Accident Analysis and Prevention 125 (2019) 70–78

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Exploring crash mechanisms with microscopic traffic flow variables: A hybrid approach with latent class logit and path analysis models☆

Rongjie Yu[a], Yin Zheng[a,*], Mohamed Abdel-Aty[b], Zhen Gao[c]

[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, 4800 Cao'an Road, 201804, Shanghai, China
[b] Department of Civil, Environmental and Construction Engineering, University of Central Florida Orlando, FL, 32826-2450, United States
[c] College of Software Engineering, Tongji University, 4800 Cao'an Road, 201804, Shanghai, China

ARTICLE INFO

Keywords:
Crash mechanism
Microscopic traffic flow variables
Heterogeneous effects
Latent class logit
Path analysis

ABSTRACT

Understanding the occurrence mechanisms of crashes is critical for traffic safety improvement. Efforts have been investigated to reveal the crash mechanisms and analyze the contributing factors from the aspects of vehicle, driver, and operational perspectives. In this study, special attention has been paid to the operational level analyses while bridging the research gaps of: (1) failing to identify the heterogeneous impact of microscopic traffic flow variables on crash occurrence, and (2) focusing on correlation effects without further investigations for the causal relationships. A hybrid modeling approach with latent class logit (LCL) and path analysis (PA) models was proposed to account for the heterogeneous influencing effects and reveal the causal relationships between crash occurrence and microscopic traffic flow variables. Data from Shanghai urban expressway system were utilized for the empirical analyses. First, the LCL model has concluded four latent subsets of crash occurrence influencing factors. Then, PA models were conducted to identify the concurrent relationships (direct and indirect eff ;ects) for the four sets of crash occurrence influencing factors separately. Finally, the results of the LCL model and PA models were compared and the crash-prone scenarios were inferred. And the potential safety improvement countermeasures were discussed.

1. Introduction

Traffic safety has become the most important issue in the transportation field worldwide. Statistics showed that, each year, about 1.24 million people died due to traffic crashes (World Health Organization, 2013). In order to improve traffic safety, it is essential to understand the crash mechanisms and their contributing factors. Currently, the crash contributing factors were mainly analyzed at three levels: (1) vehicle level, where the kinematic features prior to crash occurrence are identified and further utilized for developing advanced driving assistance systems (ADAS); (2) investigations have been conducted for the driver level to identify unsafe driving behaviors and understand their features, and thus safety improvements programs have been further designed, such as the behavior based safety (BBS) coaching (Piccinini et al., 2017); while (3) at operational level, researchers have dedicated to establish the relationships between crash occurrence and microscopic traffic flow variables, and further improve safety using proactive traffic control strategies such as variable speed limit (VSL) (Yu and Abdel-Aty, 2014). In this study, special attention has been paid to explore the crash mechanisms at the operational level.

In order to gain in-depth understanding of the operating conditions prior to crash occurrence, emerging studies such as real-time crash risk analyses have been conducted based on the microscopic traffic sensing data (Roshandel et al., 2015). Different microscopic traffic flow data sources have been utilized, including loop detector data (Kwak and Kho, 2016), probe vehicle data (Park and Haghani, 2016), and automatic vehicle identification (AVI) data (Abdel-Aty et al., 2012). Meanwhile, various modeling techniques have been employed, which includes both statistical models (Lee et al., 2003; Abdel-Aty et al., 2004) and machine learning methods (Pande and Abdel-Aty, 2006; Shi and Abdel-Aty, 2015). The former approach tries to identify the relationships between crash occurrence and contributing factors mostly based on linear correlation assumptions, while the latter one focuses on obtaining the hidden correlations and potential non-linear relationships (Mannering and Bhat, 2014). More recently, the real-time crash risk analysis studies have been focusing on solving the unobserved heterogeneity issues (Mannering et al., 2016) and further understanding the complex impacts of microscopic traffic flow variables on crash

☆ This paper has been handled by associate editor Tony Sze.

* Corresponding author.
E-mail addresses: yurongjie@tongji.edu.cn (R. Yu), 1021560261@qq.com (Y. Zheng), M.Aty@ucf.edu (M. Abdel-Aty), gaozhen@tongji.edu.cn (Z. Gao).

https://doi.org/10.1016/j.aap.2019.01.022
Received 29 September 2018; Received in revised form 28 December 2018; Accepted 19 January 2019
Available online 04 February 2019
0001-4575/

occurrence.

However, the real-time crash risk analyses could only shed lights on the correlation relationships between microscopic traffic flow variables and crash occurrence. For instance, Yu et al. (2016) found that both average speed and traffic volume had statistically significant negative impacts on crash likelihood. Lee et al. (2003) concluded that crash events were outcomes of complex interactions among various contributing factors. Furthermore, Gargoum and El-Basyouny (2016) found that traffic volume holds a mediated effect on crash occurrence through average speed in addition to the direct positive correlation. In addition, from the safety improvement perspective, understanding how the microscopic traffic flow variables affect crash occurrence is critical. For instance, based on the generalized linear models, both traffic volume and operation speed were proved to have substantial influences on crash occurrence, and it is hard to choose the core management target from the aspects of reducing crash risk with operation control strategies. Therefore, it is crucial to analyze the concurrent effects of the influencing factors rather than the correlation relationships; where the former one could reveal which factor is the mediator variable that leads to crash occurrence.

In this study, a hybrid modeling approach with latent class logit (LCL) model and path analysis (PA) models was proposed to explore the causal relationships with simultaneously considering the heterogeneous influencing effects. Within the structure, the LCL model was developed to identify the homogeneous subgroups of crash data and their unique set of influencing factors. Then, PA models were established for each subset crash data separately to reveal the confounding relationships among the microscopic traffic flow variables and their impacts on crash occurrence.

The rest of the paper is organized as follows. The second section reviews the previous studies that focused on solving the heterogeneous effects and investigating the crash mechanisms. Then, the next section provides detailed description of the data preparation procedures, which is followed by introduction of the employed methodologies. The fifth section presents the modeling results, and finally, summaries and discussion of the work are provided.

2. Background

2.1. Analyzing the heterogeneous effects

In traffic safety analyses, the unobserved heterogeneity has always been regarded as one of the most critical issues. It was concluded that if the heterogeneous effects were not being properly considered, biased coefficient estimations and erroneous crash predictions could be obtained (Greene, 2000; Mannering et al., 2016). In order to account for the heterogeneous effects, there are mainly two approaches: one is to manually divide crash data into homogeneous clusters according to their attributes, such as operating conditions (Abdel-Aty et al., 2005) and crash types (Yu and Abdel-Aty, 2013); the other one is to establish hierarchical models (Huang and Abdel-Aty, 2010; Xu et al., 2014) to account for the heterogeneous effects.

As for the former approach, Abdel-Aty et al. (2005) developed crash risk analysis models based on high-speed and low-speed operating conditions to identify the different crash contributing factors. Yu and Abdel-Aty (2013) conducted crash contributing factor analyses for single-vehicle and multi-vehicle crashes separately. Besides, Xu et al. (2014) evaluated the likelihood of crash occurrence by levels of service (LOS). However, the manual classifications of crash data would be subjective and inefficient. Recent studies have been more inclined to utilize hierarchical models to address the impact of heterogeneity.

Huang and Abdel-Aty (2010) introduced the Bayesian hierarchical modeling approach to accommodate the potential cross-group heterogeneity and spatiotemporal correlations given the multilevel data structure. Compared to the traditional generalized linear model, substantial model goodness-of-fit improvements have been obtained. Later, Xu et al. (2014) applied a Bayesian random-parameters logistic regression model to link crash likelihood with various traffic flow characteristics under different LOS, and better crash risk prediction accuracies have been achieved. However, these hierarchical models paid more effort to analyze the heterogeneous effects among individual observations rather than concluding subgroups of homogenous observations (Mannering et al., 2016), which helps to classify crash-prone traffic conditions and further understand crash mechanisms.

Latent class models, which were frequently employed to address the unobserved heterogeneity issues, have the benefit of analyzing contributing factors across subgroups (Mannering et al., 2016). The LCL model, a specific type of latent class models that deals with discrete outcome variables, has been widely adopted in the research areas of food safety (Ortega et al., 2011), environmental economics (Provencher and Bishop, 2004), and health care (Deb and K. Trivedi, 2002). In addition, the LCL model has also been utilized in crash injury-severity analyses. Xie et al. (2012) utilized an LCL model to analyze injury severities involving single-vehicle crashes on rural roads. The impact of the same explanatory variable has distinct estimation and varies by the injury severity outcome. However, to the best of our knowledge, the LCL model has not been employed for real-time crash risk analyses.

2.2. Causal relationship analyses

Given the importance of understanding crash mechanisms, currently most studies have only identified the association relationships between the microscopic traffic flow variables and crash outcomes, and there were limited studies that focused on the causal factor inference. Elvik et al. (2004) proposed a rational framework to divide the crash contributing factors into four different causal categories: mediator variables, independent variables, moderator variables and confounding variables. And a causal diagram was established to link them. Then, Gargoum and El-Basyouny (2016) applied PA models to analyze the causal relationship between average speed and crash frequency. The mediated effects of traffic volume and other geometric characteristics on speed and their impacts on crashes have been identified. However, the authors only analyzed aggregated operation features instead of microscopic traffic flow variables. Then, Xu et al. (2017) formulated latent traffic variables based on disaggregate traffic flow data and employed structural equation modeling (SEM) to capture the direct and indirect effects of latent traffic variables on crash occurrence. However, due to the data dimensionality reduction process, the authors failed to identify the confounding effects between crash occurrence and its contributing factors. In addition, the above-mentioned studies failed to account for the heterogeneous feature of crash mechanisms. Therefore, here we try to fill the gap through proposing a hybrid modeling approach with LCL and PA models to identify the causal relationships of microscopic traffic flow contributing factors on crash occurrence with consideration of their heterogeneous effects.

3. Data preparation

Data from the Shanghai urban expressway system were utilized in this study. The expressway system was chosen as the study area given its high-dense traffic sensing devices and high-quality traffic crash data. The urban expressway system holds an average of 650 m spaced loop detectors, which enable the collection of high-quality microscopic traffic flow data (Yu et al., 2018). In addition, the urban expressway system holds a high crash occurrence rate and varying traffic operation conditions that would provide enough sample data for the analyses.

3.1. Raw analyses data

A total of three datasets from the Shanghai urban expressway system were utilized, which are traffic data, crash information data, and roadway geometry characteristics:

(1) Traffic data from April to June in 2014 were employed, which were obtained from the dual loops detectors (LDs) located on both directions of the Shanghai urban expressways. The raw data, including records of speed, volume and occupancy, were collected at 20-second intervals.
(2) Crash information data were provided by the Shanghai Traffic Information Center, where crash times and locations were recorded based on video surveillance system. In this study, only two-vehicle and multi-vehicle crashes were considered, since single-vehicle crashes are more likely to be affected by the vehicle mechanical failures or erroneous driving behaviors rather than the influence of traffic flow. A total of 2485 crash records were obtained for the three-month analysis period.
(3) Roadway geometric characteristics were obtained from the online street-view map, since there were no detailed design files available. Besides, the urban expressway system has been further split into 206 sections where each section refers to the expressway mainline segment between two adjacent ramps, and each section has been assigned with a unique section ID.

Then, the three datasets were merged together. First, crash data and traffic data were joined with section ID information based on their location features. For each specific crash, with its section ID, crash data can be matched with traffic data at the section level (Yu et al., 2016). Besides, upstream and downstream section IDs can be obtained from road geometry data according to the encoding rules.

3.2. Microscopic traffic flow variable extraction

Two types of traffic conditions, crash-prone and non-crash traffic conditions, were extracted. The crash-prone conditions refer to the traffic operating conditions prior to crash occurrence at three consecutive roadway segments: crash section (C), adjacent upstream section (U) and downstream section (D). And the crash-prone conditions were depicted by microscopic traffic flow variables, such as operation speed and traffic volume at five minutes' intervals. Besides, the non-crash traffic conditions refer to the control groups that corresponding to the crash-prone conditions with considering the confounding effects like roadway geometric characteristics, time of day, and day of week.

As for the crash-prone condition extraction, based on the crash reporting time, a 30-mintue interval raw traffic data of the crash section (C), upstream section (U), and downstream section (D) prior to crash occurrence were extracted, which were further split into six 5-minute time slices. These raw data were collected at 20-second intervals and further aggregated on 5-minute intervals to obtain the microscopic traffic flow variables. Then, the 30-minute interval was processed into six time slices, which were named as time slice 1 to time slice 6 with time slice 1 refers to the 5-minute time slice closest to crash occurrence. For example, if a crash occurred on April 21 st, 2014 (Monday) at 9:30 am, the traffic data of crash section (C), upstream section (U), and downstream section (D) from 9:00 am to 9:30 am (30-minute interval, or six 5-minute time slices) were extracted, and time slice 1 here refers to 9:25 am to 9:30 am. The nomenclature for defining road sections and time slices is given in Fig. 1.

As for the non-crash traffic condition extractions, matched case-control data structure was adopted in this study. The matched case-control data structure was frequently utilized in the disaggregate crash occurrence studies (Abdel-Aty et al., 2004; Yu and Abdel-Aty, 2013; Chen et al., 2018), and it was claimed that this approach could account for the confounding factors such as roadway geometric characteristics, time of day, and day of week through the matching processes. For each specific crash case, a random selection of m controls (non-crash cases) was chosen. Ahmed et al. (2012) examined the different 1:m (crash to non-crash) ratios and identified that when m is equal to 4, the model provided higher crash prediction precision. Hence, in this study, 4 non-crash cases were collected for each crash as in consistent with the majority studies (Lord and Washington, 2018), considering the same time of day, the same day of week, and the same location but different weeks (two weeks before and two weeks after the crash). For the same above-mentioned example, the non-crash traffic conditions would be collected at the same roadway sections and time periods for April 7th, April 14th, April 28th, and May 5th. After the data processing, the final dataset contains of 2485 crash cases and 7204 non-crash cases (the non-exact 1:4 ratio of crash and non-crash is due to loop detector malfunctions and that the non-crash cases discarded during the data processing procedure).

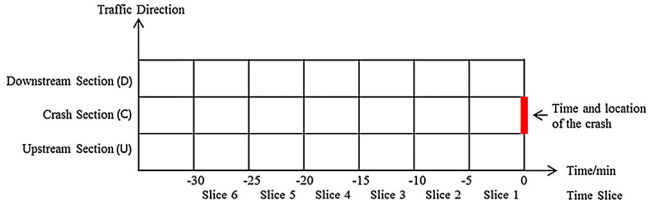

Fig. 1. Nomenclature for defining sections and time slices.

Furthermore, for each crash and non-crash case, the calculated microscopic traffic flow variables for modeling include the descriptive statistics (average and standard deviation) for speed, volume and occupancy at three sections for a total of six time slices. And a unified nomenclature method was proposed for the traffic variables. The nomenclature includes three letters and one numeric characteristic, as shown in Fig. 2, where the first letter takes the value of A or S for average or standard deviation, respectively. The second letter takes the value of S, V, or O for speed, volume, or occupancy. The third letter takes the value of U, C, or D for upstream section, crash section, or downstream section. And the last number takes the value of 1, 2, 3, 4, 5, or 6 referring to the six 5-minute time slices. Finally, a total of 108 (2 descriptive statistics × 3 traffic flow parameters × 3 road sections × 6 time slices) microscopic traffic flow variables were calculated for each crash or non-crash observation.

3.3. Modeling variable selection

The calculated microscopic traffic flow variables mentioned above may correlate with each other and some of them may not have substantial influences on crash occurrence. Therefore, in order to enhance the modeling efficiency, a variable selection procedure was conducted as follows.

First, the 18 variables of time slice 1 (0–5 min prior to the crash occurrence) were dropped to avoid the misreporting of crash time according to previous research (Yu et al., 2016), and there were 90 of the original 108 variables left.

Then, random forest (RF) models were employed to rank the variables' importance. RF, proposed by Breiman (2001), has been widely used for variable selection in crash risk analyses (Pande et al., 2011; Shi and Abdel-Aty, 2015). In this study, considering the correlations between the 90 microscopic traffic flow variables, *cforest* from party package (Hothorn et al., 2006) in R is used here and the permutation variables' importance were obtained (Strobl et al., 2009). The RF model was conducted for 16 times with different random seeds to eliminate the effects of random experiments. After cross comparisons of the variables' ranking consistency and check for the correlation effects, five variables were identified for further analyses, which were ASC2, SOC2, AOD2, ASU2, and SSC2. The descriptive statistics of the modeling variables are shown in Table 1.

4. Methodology

A hybrid modeling approach with LCL and PA models was proposed to simultaneously consider the unobserved heterogeneity issues while exploring the confounding relationships between crashes and

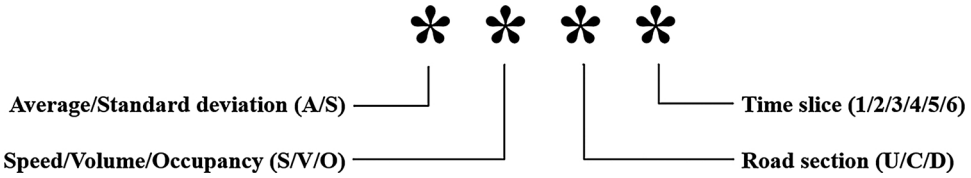

Fig. 2. Nomenclature of the microscopic traffic flow variables.

contributing factors.

4.1. Latent class logit model

The LCL model was initially introduced by Henry and Neil (1968). In this study, suppose there are C distinct subgroups of crash data. The probability of observation n been identified in class c can be expressed as:

$$P_n(\beta_c) = exp(y_n \beta_c x_n)/1 + exp(\beta_c x_n) \quad (1)$$

where y_n is a binary variable denoting the outcomes of crash and non-crash cases, which equals 1 if observation n is a crash case, and 0 otherwise.

Since the class membership status is unknown, the unconditional likelihood of observation' s choice needs to be specified; which equals to the weighted average of the above-mentioned probabilities over classes. For instance, the weight for class c, $\pi_{cn}(\theta)$, is the population share of that class and usually modeled as fractional multinomial logit:

$$\pi_{cn}(\theta) = \exp(\theta_c z_n)/1 + \sum_{l=1}^{C-1} \exp(\theta_l z_n) \quad (2)$$

where θ_l is class membership model parameter, θ_c is normalized to 0 for identification, and z_n is a constant of observation-specific characteristics.

Finally, the log-likelihood for the sampled data can be obtained by adding each individual observation's log unconditional likelihood as:

$$lnL(\beta, \theta) = \sum_{n=1}^{N} ln \sum_{c=1}^{C} \pi_{cn}(\theta) P_n(\beta_c) \quad (3)$$

Maximizing the above-mentioned likelihood directly was concluded to be a computational difficulty, then an expectation-maximization (EM) algorithm was proposed to estimate β and θ for likelihood maximization (Bhat, 1997; Train, 2008). In this study, the EM inference approach was completed in Stata (Pacifico and il Yoo, 2012).

4.2. Path analysis model

The PA model can be viewed as a special case of SEM where all variables are manifest (measurable/observed) variables (Gargoum and El-Basyouny, 2016). Fig. 3 shows a typical schematic of the PA model, which comprises an inner and an outer model. Variables in the outer model are called "exogenous", such as ; variables in the inner model are referred to as "endogenous", such as X_m and Y. In the model below, the exogenous independent variables X are modeled as simultaneously having both direct and indirect (through mediator independent variable X_m) effects on dependent variable Y.

The above showed model can be expressed as:

$$Y = \beta_0 + \beta_1 X_m + \beta_2 X + \varepsilon_1 \quad (4)$$

$$X_m = \gamma_0 + \gamma_1 X + \varepsilon_2 \quad (5)$$

where, Y denotes the dependent variable; X_m is the mediator independent variable; X is the exogenous independent variable matrix; ε_1 and ε_2 are the errors; β_0 and γ_0 are the intercepts; β_1, β_2, and γ_1 are the regression coefficients to be estimated.

The estimated coefficients β_1, β_2, and γ_1 are used to quantify the impacts of independent variables on dependent variables, where β_2 denotes the direct effect of X on Y and the magnitude of the indirect effect of X on Y can be estimated by $\gamma_1 \beta_1$.

In this study, the dependent variable Y was set to be crash occurrence, which is a binary outcome variable. Then, according to the model definition, mediator independent variable refers to the variable that is more directly causing the crash occurrence than exogenous independent variable. Since the crash occurred in the crash section, the crash section variable is considered to be more directly causing the crash occurrence than other section variables. Hence, the mediator independent variable X_m was set to be crash section related variable. That is to say, except AOD2 and ASU2, the other three microscopic traffic flow variables ASC2, SOC2, and SSC2 are all candidate mediator independent variables. Finally, the exogenous independent variables X, were set to be other (except the mediator independent variable) significant microscopic traffic flow variables identified by the LCL model. One point to mention is that both the direct and indirect effects of exogenous independent variables on dependent variables have been tested.

In addition, in order to compare the various potential model structures, the following criteria were set up to obtain the best-fitted PA model. First, the statistically insignificant paths were dropped based on confidence level ($< 95\%$). Then, Root Mean Square Error of Approximation (RMSEA) was used for model selection (Browne and Cudeck, 1992). RMSEA, which represents the under-fitted between the proposed model and the population, can be defined as:

$$RMSEA = \sqrt{(\chi_S^2/df_S) - 1/N} \quad (6)$$

where, χ_S^2 denotes the model Chi-square statistic, df_S denotes the degree of freedom; N denotes the sample size.

In general, the interpretation of RMSEA values is: 0 means perfect fit; $RMSEA \leq 0.08$ is recommended as the threshold for acceptable model structure (Browne and Cudeck, 1992).

As for the model inference, Mplus (Muthén and Muthén, 2012) was used. Besides, since the dependent variable, crash occurrence, is a binary outcome variable, Mean and Variance-adjusted Weighted Least Square (WLSMV) estimation was used to compute the estimated coefficients (Aarts and van Schagen, 2006; Muthén and Muthén, 2012).

Table 1
Summary descriptive statistics for modeling variables.

Variable	Description	Mean	Std. Dev.
ASC2	Average speed at crash section 5–10 min prior to crash occurrences	43.32	18.32
SOC2	Standard deviation of occupancy at crash section 5–10 min prior to crash occurrences	3.66	2.43
AOD2	Average occupancy at downstream section 5–10 min prior to crash occurrences	20.79	13.75
ASU2	Average speed at upstream section 5–10 min prior to crash occurrences	48.52	18.92
SSC2	Standard deviation of speed at crash section 5–10 min prior to crash occurrences	4.08	2.48

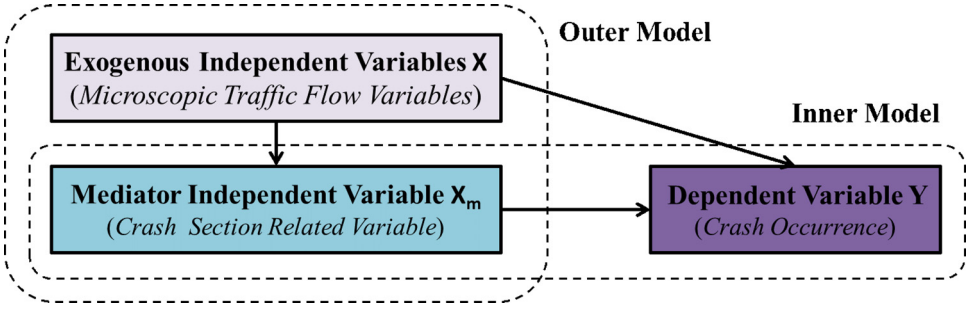

Fig. 3. Schematic of PA model.

5. Modeling results

5.1. Latent class logit modeling results

In order to identify the optimal number of latent classes, evaluation indicators including consistent Akaike's information criterion (CAIC), Bayesian information criterion (BIC), log likelihood, and the area under the ROC (receiver operating characteristics) curve (AUC) were utilized. And after testing the latent class numbers from 2 to 9, the optimal number of latent classes is chosen to be 4.

The 4-class LCL model estimation results are shown in Table 2. From the table it can be seen that the statistically significant variables in each class are inconsistent, which indicates the existence of the heterogeneity among the crash occurrence contributing factors. The modeling results for each latent class have been explained as follows.

(1) For latent class one, a total of three microscopic traffic flow variables are statistically significant at 95% confidence level, which are ASC2 (average speed at crash section in time slice 2), AOD2 (average occupancy at downstream section in time slice 2), and SSC2 (standard deviation of speed at crash section in time slice 2). ASC2 has a negative coefficient, which indicating that lower values of operating speed at crash section would lead to increased crash risk. Besides, both AOD2 and SSC2 are found to be positively correlated with crash likelihood, which means higher average occupancy at downstream section and higher variation of speed at crash section would increase crash risk.
(2) As for the latent class two, three microscopic traffic flow variables are statistically significant at 99% confidence level, which are ASC2 (average speed at crash section in time slice 2), ASU2 (average speed at upstream section in time slice 2), and SSC2 (standard deviation of speed at crash section in time slice 2). Both ASC2 and ASU2 have negative coefficients, indicating that lower values of operating speed prior to crash occurrence at both crash section and upstream section would lead to increased crash risk. Similar to latent class one, SSC2 is found to have a positive effect.
(3) For latent class three, three microscopic traffic flow variables are found to be statistically significant at 95% confidence level, which are ASC2 (average speed at crash section in time slice 2), SOC2 (standard deviation of occupancy at crash section in time slice 2), and AOD2 (average occupancy at downstream section in time slice 2). Similar to latent class one, ASC2 and AOD2 are found to have a negative and a positive effect on crash occurrence, respectively. Besides, SOC2 has a positive coefficient, suggesting that there would be high crash occurrence probability when variation of occupancy is large.
(4) As for latent class four, the statistically significant variables are exactly the same with latent class two at 95% confidence level. While at 99% confidence level, only two significant variables left, which are ASC2 (average speed at crash section in time slice 2) and ASU2 (average speed at upstream section in time slice 2), and they are both negatively related to crash occurrence.

Table 2
Modeling results of the LCL model.

Latent Class #	Class Share	Variable	Mean	Std. Dev.	P-value	Significance[1]
One	0.212	ASC2	−0.021	0.010	0.042	**
		SOC2	−0.026	0.050	0.606	
		AOD2	0.067	0.014	0.000	****
		ASU2	0.001	0.012	0.916	
		SSC2	0.095	0.048	0.049	**
Two	0.039	ASC2	−0.114	0.034	0.001	***
		SOC2	−0.150	0.241	0.533	
		AOD2	0.036	0.044	0.409	
		ASU2	−0.107	0.038	0.005	***
		SSC2	0.695	0.217	0.001	***
Three	0.148	ASC2	−0.056	0.020	0.005	***
		SOC2	0.269	0.117	0.022	**
		AOD2	0.078	0.022	0.000	****
		ASU2	0.017	0.013	0.203	
		SSC2	0.079	0.116	0.496	
Four	0.601	ASC2	−0.047	0.006	0.000	****
		SOC2	0.037	0.032	0.246	
		AOD2	0.006	0.005	0.260	
		ASU2	−0.019	0.004	0.000	****
		SSC2	0.052	0.026	0.041	**

Note: Significance[1], ****: 99.9% confidence level is used (i.e., p-value < 0.001 is statistically significant); ***: 99% confidence level is used; **: 95% confidence level is used.

The estimation results for the LCL analyses have showed that each class holds a unique set of significant microscopic traffic flow variables correlated with crash occurrence. Moreover, the same specific independent variable may have distinct impact on the crash likelihood across the latent classes, which further indicates the existence of unobserved heterogeneity. In order to further investigate the causal relationships among microscopic traffic flow variables, path analyses have been separately conducted for each latent class in the following process.

One point that needs to be mentioned is that since there are several potential PA structures given the different candidate mediator independent variables and the potential direct and indirect paths, the best-fitted PA models were first identified based on the confidence level and RMSEA. Table 3 lists all the tested PA model structures.

Take latent class one for example, among the three statistically significant independent variables, ASC2 and SSC2 were selected as the candidate mediator independent variable, considering that they are crash section related, especially speed related. When ASC2 was the mediator independent variable, both direct and indirect (mediated through ASC2) effects of exogenous independent variables AOD2 and SSC2 on crash occurrence were considered. When SSC2 was the mediator independent variable, both direct and indirect (mediated through SSC2) effects of exogenous independent variables ASC2 and AOD2 on crash occurrence were considered. Hence, a total of 8 potential model structures were proposed, and the best-fitted model has been obtained

Table 3
The process of obtaining the best-fitted PA model for each latent class.

Latent Class #	Statistically Significant Independent Variables	Candidate Mediator Independent Variable	# of Potential Model Structures	Best-Fitted Model Mediator Independent Variable
One	ASC2, AOD2, SSC2	ASC2/ SSC2	8	ASC2
Two	ASC2, ASU2, SSC2	ASC2/ SSC2	8	SSC2
Three	ASC2, SOC2, AOD2	ASC2/ SOC2	8	ASC2
Four	ASC2, ASU2	ASC2	2	ASC2

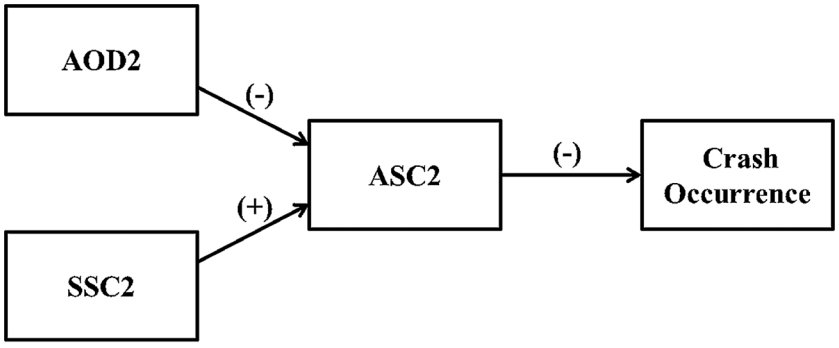

Fig. 4. Final PA model structure of latent class one.

Table 4
Estimation result of PA model of latent class one.

Latent class one		Mean	Std. Dev.	P-value
Effects of AOD2 on ASC2		−0.646	0.032	0.000
Effects of SSC2 on ASC2		0.733	0.178	0.000
Effects of ASC2 on Crash Occurrence		−0.012	0.002	0.000
Effects of AOD2 on Crash Occurrence	Total	0.008	0.001	0.000
	Direct	–	–	–
	Indirect	0.008	0.001	0.000
Effects of SSC2 on Crash Occurrence	Total	−0.009	0.003	0.001
	Direct	–	–	–
	Indirect	−0.009	0.003	0.001

Table 5
Estimation result of PA model of latent class two.

Latent class two		Mean	Std. Dev.	P-value
Effects of ASU2 on SSC2		−0.077	0.018	0.000
Effects of ASC2 on SSC2		0.073	0.011	0.000
Effects of SSC2 on Crash Occurrence		0.076	0.028	0.006
Effects of ASU2 on Crash Occurrence	Total	−0.006	0.003	0.019
	Direct	–	–	–
	Indirect	−0.006	0.003	0.019
Effects of ASC2 on Crash Occurrence	Total	−0.028	0.007	0.000
	Direct,121%	−0.034	0.007	0.000
	Indirect,21%	0.006	0.002	0.011

after comparison. The best-fitted PA models for each latent class are further explained as follows.

5.2. Path analysis model for latent class one

As for latent class one, the best-fitted PA model is shown in Fig. 4, and the estimation results are listed in Table 4. ASC2 is the mediator independent variable while AOD2 and SSC2 are exogenous independent variables. From the results it can be seen that ASC2 and SSC2 are negatively correlated with crash likelihood. AOD2 have positive impacts on the crash occurrence risk, which is the appositive findings compared to the LCL model. In addition, both the effects of AOD2 and SSC2 on crash occurrence are fully mediated through ASC2, without statistically significant direct effects.

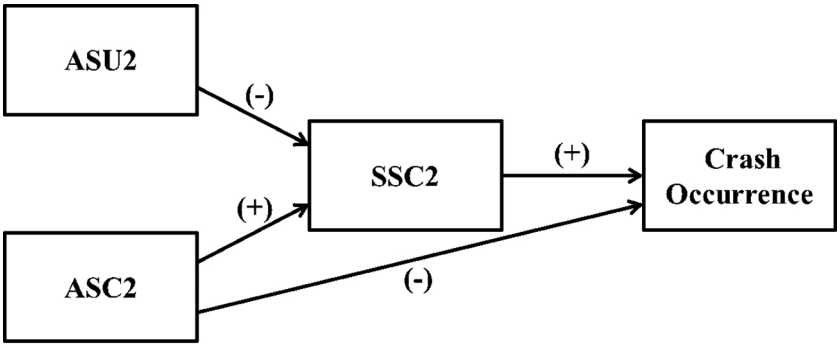

Fig. 5. Final PA model structure of latent class two.

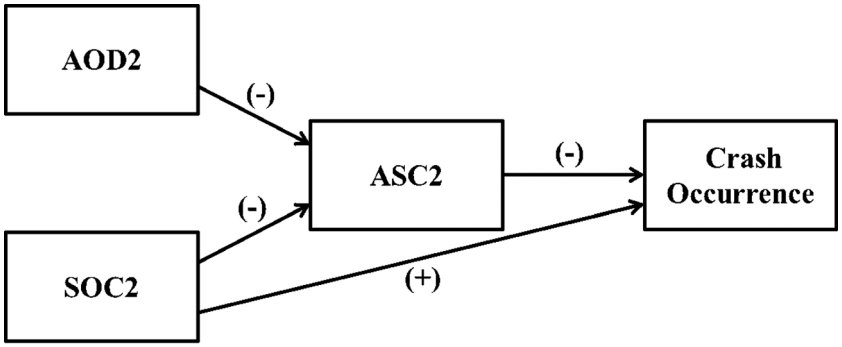

Fig. 6. Final PA model structure of latent class three.

Table 6
Estimation result of PA model of latent class three.

Latent class three		Mean	Std. Dev.	P-value
Effects of AOD2 on ASC2		−0.554	0.031	0.000
Effects of SOC2 on ASC2		−3.798	0.144	0.000
Effects of ASC2 on Crash Occurrence		−0.019	0.003	0.000
Effects of AOD2 on Crash Occurrence	Total	0.010	0.002	0.000
	Direct	–	–	–
	Indirect	0.010	0.002	0.000
Effects of SOC2 on Crash Occurrence	Total	0.138	0.016	0.000
	Direct,49%	0.067	0.021	0.002
	Indirect,51%	0.071	0.010	0.000

5.3. Path analysis model for latent class two

As for latent class two, Fig. 5 shows the best-fitted PA model and the estimation results are listed in Table 5. SSC2 is the mediator independent variable; ASU2 and ASC2 are the exogenous independent variables. As concluded, SSC2 has a positive effect, while ASU2 and ASC2 have negative effects. The effect of ASU2 on crash occurrence is fully mediated through SSC2. While ASC2 has both direct and indirect (mediated through SSC2) effects on crash occurrence. Interestingly, the direct effect is negative, but the indirect effect is positive. Since the direct effect is much larger than the indirect one, -0.034 vs. 0.006, the total effect is negative.

5.4. Path analysis model for latent class three

As for latent class three, the best-fitted PA model is shown in Fig. 6 and the estimation results are listed in Table 6. ASC2 is the mediator independent variable; AOD2 and SOC2 are the exogenous independent variables. ASC2 has a negative effect, while AOD2 and SOC2 have positive effects. The effect of AOD2 on crash occurrence is fully mediated through ASC2, while SOC2 has nearly same direct and indirect (mediated through ASC2) effects on crash occurrence, which are 49% and 51%, respectively.

Table 7
Estimation result of PA model of latent class four.

Latent class four		Mean	Std. Dev.	P-value
Effects of ASU2 on ASC2		0.554	0.012	0.000
Effects of ASC2 on Crash Occurrence		−0.016	0.001	0.000
Effects of ASU2 on Crash Occurrence	Total	−0.009	0.001	0.000
	Direct	–	–	–
	Indirect	−0.009	0.001	0.000

5.5. Path analysis model for latent class four

As for latent class four, Fig. 7 shows the best-fitted PA model and Table 7 listed the estimation results. ASC2 is the mediator independent variable; ASU2 is the only exogenous independent variable. These two variables are both negatively correlated with crash occurrence while the effect of ASU2 is fully mediated through the mediator independent variable.

6. Summary and discussions

In this study, a hybrid modeling approach with LCL and PA models has been developed to understand the crash mechanisms with consideration of the heterogeneous occurrence scenarios. First, the LCL model was developed and four latent classes with different sets of contributing factors were obtained. Then, the PA models were developed for each latent class to understand the confounding impacts among the microscopic traffic flow variables on crash occurrence.

Based on the proposed modeling scheme, more in-depth understandings of how the operation conditions impact crash occurrence have been concluded. Table 8 compares the modeling results from the both models, and interesting results have been found:

(1) For latent class one, as indicated by the PA model that SSC2 has no direct impacts on crash occurrence since it was proved to contain only indirect and positive impacts on ASC2, that further influence the crash occurrence. This finding is totally opposite to the LCL model.
(2) For latent class two, the difference is that ASU2 only has indirect

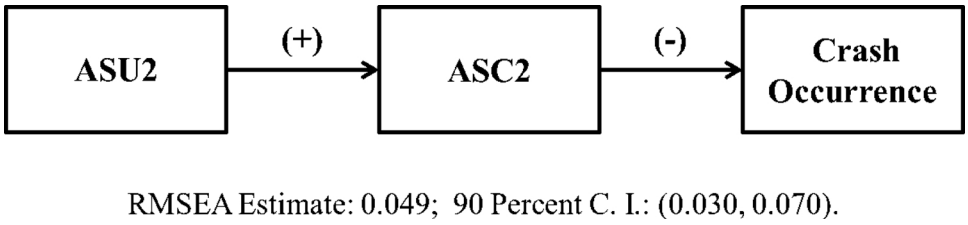

Fig. 7. Final PA model structure of latent class four.

Table 8
Comparison of the results of the LCL model and PA models for four latent classes.

		ASC2	SOC2	AOD2	ASU2	SSC2
Latent Class One	LCL	–	N.S.[1]	+	N.S.	+
	PA	-(direct)*	N.S.	+	N.S.	-(indirect)
Latent Class Two	LCL	–	N.S.	N.S.	–	+
	PA	-(direct) > +(indirect)	N.S.	N.S.	-(indirect)	+(direct)*
Latent Class Three	LCL	–	+	+	N.S.	N.S.
	PA	-(direct)*	+(direct/ indirect)	+(indirect)	N.S.	N.S.
Latent Class Four	LCL	–	N.S.	N.S.	–	N.S.
	PA	-(direct)*	N.S.	N.S.	-(indirect)	N.S.

Note: N.S.[1]: not significant. *: the mediator independent variable in PA model. Note: N.S.[1]: not significant. *: the mediator independent variable in PA model.

influences on crash occurrence through the mediator factor of SSC2 while ASC2 simultaneously holds the direct and indirect effects.

(3) For latent class three, the impact directions of the microscopic traffic flow variables are consistent for the two types of models. However, the PA model has concluded that AOD2 did not contain direct impacts on crash occurrence, and its influencing mechanisms is through the ASC2.

(4) Finally, for the latent class four, the ASU2 was concluded to only have indirect impacts on crash occurrence, and its influencing mechanisms is also through the ASC2.

Interestingly, although the upstream and downstream microscopic flow variables have been widely included in the crash risk evaluation models. However, they were concluded to mainly have indirect impacts on crash occurrence from the above-mentioned results. Besides, for the crash occurrence sections, since the flow variables of operation speed, standard deviation of speed and occupancy have inter-correlations through the traffic flow theory, the PA models have identified that they could have both direct impacts on crash risk and indirect influences through the mediator variables. Moreover, the revealed confounding effects of microscopic traffic flow variables could shed lights on understanding the pre-crash traffic operation scenarios.

For latent class one, according to the PA modeling result, AOD2 and SSC2 were concluded to have only indirect influences on crash occurrence through ASC2. In addition, the crash-prone scenario can be inferred as that it is likely to have a queue formation at downstream section, where it becomes congested with increasing occupancy. Then the shockwave propagates to the crash section and further reduces the operating speed there. Under this scenario, the inconsistent braking behaviors or increased sudden lane change behaviors would increase the crash likelihood. And it is suggested to implement a queue warning system one section ahead or an in-vehicle speed advisory system to avoid this type of crash occurrence scenario.

For latent class two, it can be summed up from PA model that the negative effect of ASU2 on crash occurrence is fully mediated through ASC2, while the final negative effect of ASC2 on crash occurrence is the result of a 121% direct negative effect superimposed by a 20% indirect (mediated through ASC2) positive effect. This crash occurrence scenario can be seen that the crash occurrence was due to heterogeneous travelling speeds at the crash segment, and improvement strategies like variable speed limit (VSL) control are recommended to smooth the traffic and further improve safety.

For latent class three, according to the PA modeling results, the positive effect of AOD2 on crash occurrence is fully mediated through ASC2, while for SOC2, direct and indirect (mediated through ASC2) effects share together. Moreover, a specific crash-prone scenario can be described as a congestion dissipating process, where the different acceleration and deceleration behaviors have increased the traffic turbulence at crash sections. In this case, control strategies can be designed to provide more smooth dissipating at downstream sections.

Finally, for latent class four, it can be summed up from PA model that the negative effect of ASU2 is fully mediated through ASC2. Moreover, this is a typical crash-prone scenario with 60.1% latent class sharing. The scenario is a simple congested roadway segments at both upstream and crash sections, where temporary hard shoulder running and lane change prohibitions could be useful.

However, given the interesting findings from the proposed modeling approach, there are still several limitations of the current study.

(1) First, this study employed matched case-control data structure with a fixed crash and non-crash ratio (1:4) to deal with the imbalanced crash risk analysis dataset. In future studies, other sampling techniques, such as over-sampling methods of synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) and full-size data would be tested. And the effects of sampling methods on the concurrent microscopic influencing factors need to be revealed.

(2) Second, this study utilized the LCL model to group observations to account for the heterogeneity issues, and future efforts would be conducted to identify how to improve the proposed hybrid approach from the aspects of incorporating other clustering methods, such finite mixture model (FMM) (Park and Lord, 2009).

(3) Moreover, after applying the LCL model to solve the heterogeneity issues, repeated measures still exist within a latent class. In this study, the matched case-control data preparation process can greatly reduce the impacts of repeated measures on the analysis results. While in future studies, other methods would be tested to solve this repeated measurement issues, such as incorporating the LCL model with roadway section unique random effect terms.

(4) In addition, in this study, the PA models were employed to conduct the causal inference, which is a typical mediation analysis method. Additional efforts are needed to utilize other causal inference methods (Pearl, 2018), like propensity score matching (PSM) (Li et al., 2013; Li and Graham, 2016), to further explorations the causal relationships between crash occurrence and the microscopic traffic flow variables.

Acknowledgments

This study was jointly sponsored by the Chinese National Natural Science Foundation (NSFC 71771174 and 71401127) and the 111 Project (B17032).

References

Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: a review. Accid. Anal. Prev. 38 (2), 215–224.
Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec. 1897, 51–58.
Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. Transp. Res. Rec. 1908, 51–58.
Abdel-Aty, M., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transp. Res. Part C Emerg. Technol. 24 (9), 288–298.
Ahmed, M., Abdel-Aty, M., Yu, R., 2012. Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. Transp. Res. Rec. 2280 (1), 60–67.

Bhat, C.R., 1997. An endogenous segmentation mode choice model with an application to intercity travel. Transp. Sci. 31 (1), 34–48.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Browne, M.W., Cudeck, R., 1992. Alternative ways of assessing model fit. Sociol. Methods Res. 21 (2), 230–258.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16 (1), 321–357.
Chen, Z., Qin, X., Shaon, M.R.R., 2018. Modeling lane-change related crashes with lane-specific real-time traffic and weather data. J. Intell. Transp. Syst. Technol. Plan. Oper. 22 (4), 291–300.
Deb, P., K. Trivedi, P., 2002. The structure of demand for health care: latent class versus two-part models. J. Health Econ. 21 (4), 601–625.
Elvik, R., Christensen, P., Amundsen, A., 2004. Speed and Road Accidents: An Evaluation of the Power Model 740. The Institute of Transport Economics Report, pp. 2004.
Gargoum, S.A., El-Basyouny, K., 2016. Exploring the association between speed and safety: a path analysis approach. Accid. Anal. Prev. 93, 32–40.
Greene, W.H., 2000. Econometric Analysis. Pearson Education, India.
Henry, Neil, W., 1968. Latent structure analysis. Am. Sociol. Rev. 34 (2).
Hothorn, T., Hornik, K., Zeileis, A., 2006. Party: A Laboratory for Recursive Part(y) itioning. R Package Version 0.9-11.
Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42 (6), 1556–1565.
Kwak, H.C., Kho, S., 2016. Predicting crash risk and identifying crash precursors on korean expressways using loop detector data. Accid. Anal. Prev. 88, 9–19.
Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-time crash prediction model for application to crash prevention in freeway traffic. Transp. Res. Rec. 1840, 67–77.
Li, H., Graham, D.J., 2016. Quantifying the causal effects of 20mph zones on road casualties in london via doubly robust estimation. Accid. Anal. Prev. 93, 65–74.
Li, H., Graham, D.J., Majumdar, A., 2013. The impacts of speed cameras on road accidents: an application of propensity score matching methods. Accid. Anal. Prev. 60, 148–157.
Lord, D., Washington, S., 2018. Safe Mobility: Challenges, Methodology and Solutions. Emerald Publishing Limited, pp. 175–204.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.
Muthén, L.K., Muthén, B.O., 2012. Mplus User's Guide, 7th edition. Muthén & Muthén, Los Angeles, CA.
Ortega, D.L., Wang, H.H., Wu, L., Olynk, N.J., 2011. Modeling heterogeneity in consumer preferences for select food safety attributes in China. Food Policy 36 (2), 318–324.
Pacifico, D., Il Yoo, H., 2012. Lclogit: a stata module for estimating latent class conditional logit models via the expectation-maximization algorithm. Stata J. 13 (3), 625–639.
Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38 (5), 936–948.
Pande, A., Das, A., Abdel-Aty, M., Hassan, H., 2011. Estimation of real-time crash risk. Transp. Res. Rec. 2237 (1), 60–66.
Park, H., Haghani, A., 2016. Real-time prediction of secondary incident occurrences using vehicle probe data. Transp. Res. Part C 70, 69–85.
Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41 (4), 683–691.
Pearl, J., 2018. The Seven Tools of Causal Inference with Reflections on Machine Learning. https://doi.org/10.1145/nnnnnnn.nnnnnnn.
Piccinini, G.B., Engström, J., Bärgman, J., Wang, X., 2017. Factors contributing to commercial vehicle rear-end conflicts in China: a study using on-board event data recorders. J. Saf. Res. 62, 143.
Provencher, B., Bishop, R.C., 2004. Does accounting for preference heterogeneity improve the forecasting of a random utility model? A case study. J. Environ. Econ. Manage. 48 (1), 793–810.
Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. Accid. Anal. Prev. 79, 198–211.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C 58, 380–394.
Strobl, C., Hothorn, T., Zeileis, A., 2009. Party on! A new, conditional variable importance measure for random forests available in the party package. R J. 1 (2), 14–17.
Train, K.E., 2008. Em algorithms for nonparametric estimation of mixing distributions. J. Choice Model. 1 (1), 40–69.
World Health Organization, 2013. Global Status Report on Road Safety 2013: Supporting a Decade of Action. World Health Organization.
Xie, Y., Zhao, K., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accid. Anal. Prev. 47 (3), 36–44.
Xu, C., Liu, P., Wang, W., Li, Z., 2014. Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. Transp. Res. Part A 69, 58–70.
Xu, C., Li, D., Li, Z., Wang, W., Liu, P., 2017. Utilizing structural equation modeling and segmentation analysis in real-time crash risk assessment on freeways. Ksce J. Civ. Eng. 1–9.
Yu, R., Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. Accid. Anal. Prev. 58, 97–105.
Yu, R., Abdel-Aty, M., 2014. An optimal variable speed limits system to ameliorate traffic safety risk. Transp. Res. Part C Emerg. Technol. 46, 235–246.
Yu, R., Wang, X., Yang, K., Abdel-Aty, M., 2016. Crash risk analysis for shanghai urban expressways: A Bayesian semi-parametric modeling approach. Accid. Anal. Prev. 95, 495–502.
Yu, R., Quddus, M., Wang, X., Yang, K., 2018. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. Accid. Anal. Prev. 120, 304–310.