1.0–1.5 mile, so it is logical to get significant variable(s) from the crash segment only.

As shown in Table 4, a hazard ratio of 2.19 corresponding to (Log. CSC_2) means that the risk of a VR crash increases about 2.2 times for each unit increase in (Log. CSC_2). Also the table shows that the sensitivity and the specificity of the model are 69.23% and 61.03%, respectively. As discussed earlier, due to the absence of informative priors, all the three models presented in the present study were estimated using uniform prior which is favored by many statisticians (SAS Institute Inc., 2009). However, it is worth mentioning that we re-estimated Model-3 using the results of Model-2 as informative priors (specifically, Log. coefficient of variation in speeds). Note that the datasets used to develop Model-2 and Model-3 is comparable as both of them are space-mean speeds data. It was found that the results of Model-3 had not significantly improved when using the informative priors possibly because the configurations of LDs and AVI are different. The LDs stations are spaced approximately at 0.5–0.8 mile while the lengths of AVI segments vary from 1–1.5 miles. Also, LDs have upstream and downstream stations only while, AVI has crash segment in addition to the upstream and downstream segments. The results imply that it may not be advisable to use informative priors from other corridors that probably have different characteristics. Therefore, the results based on uniform prior of Model-3 are only presented here.

It is worth noting that several agencies were contacted to obtain historical visibility measurements at the same period and study area. The aim was to confirm that the VR crashes under investigation did indeed occur in reduced visibility conditions and to determine non-crash cases at reduced visibility. Among the agencies contacted, it was found that National

Table 4
Results of Bayesian matched case-control logistic regression (Model 3) (based on AVI data; space mean speeds).

Parameter	Mean	Standard deviation	Credible interval	
			2.5%	97.5%
Parameters estimates				
Log. CSC_2	0.7588	0.2177	0.3489	1.2062
Hazard ratios				
Log. CSC_2	2.1877	0.4943	1.4174	3.3406
Model fit statistics				
DIC				91.122
pD (effective number of parameters)				0.990
		Predicted *Y*		
	Frequency percent Row percent Col percent	0	1	Total
Classification results of actual and predicted VR crashes				
Actual *Y*	0	119 50.85 61.03 90.84	76 32.48 38.97 73.79	195 83.33
	1	12 5.13 30.77 9.16	27 11.54 69.23 26.21	39 16.67
	Total	131 55.98	103 44.02	234 100.00

Climate Data Center (NCDC) provides the historical visibility data. NCDC's website provides access to their database that consists of hourly weather data for many stations across the United States. Visibility measurements for the same period and study area were successfully obtained for six airport weather stations surrounding the study area: Daytona Beach, Orlando Sanford, Orlando Kissimmee, Orlando Executive, Orlando International, and Melbourne.

The average visibility measurements obtained from the two closest weather stations to every visibility related crash location were estimated for all crashes and the corresponding non-crash cases. The closest stations to every visibility related crash were identified using GIS software. A threshold of 250 m (about 820 ft.) was selected as the criterion for determining non-crash cases at reduced visibility (Rockwell, 1997). Therefore, non-crash cases at reduced visibility were considered if the corresponding average visibility measurement obtained from the two closest weather stations to the crash location was 250 m or less. By this, we confirmed that the VR crashes used in this study occurred under reduced visibility conditions.

Consequently, two stratified case-control datasets consisting of traffic data corresponding to every VR crash (case) and three random non-crash cases (controls) both under reduced visibility conditions were created for both LDs and AVI data. Subsequently, two models were estimated for these two datasets. Regarding the model estimated based on LDs data, the results showed that an increase in the average speed observed at the nearest upstream station along with an increase in the standard deviation in speed observed at the nearest downstream station, all at 5–10 min prior to the crash time, were found to have significant effect on VR crash risk. However, for the model developed based on AVI data, the findings indicated that an increase in the standard deviation in speed observed at the crash segment, at 5–10 min prior to the crash time, affected the likelihood of VR crash occurrence. These two models developed based on LDs and AVI datasets showed lower prediction accuracy of VR crashes (63% and 60%, respectively) compared to the models presented earlier in this paper (which were estimated based on comparing VR crashes to non-crash cases at normal visibility conditions). One possible explanation is that when comparing crashes to non-crash cases both in poor visibility conditions using speed data only, less variation in speed between crashes and non-crash-cases should be expected and this may explain the reason of having lower prediction accuracy compared to the models presented in this paper. Therefore, in this study, the results of comparing VR crashes to non-crash cases at normal visibility conditions were only presented and discussed.

It is worth mentioning that we did not include correlated factors in the same model. For example, no two speed variables, from the same loop detector station were used in the same model, but speed upstream and downstream would be acceptable as there is a gap in between (0.5–0.8 mile). In addition, Pearson correlation was checked and the results confirmed that there was no collinearity problem (i.e., the correlation between predictors used in the models presented in this paper ranges from 0.05 to 0.2). Also, it was found that the correlation between crashes (target variable) and traffic data (predictors) used in the final three models presented here ranges from 0.7 to 0.9.

8. Conclusions

This study aimed at identifying patterns (i.e., turbulence in traffic flow) in the expressway AVI traffic data that potentially precede visibility related (VR) crashes. Also, it investigated which traffic data is advantageous for predicting VR crashes; data collected from LDs sensors installed on freeways or data collected from AVI sensors installed on expressways. Statistical links between turbulent traffic conditions and VR crash occurrences were established through a detailed analysis of LDs/AVI traffic data corresponding to VR crashes that occurred on freeways (I-4 and I-95) and on expressways (SR408 and SR417) in Central Florida during the study time.

The approach adopted in this study involves developing Bayesian matched case-control logistic regression models using the historical crash, LDs and AVI data. The purpose of adopting this statistical approach was to explore the effects of traffic flow variables on VR crashes while controlling for the effects of other confounding variables such as crash time and the geometric design elements of freeway/expressway sections. To achieve the objectives of the present study, three models were estimated and discussed.

Historical VR crashes along with traffic data (time-mean speeds) collected from LDs on freeways were used to calibrate the first model (Model-1). The second model (Model-2) was calibrated using the same data but after converting it into space-mean speeds (to make it equivalent to AVI data). The results of both models indicated that the average speed observed at the nearest downstream station coupled with the coefficient of variation in speed observed at the nearest upstream station, all at 5–10 min prior to the crash time, were found to have significant effect on VR crash risk. It has been shown that Model-1 and Model-2 achieved over 73% and 71% of VR crash identification, respectively. The performance of model-2 was then tested using historical VR crashes and AVI traffic data (space-mean speeds) collected from expressway (SR417 and SR408). It was found that about 65% of VR crashes were correctly identified. It can be noted that this classification accuracy is relatively comparable to the accuracy 71.64% obtained previously by Model-2 which may imply that Model-2 is performing well in correctly predicting the occurrences of VR crashes, however, one possible explanation for obtaining relatively lower classification accuracy when using the tested dataset is the differences between LDs and AVI arrangements (configurations). LDs sensors are spaced at approximately 0.5–0.8 mile compared to AVI sensors that are spaced at approximately 1.0–1.5 mile and hence, AVI data and LDs data may not match exactly.

Also historical VR crashes and space-mean speeds data collected from AVI sensors located on expressways (SR417 and SR408) were used for developing prediction model of VR crashes on expressways (Model-3). The results of the model revealed that an increase in the coefficient of variation in speed at the crash segment, 5–10 min before the crash time increases the likelihood of VR crashes. No variables from the upstream or downstream AVI segments were found significant possibly

because the effect of fog/smoke or heavy rain is most likely localized and the longer Expressway segments. Model-3 achieved over 69% of VR crash identification.

One objective of this study was to investigate which data (LDs or AVI) are advantageous for predicting VR crashes. Considering the results of Model-3 and compared to the results of Model-1 and Model-2, it can be realized that LDs data is working slightly better than AVI data regarding the prediction of VR crashes possibly due to three reasons. First, the configuration (arrangement) of LDs and AVI sensors is different as discussed above (i.e., the distances between LDs sensors are less than the lengths of AVI segments). Second, AVI measures space-mean speeds by tracking the speed of vehicles through successive AVI sensors while, LDs measures time-mean speed (spot speeds) of vehicles at certain point (LDs stations) on a roadway. Third, the AVI sensors can only record and archive traffic data for vehicles that have AVI tags (i.e., transponders, E-pass, etc.). It is well established that about 80% of vehicles using expressways have AVI tags. On the other hand, LDs record and archive traffic flow data for all vehicles traveling on the roadway. The findings from this study led us to infer that it may be better to develop VR crash risk assessment models based on LDs traffic data. However, the main disadvantage of LDs is that it sometimes fails due to sudden hardware problems which may lead to large missing data. In this case, using AVI or Radar data might be a good alternative for predicting VR crashes.

The results of the present study shed light on the possibility of using LDs and AVI traffic data for predicting VR crashes on freeways/expressways. However, further validation with larger samples might be needed. Finally, one should remember that both systems are installed without safety predictive application in mind. In other words, the results of our work indicate that both systems could be used for safety applications, although there is room for improvement in the system (e.g., shorten AVI segments). Given that most roadways will have either systems, this study showed that risk predictive models could be implemented in both cases.

Acknowledgements

The authors wish to thank Florida Department of Transportation, District 5, for funding this research. The authors are also thankful to the three anonymous reviewers for their valuable comments that helped in revising and significantly improving the paper. All opinions and results are those of the authors.

References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. Safety Research 36, 97–108.
Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., Hsia, L., 2004. Predicting freeway crashes based on loop detector data using matched case–control logistic regression. Transportation Research Record 1897, 88–95.
Abdel-Aty, M., Pande, A., Das, A., Knibbe, W., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transportation Research Record 2083, 153–161.
Agresti, A., 2002. Categorical Data Analysis, second ed. John Wiley & Sons, Inc., Hoboken, New Jersey.
Al-Ghamdi, A., 2007. Experimental evaluation of fog warning system. Accident Analysis and Prevention 39, 1065–1072.
Dion, F., Rakha, H., 2006. Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. Transportation Research Part B 40, 745–766.
Espino-Hernandez, G., Gustafson, P., Burstyn, I., 2011. Bayesian adjustment for measurement error in continuous exposures in an individually matched case-control study. BMC Medical Research Methodology 11, 67.
Golob, T., Recker, W., 2001. Relationships among urban freeway accidents, traffic flow, weather and lighting conditions. California PATH program, institute of transportation studies, Berkeley, ISSN 1055-1417, UCB-ITS-PWP-19.
Golob, T., Recker, W., Alvarez, V., 2004. Freeway safety as a function of traffic flow. Accident Analysis and Prevention 36, 933–946.
Hassan, H., Abdel-Aty, M., 2011. Exploring visibility related crashes on freeways based on real-time traffic flow data. In: Transportation Research Board, 90th Annual Meeting, Washington DC, paper no. 11-0920.
Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. Transportation Research Record 1784, 1–8.
Lee, C., Saccomanno, F., Hellinga, B., 2003. Real-time crash prediction model for the application to crash prevention in freeway traffic. Transportation Research Record 1840, 68–77.
Madanat, S., Liu, P., 1995. A prototype system for real-time incident likelihood prediction. IDEA project final report (ITS-2). Transportation Research Board, National Research Council, Washington, DC.
Oh, C., Oh, J., Ritchie, S., Change, M., 2001. Real time estimation of freeway accident likelihood. In: Transportation Research Board, 80th Annual Meeting. Washington DC.
Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accident Analysis and Prevention 38, 936–948.
Pande, A., Abdel-Aty, M., Hsia, L., 2005. Spatiotemporal variation of risk preceding crashes on freeways. Transportation Research Record 1908, 26–36.
Pande, A., Das, A., Abdel-Aty, M., Hassan, H., 2011. Estimation of real-time crash risk. Are All Freeways Created Equal? Transportation Research Record 2237, 60–66.
Rakha, H., Zhang, W., 2005. Estimating traffic stream space mean speed and reliability from dual- and single-loop detectors. Transportation Research Record 1925, 38–47.
Rao, J., 2003. Small Area Estimation. Wiley, New York.
Rockwell Transportation Systems, Anaheim, CA, 1997. Adverse Visibility Information System Evaluation (ADVISE), System Design Document. Prepared for Utah Department of Transportation, January 14, 1997.
SAS Institute Inc., 2009. SAS/STAT ® 9.2. User's Guide, second ed. SAS Institute Inc., Cary, NC.
Singh, S., 2003. Driver attributes and rear-end crash involvement propensity. NHTSA Technical Report, DOT HS 809 540 <http://www-nrd.nhtsa.dot.gov/Pubs/809-540.pdf>, (accessed January 2012).
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, V.D., 2003. Bayesian measures of model complexity and fit. Royal Statistical Society B 64 (4), 583–616.
Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers 1–2, 325–378.
Whiffen, B., Delannoy, P., Siok, S., 2004. Fog impact on road transportation and mitigation options. In: National Highway Visibility Conference, Madison, Wisconsin, pp. 18–19.
Zhang, E., Feissel, P., Antoni, J., 2011. A comprehensive Bayesian approach for model updating and quantification of modeling errors. Probabilistic Engineering Mechanics 26, 550–560.
Zhou, M., Sisiopiku, V., 1997. Relationship between volume-to-capacity ratios and accident rates. Transportation Research Record 1581, 47–52.