Accident Analysis and Prevention 123 (2019) 29–38

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Exploring the impacts of speed variances on safety performance of urban elevated expressways using GPS data

Chuan Xu[a,b], Xuesong Wang[c,d,*], Hong Yang[e], Kun Xie[f], Xiaohong Chen[c,d]

[a] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan, 610031, China
[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu, Sichuan, 610031, China
[c] School of Transportation Engineering, Tongji University, Shanghai, 201804, China
[d] Road and Traffic Key Laboratory, Ministry of Education, Shanghai, 201804, China
[e] Department of Modeling, Simulation & Visualization Engineering, Old Dominion University, Norfolk, VA, 23529, United States
[f] Department of Civil and Natural Resources Engineering, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Keywords:
Spatial speed variance
Urban expressway
GPS mixing problem
FCD
Measurement error
Hierarchical model

ABSTRACT

Speed variation on urban expressways has been frequently noted as a key factor associated with high crash risk. However, it was often difficult to capture the safety impact of speed variance with spaced sensor measurements. As an alternative, this paper aims to leverage the use of the floating car data (FCD) to capture the speed variance in a morning rush hour on urban elevated expressways and examine its effect on safety. A semi-automatic filtering process was introduced to distinguish taxi GPS data points on the elevated expressways from the ones on the surface roads under the expressways. Subsequently, the standard deviation of the cross-sectional speed mean (SDCSM) and the cross-section speed standard deviation (MCSSD) were derived to capture the spatial and temporal speed variances, respectively. Together with other explanatory variables, both hierarchical and non-hierarchical Poisson-gamma measurement error models were developed to model the crash frequencies of the expressways. The modeling results showed that the hierarchical model performed better and both SDCSM and MCSSD were found to be positively related to the crash occurrence. This secures the need for addressing the impact of speed variation when modeling crashes occurred on the elevated expressways.

1. Introduction

Urban expressways usually maintain high speed limits with full access control that largely improve traffic operations in urban areas. Nonetheless, its safety performance has drawn great attention from transportation researchers and practitioners, spurring a great number of studies related to this subject (Kononov et al., 2008; Hossain and Muromachi, 2012, 2013b, 2013a; Zhang et al., 2013; Qu and Meng, 2014; Shi and Abdel-Aty, 2015; Wu and Sun, 2015; Shi et al., 2016a, b; Sun et al., 2016; Suzuki et al., 2016; Yu et al., 2016, 2017; Basso et al., 2018).

Unlike freeways, urban expressways typically have dense ramps to connect with auxiliary roads and their travel lanes configurations are often restricted from having wide lanes and/or shoulders. These special characteristics unavoidably induce many safety concerns among the public (Hossain and Muromachi, 2013a; Qu et al., 2014; Yun et al., 2017). Existing studies have reported three types of factors that are frequently associated with the safety performance of urban expressways: (1) factors related to roadway design and construction, e.g., road geometric factors, lane setting, divisions, and lighting; (2) factors on traffic operations, e.g., traffic flow parameters (i.e., volume, speed, and density) and traffic composition; and (3) factors that can be barely interposed currently, such as severe weather. Although it is difficult to intervene/control the first and the third types of factors post the built of urban expressways, the advances in active traffic management systems (ATMSs) empower operators with greater potential to manage the traffic operational factors. Nevertheless, an effective management scheme requires a clear understanding of the impact of these operational factors, including speed variation.

In ATMSs practices, traffic speed homogenization in space and time has been considered as an important scheme to improve traffic operations and safety. The measure of speed variance of different locations along the road can be called the spatial speed variance. Compared with interrupted-flow facilities (American Association of State Highway and Transportation Officials, 2010; Transportation Research Board, 2010), the safety impact of the spatial speed variance is expected to be more

* Corresponding author at: School of Transportation Engineering, Tongji University, Shanghai, 201804, China.
E-mail address: wangxs@tongji.edu.cn (X. Wang).

https://doi.org/10.1016/j.aap.2018.11.012
Received 15 March 2015; Received in revised form 9 November 2018; Accepted 12 November 2018
Available online 17 November 2018
0001-4575/

significant for uninterrupted-flow facilities such as freeways and urban freeways. This is because a driver may maintain a relatively high speed when approaching a congested site before becoming fully aware of the road situation downstream. Such large variations in speed between adjacent vehicles induce conflicts, which in turn can lead to crashes. Owing to the heavy traffic on urban expressways, this circumstance is quite common. This motivates a number of researchers to study the safety impact of the spatial speed variance based on fixed cross-sectional sensors, e.g., loop detector data and microwave vehicle detection system (Hossain and Muromachi, 2012; Xu et al., 2012; Hossain and Muromachi, 2013b; Wang et al., 2015a; Yu et al., 2016). Despite the progress, the insight into spatial speed variance gained from these studies can be limited at the locations with widely spaced detectors. It cannot be accurately monitored on roads without traffic sensors as well as the sections between two adjacent sensors. Temporal speed variance, which also called speed variation, is typically defined as standard deviation of speeds during a certain period (Kockelman and Ma, 2010; Quddus, 2013). However, it usually was measured from a single cross-section using loop detector, which was also highly restricted by the locations of the detectors and the overall temporal speed variance of a road segment couldn't be addressed.

In recent years, many commercial vehicles (taxis and buses) are equipped with the Global Position System (GPS) devices to record their travel information such as location and speed. This type of data is usually called probe vehicle data or floating car data (FCD). The FCD typically covers a large spatial area and enables users to estimate both spatial and temporal speed variations along a roadway. Thus, it offers a new valuable source to probe the relationship between the speed variance and crash frequency on the expressways. If the speed variance can serve as a surrogate safety measure, such information could be used to augment the decision-makings in developing ATM schemes.

Therefore, this paper aims to extract the speed variance from massive FCD and explore its safety effect on urban elevated expressways while controlling other contributing factors such as the traffic exposure and road design. It intends to leverage the use of the massive trajectory data generated by thousands of taxis in a megacity for safety purpose.

2. Literature review

The safety issues of urban expressways have been investigated in a number of studies. In general, these studies can be categorized into two sets based on the nature of the used datasets: (i) real-time disaggregated datasets (Hossain and Muromachi, 2012; Xu et al., 2012, 2013; Shi and Abdel-Aty, 2015; Sun and Sun, 2015; Wang et al., 2015a; Yu et al., 2016, 2017; Basso et al., 2018); and (ii) long-term aggregated datasets (Shi et al., 2016a, b; Sun et al., 2016).

To model the real-time crash risk of urban expressways, researchers typically linked the pre-crash traffic flow characteristics with the occurrence of crash. Assuming that the same traffic conditions will be associated with the same risk of crashes, these studies require the assemble of similar data (e.g., speed, occupancy, etc.) from both non-crash and crash conditions to construct a “matched-case control” dataset. For example, Abdel-Aty et al. (2004) and Yu et al. (2017) considered to match data of one crash case with that of four non-crash periods at the same location. The data for the non-crash periods are often selected at the same day of the week and the same time of the day (Abdel-Aty et al., 2004; Xu et al., 2012; Hossain and Muromachi, 2013b; Yu et al., 2016). Such selection attempts to control the variability caused by the time of the day and day of the week.

Instead of focusing on individual crash risk on urban expressways, a number of studies also explored the possibility to model the long-term crash frequency (Shi et al., 2016a, b; Sun et al., 2016). These modeling practices, like other safety research, require the use of long-term aggregated crash count data over a spatial unit (e.g., road segments, corridors, etc.) (Lord and Mannering, 2010; Mannering and Bhat, 2014). Correspondingly, many explanatory variables in the model also need to be aggregated. For example, to account for the impact of traffic flow researchers usually use the annual average daily traffic (AADT) instead of the real-time traffic volume as the exposure. This inevitably leads to questions on the simplified use of the data. For example, the averaged values cannot capture the differences of traffic factors at different periods (e.g., weekday versus weekend, rush hour versus non-rush hour, etc.). Alternatively, previously studies were often limited to analyze crashes occurred at specific days (Shi et al., 2016a, b) or peak hours (Qu et al., 2014; Shi et al., 2016a).

Despite the influence of various factors, it deserves to note that speed variance has been frequently reported as a critical risk factor in estimating crash risk on expressways (Xu et al., 2012; Hossain and Muromachi, 2013b; Quddus, 2013; Xu et al., 2013; Wang et al., 2015a; Yu et al., 2016). Arguably, many crashes occurred on roads with ideal geometry are owing to the sudden traffic disruptions (e.g., incidents at downstream, vehicle merging at weaving sections, etc.) (Hossain and Muromachi, 2012). Xu et al. (2012) identified a risky situation when the upstream traffic was a free flow while the downstream traffic was a congested flow. Under such conditions, the spatial speed variance is large. Hossain and Muromachi (2012) found that the differences between the traffic flow parameters (including speed) of upstream and downstream traffic flows had a significant effect on the precise detection of hazardous conditions. Hossain and Muromachi (2013b) also demonstrated that the speed difference between upstream and downstream traffic was one of the best factors to explain crashes. Xu et al. (2013) showed that property damage only (PDO) crashes were more likely to occur under congested traffic flow conditions with a highly variable speed. Wang et al. (2015a) exhibited that the speed difference between the start and end of a weaving segment had a major impact on the crash risk. Quddus (2013) found temporal speed variance (standard deviation of 1-h speeds) is positively associated with accident rates.

Although speed variance is commonly recognized as a key factor, it is difficult to consistently obtain such measurements for modeling due to the limit of available traffic data sources. Existing studies mainly exploited cross-sectional speed data from loop detectors or microwave sensors for such purpose. Consequently, the speed variance was described as the point speed variance during a certain period (Xu et al., 2012, 2013; Shi et al., 2016a,b; Yu et al., 2016, 2017; Basso et al., 2018) or speed difference between two consecutive cross-sections (Hossain and Muromachi, 2012; Xu et al., 2012; Hossain and Muromachi, 2013b; Xu et al., 2013; Wang et al., 2015a; Yu et al., 2016). The former measures the speed temporal variance at a specific site with equipped sensors, whereas the latter measures the speed spatial variance between two different sites. The accuracy of captured speed spatial variance is affected by the sensor/detector gaps, which are often not small (e.g. 0.5 to 2 miles). Also, many roads may not have the necessary data collection facilities. More importantly, the derived speed variance based on measurements from two adjacent sensors cannot continuously capture the possible variation of speed along the road section. Therefore, alternative solutions are expected to address these issues.

Fortunately, the emerging trajectory data from vehicles with equipped GPS provides a new avenue for analyzing a number of speed related issues in a transportation network. For example, taxi GPS (or FCD) data of Shanghai has already been used as a way to collect road speed information in previous studies (Xie et al., 2013; Wang et al., 2014, 2015b; Wang et al., 2016). In megacities like New York, Beijing, and Shanghai, taxis share a large proportion of traffic flows. This offers an important source to collect the speed data of urban expressways. The use of these taxi data may nevertheless be challenging: (1) there is no elevation information in the GPS dataset; and (2) most urban expressways in cities like Shanghai are elevated above the surface arterials. For example, Mei et al. (2015) conducted a study to deal with this data mixing problem using a semi-supervised learning technique. Although the algorithm worked well under the uncongested condition (yielding an approximately 100% clustering purity), the clustering purity was

relatively low (approximately 80%) under the congested conditions, which is common in rush hours. To facilitate the use the taxi GPS data for quantifying urban expressway speed variance, additional efforts for addressing the aforementioned challenges are expected. This motivates us to tackle the issues and augment the use of the valuable data for safety purpose.

3. Data description

To measure the speed variance and investigate its safety impact on urban expressways while controlling other contributing factors, taxi FCD, loop detector data, road feature data, and crash data were collected from five main urban elevated expressways in Shanghai. The road segments were defined by adjacent ramps. After removing all the segments having at least one type of data missing (e.g., for tunnels, the FCD is missing), the sampled urban expressways finally consisted of 199 segments with a total length of 200.12 km. The study time window was set as the weekday morning rush hour from 8:00 a.m. to 9:00 a.m.

The FCD and loop detector data of five weekdays (from Monday to Friday: 3/16/2009, 3/17/2009, 3/11/2009, 3/26/2009, 3/20/2009) were used to extract the speed variance and other traffic flow features. The FCD were obtained from the Qiangsheng taxi company, a major taxi service provider in Shanghai. The FCD dataset includes the vehicle identification number, time stamp, date, longitude and latitude, and vacant or occupied information. In this research, only the GPS data of the occupied taxis were utilized. The original GPS sampling time interval was 10 s. Owing to some missing GPS points, the average interval was increased to 10.91 s.

Traffic volume is one of the most important parameters when analyzing crash frequency. However, it could not be extracted from the taxi GPS data. Therefore, the loop detector data was acquired to collect traffic volume. In the original loop detector data, vehicle counts were categorized as three types based on the vehicle length (unit: meter): large vehicle (length $\geq$ 10 m), medium vehicle (5 m $\leq$ length $<$ 10 m), and small vehicle (length $<$ 5 m). The original traffic counts were converted into the equivalent passenger car units (pcu) using the following simple rules: 1 large vehicle = 2 pcu, 1 medium vehicle = 1.5 pcu, and 1 small vehicle = 1 pcu.

The crash records of the sampled segments were obtained from the crash database of Shanghai. The information concerning the time, location, type of crash, and severity was extracted for each record. Regarding the crash severity, only ten crashes involved fatal or serious injury on the urban expressways in 2009, and the remaining ones were PDO crashes. Thus, the injury and fatal crashes were excluded from consideration. Accordingly, the PDO crash frequency was calculated for each segment. For road segment features, the segment length, ramp type, and number of lanes were also collected.

Both FCD data and traffic volume data used one-week (five different weekdays) data. For FCD data, the manual work to identify on-expressway GPS was time-consuming and using one-year FCD data was impossible based on current data process method. Another limitation was for both FCD data and traffic volume data in our dataset, only few days were available. So, we selected five different weekdays with both FCD and traffic volume. In terms of the week periodicity of the traffic (Dendrinos, 1994), we used the average value of the factors extracted from five different weekdays. However, for crash frequency, it would be too rare when only use crashes happened during five weekdays. But if considering one-year crash data, it's unreasonable to use five-day traffic data to represent that one year situation. Therefore, as a balance, crashes at 8:00 a.m. – 9:00 a.m. from March to June were filtered. January and February were excluded because their average temperature were below 0 C, while that of March to June were close to each other, and above 0 C. Then to avoid the mixing of working and non-working days, crashes on holidays and vacations were eliminated. Finally, 347 PDO crashes in total were included in the final crash data set.

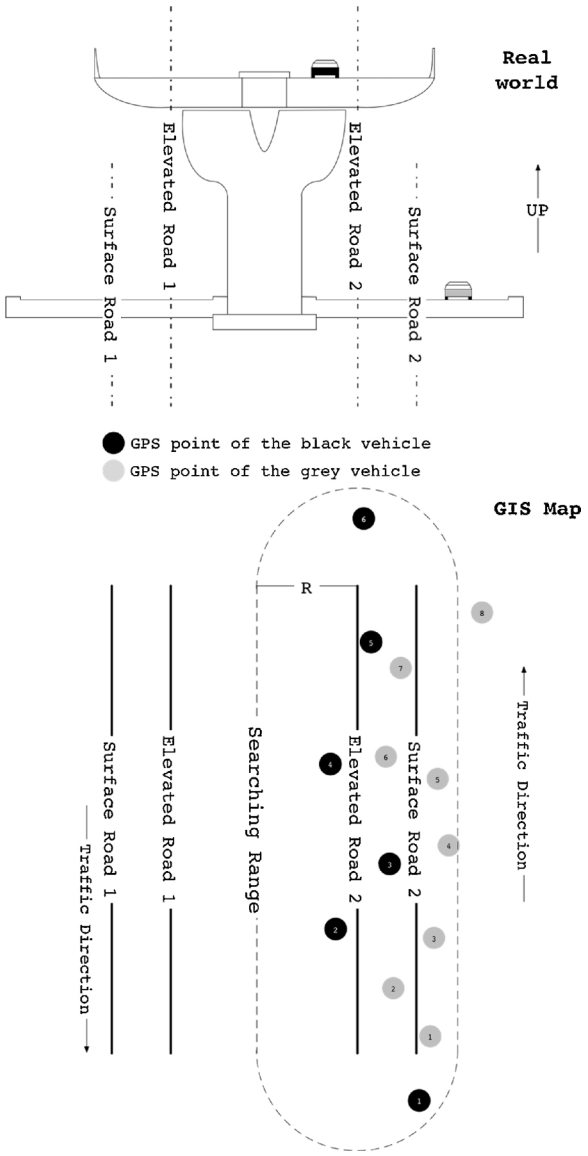

Fig. 1. GPS mixing problem attributing to multiple-layered roads.

4. Data processing

4.1. Elevated expressway GPS data filtration

Most urban expressway segments in the study area are elevated, with surface arterials below. Due to the missing of the altitude information, GPS data collected on the urban elevated expressways and the surface roads below are mixed together. Fig. 1 illustrates this GPS mixing problem. In the figure, the black car is on the elevated expressway, whereas the gray car is on the surface arterial under it. The GPS trajectory points of each vehicle do not have any attribute that reflect the elevation information. Users cannot easily distinguish the trajectory of the black vehicle from the other one when their trajectories are projected on a map. In order to extract the trajectories on the elevated expressways, a filtering effort is needed.

In this paper, we implemented a semi-automatic filtering process to extract the needed trajectory data. This process involves two stages of tasks. Firstly, all trips that include the GPS points along the elevated expressways were extracted and visualized with a Geographic Information System (GIS) map of the road network. Subsequently, the first and last points of the portion that overlap with the elevated expressways were manually labeled. Finally, all the GPS points between

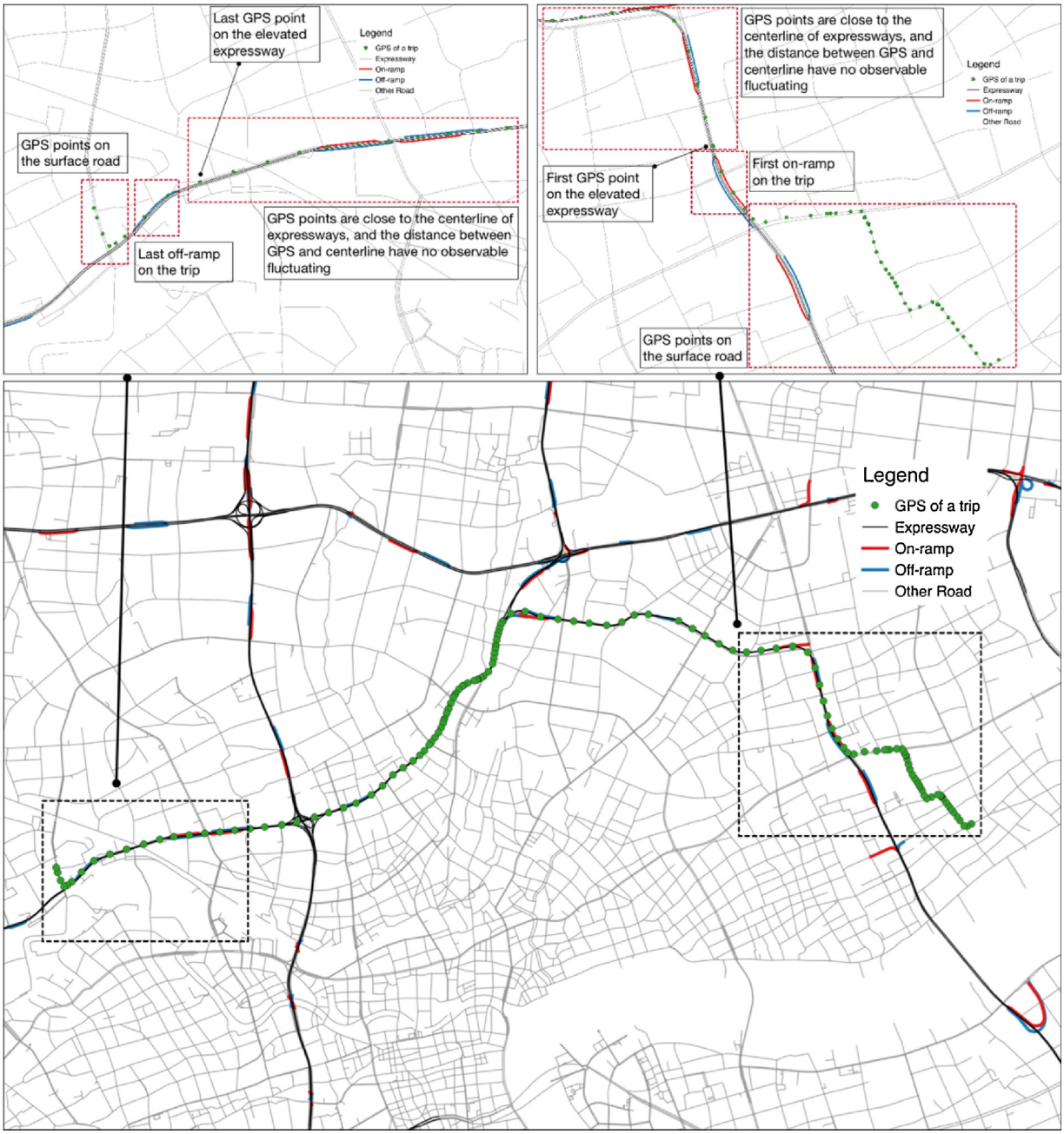

Fig. 2. Labeling the first and last on-expressway GPS points of a typical trip.

these two identified points were labeled according to time stamp (see Fig. 2 for an illustration). The detailed steps are as follows:

(1) All the occupied trips that were completed in the study period are extracted and labeled with the trip id.
(2) A 20-m lateral buffer is generated on the urban expressways in each studied direction.
(3) The GPS points inside the buffer are counted for each trip and the trips that have N ≤ 10 counted GPS points are removed.
(4) Shape files are generated for all the trips and visualized on a GIS map sequentially. The first and last points that seem to be on the elevated expressway are labeled based on the rules defined later in this section.
(5) For each trip, the points between the first and last points identified in step (4) as the elevated expressway GPS points are also labeled.

The rules to identify the start and end points on the elevated expressway are based on the visual verification of the trajectory, which are presented as below.

(1) Direction rule: For a given travel direction, the vehicle should maintain its path on the right side of the road.
(2) Ramp rule: for the on-expressway GPS points of each trip, the first and last point is assumed to be close to the first entrance ramp and last exit ramp along the path of the vehicle, respectively.
(3) Drift rule: the GPS points on and under the elevated expressway are more likely to be close to and far away from the expressway centerline, respectively (Fig. 3).

Based on aforementioned procedure, the candidate trajectories of over 10,000 trips were reviewed. In summary, 4147 trips were filtered and 244,734 GPS points were labeled as on-expressway points.

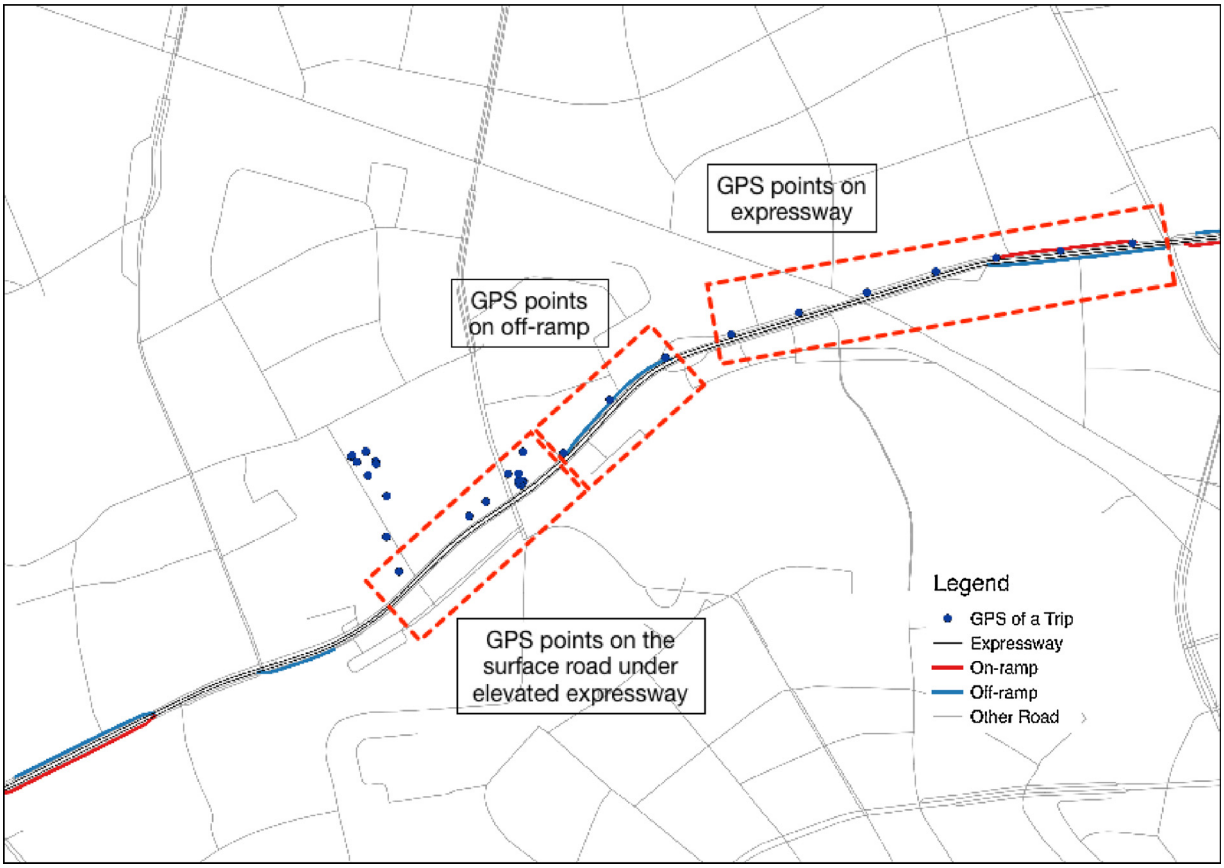

Fig. 3. The difference between GPS points on and under expressway.

4.2. Speed variance extraction

Following the identification of on-expressway GPS points, the speed calculation was implemented based on the sequence of adjacent points from the same vehicle. The basic concept is that the average speed between two adjacent GPS points equals the route length difference divided by the time difference. Subsequently, the erroneous speed values (out of the range from 0 to 120 km/h) were removed.

These between-GPS-pair speed estimates were then used for computing the speed variance. To measure the speed variance, a series of fixed-interval cross-sections were generated for each road segment. The first cross-section was located at the starting point of a road segment and the interval between two adjacent cross-sections is 10 m (the average number of cross-sections per segment is 100.6). To generalize the GPS speed characteristics of each cross-section, the instantaneous speed estimates associated with the GPS points passing through a given cross-section were aggregated to obtain the mean speed and speed standard deviation (SD) for that short section. Fig. 4 presents the visualization of the cross-sectional average speed and speed SD for a series of consecutive segments on the Yan'an expressway.

Further, two parameters, which were the SD of the cross-sectional speed mean (SDCSM) and the mean of the cross-sectional speed SD (MCSSD), were derived to measure the speed variance of a road segment. Mathematically, these two parameters are calculated based on the equations below and in conjunction with the illustration of Fig. 5.

Definition 1. (The pair of GPS Points). The pair of GPS points (POGP) is defined as two GPS points from the same vehicle, which are adjacent in the order of timestamp.

The POGP is the minimum information unit of GPS speed. To get SDCSM and MCSSD, we need to calculate the speed of each POGP in the database.

First, the milepost and time of a GPS point are denoted by $(x_{j,k}, t_{j,k})$, which represent the milepost and time of the k_{th} GPS point (in ascending order of timestamp) of the j_{th} vehicle, respectively. Then, a POGP can be written as $[(x_{j,k}, t_{j,k}), (x_{j,k+1}, t_{j,k+1})]$, while j > 0, k > 0. The speed of this POGP is calculated as

$$v_{jk} = (x_{j,k+1} - x_{j,k})/(t_{j,k+1} - t_{j,k}) \tag{1}$$

Definition 2. (The Set of POGP Related to a Cross-section). If one POGP covered a cross-section, then this POGP is related to the cross-section. The set of POGP related to a cross-section includes all the POGP related to this cross-section.

Mathematically, let's assume that a series of cross-sections are generated on a road segment, the milepost of the i_{th} cross-section is m_i. The set of POGP related to the i_{th} cross-section P_i can be represent as:

$$P_i = \{[(x_{j,k},\ t_{j,k}),\ (x_{j,k+1},\ t_{j,k+1})] | x_{j,k} \le m_i \le x_{j,k+1}\} \tag{2}$$

With Definition 2, we can calculate the mean and SD of GPS speed related to a cross-section *i*.

$$\bar{V}_i = \frac{\sum_{j=1}^{N_i} (x_{j,k+1} - x_{j,k})/(t_{j,k+1} - t_{j,k})}{N_i} \tag{3}$$

$$\mathrm{SD}(V_i) = \frac{\sqrt{\sum_{j=1}^{N_i} [(x_{j,k+1} - x_{j,k})/(t_{j,k+1} - t_{j,k}) - \bar{V}_i]^2}}{N_i - 1} \tag{4}$$

while $[(x_{j,k},\ t_{j,k}),\ (x_{j,k+1},\ t_{j,k+1})] \in P_i$, N_i is the total number of POGP in the set of POGP Related to a Cross-section *i*.

Definition 3. (The Set of Fixed-interval Cross-sections inside a Segment) Assume the starting and ending mileposts of segment *z* are S_z and E_z, respectively, the fixed interval is L (in our case, L is 10 m). Then the set of fixed-interval cross-sections inside segment *z* can be

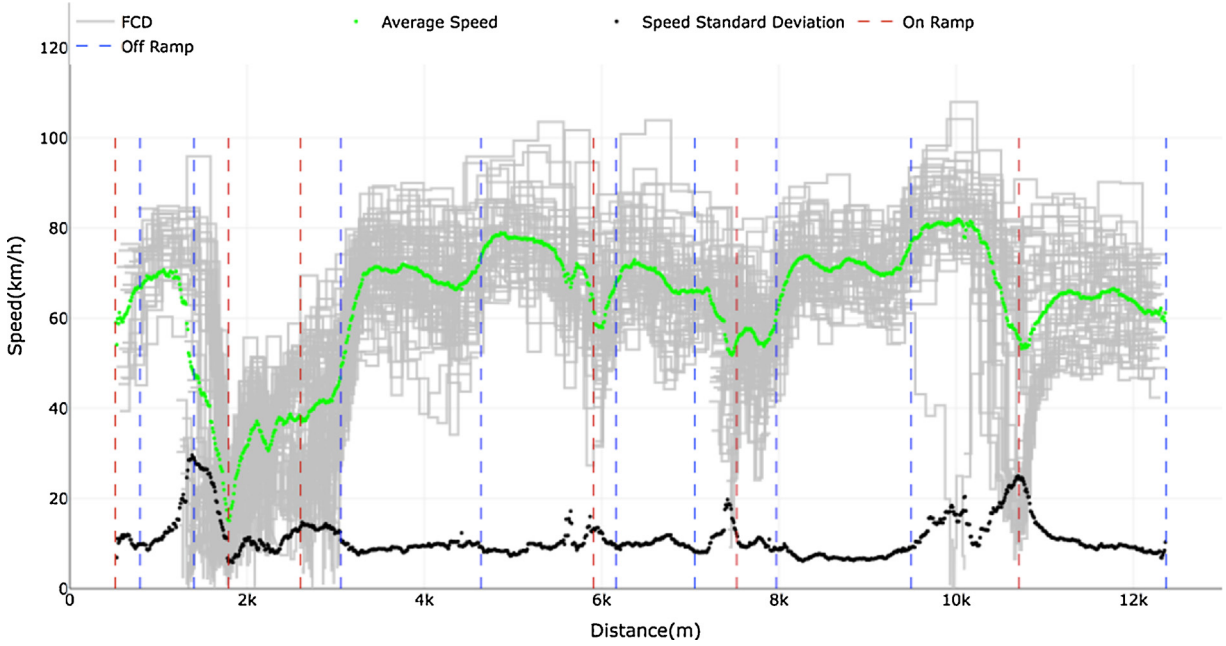

Fig. 4. Cross-sectional average speed and speed standard deviation.

denoted as:

$$C_z = \{m_i | m_i = S_z + \mathrm{L}\times(\ \mathrm{i}-1)\ \text{and}\ m_i \le E_z\} \tag{5}$$

while m_i is the milepost of a cross-section, i is the order number of a cross-section, and it begins from 1. The milepost of the first cross-section is always S_z. The maximum i can be calculated by:

$$\mathrm{I} = \max(i) = \lfloor (E_z - S_z)/\mathrm{L} \rfloor + 1 \tag{6}$$

For any cross-section $m_i \in C_z$, we can calculate the $\bar{V}_i$ and $\mathrm{SD}(V_i)$ based on Eqs. (3) and (4).

Finally, the SDCSM for segment z can be calculated as

$$SDCSM_z = \sqrt{\frac{\sum_{i=1}^{I}(\bar{V}_i - \frac{\sum_{i=1}^{I}\bar{V}_i}{I})^2}{\mathrm{I}-1}} \tag{7}$$

and the MCSSD can be calculated by Eq. (8)

$$MCSSD_z = \frac{\sum_{i=1}^{I}\mathrm{SD}(V_i)}{\mathrm{I}} \tag{8}$$

SDCSM was used as an indicator of the spatial speed variance. MCSSD was used as a temporal speed variance factor which measuring the overall temporal speed variance during the analysis period. The SDCSM and MCSSD are calculated for the period 8:00 a.m. – 9:00 a.m. of each weekday at first. Then, the five-day averaged SDCSM and MCSSD are used in the final analysis. The average FCD sample size for each segment is 122.4 (SD: 78.4). The summary of the final dataset is presented in Table 1.

5. Statistical models

Conventionally, the equivalence of the variance to the mean is the limitation of the Poisson model, which is not suitable for crash data because of its over-dispersed characteristic. This study proposed to use a Poisson-gamma model for the crash counts associated with urban elevated expressways. An independent error term, ε_{ik}, was added when estimating the expected crash frequency at each site. The overall Poisson-gamma model is shown as follows:

$$Y_{ik} \sim \mathrm{Poisson}(\lambda_{ik}) \tag{9}$$

$$\ln\lambda_{ik} = \beta_0 + \sum_{p=1}^{P}\beta_p X_{pik} + \varepsilon_{ik} \tag{10}$$

where Y_{ik} is the frequency of crashes on expressway i and road segment k;

λ_{ik} is the expected crash frequency;

X_{pik} represents the explanatory variables;

β_{pik} represents the regression coefficients to be estimated (p = 0, 1, …P);

P is the number of explanatory variables; and

ε_{ik} is the error term, and $\exp(\varepsilon_{ik})$ is assumed to follow a gamma distribution with mean and variance σ_ε^2.

However, the independence assumption of the Poisson-gamma model could be violated by the possible heterogeneity of the crash data. For different expressways, the rural/urban location, traffic volume on the ramp, traffic composition, nearby land use, and road side objects may differ. Many unobserved factors could directly affect the crash frequency. To account for the potential heterogeneity across homogeneous groups, the random effect hierarchical model was considered. The hierarchical model can be specified as

$$\ln\lambda_{ik} = \beta_{0i} + \sum_{p=1}^{P}\beta_p X_{pik} + \mu_i + \varepsilon_{ik} \tag{11}$$

where μ_i is the random effect across the expressways, following a normal distribution with $\mu_i \sim N(0, \sigma_\rho^2)$.

In addition, three traffic characteristic factors (average speed, SDCSM, and MCSSD) were extracted from the GPS data. Although the manual labeling of expressway GPS points was carefully designed, some surface road GPS points still could be mistakenly labeled as on-expressway points. To consider the deficiencies in the measurements, measurement error terms (Yang et al., 2013; Xie et al., 2015) of the average speed, SDCSM, and MCSSD were introduced into the model.

$$\mathrm{X}^*_{qik} = \mathrm{X}_{qik} + \tau_{ik} \tag{12}$$

where X^*_{qik} denotes the explanatory variables extracted from the GPS data on expressway i and segment k, and τ_{ik} is the measurement error term. It is assumed that τ_{ik} follows a normal distribution, $N(0, \sigma_q^2)$

$$\mathrm{X}^*_{qik} \sim N(\mathrm{X}_{qik}, \sigma_q^2) \tag{13}$$

where σ_q^2 is the variation parameter for the q_{th} explanatory variable.

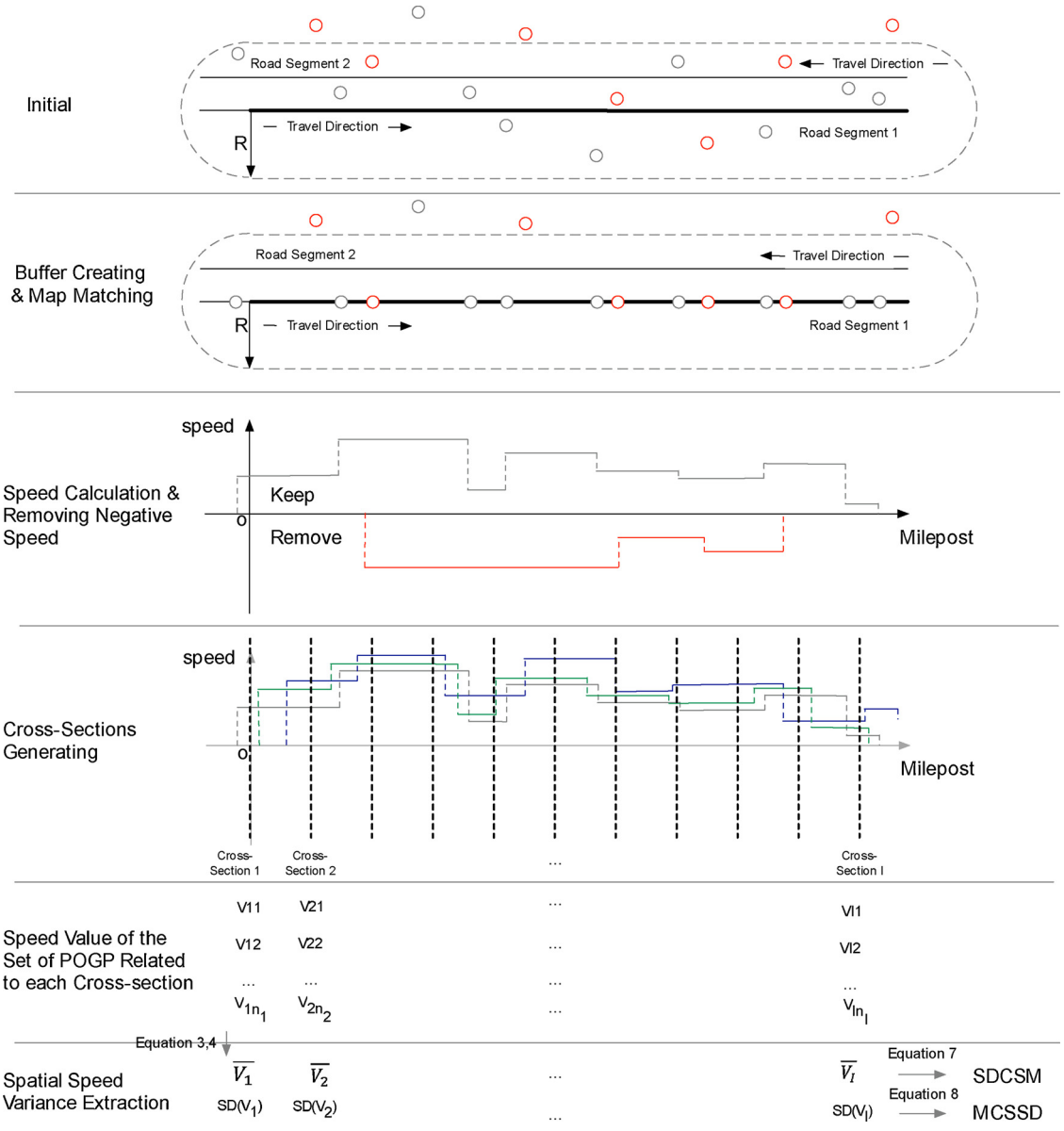

Fig. 5. The process of speed variance indicators extraction.

Table 1
Summary Statistics of the Variables.

Type	Variables	Description	Mean	S.D	Min	Max
Response Variable	Crash	Frequency of all PDO crashes per segment	1.44	2.10	0.00	12.00
Traffic Characteristic	Average speed	Average speed for the segment(km/h)	59.5	16.4	17.2	86.8
	SDCSM	SD of cross-section speed mean	4.52	3.58	0.38	25.78
	MCSSD	Mean of cross-section speed SD	13.91	5.46	5.78	39.74
	TV	Hourly traffic volume (pcu/h)	3778	1374	367	7868
Road Features	Expressway	Which expressway the segment belongs to: 1-Yan'an expressway, 2-North-South expressway, 3-Inner ring, 4-Midium ring, 5-Yixian expressway	2.95	1.12	1.00	5.00
	SL	Length of the road segment (meter)	1006	591	117	2899
	Number of lanes	Number of lanes for the road segment	3.22	0.94	2.00	5.00
	Ramp type	The upstream and downstream ramp type: 1-upstream off, downstream off 2-upstream on, downstream off 3-upstream off, downstream on 4-upstream on, downstream on	2.45	0.86	1.00	4.00

Replacing measured explanatory variables extracted from the GPS data X_{qik} with real values of these explanatory variables X^{*}_{qik}, the mean function can be rewritten as follows:

$$\ln\lambda_{ik} = \beta_{0i} + \sum_{p=1}^{P} \beta_p X_{pik} + \sum_{q=1}^{Q} \beta_q X^{*}_{qik} + \mu_i + \varepsilon_{ik} \quad (14)$$

where Q is the number of explanatory variables extracted from GPS data and P is the number of other explanatory variables.

The deviance information criterion (DIC) is used as a Bayesian measure of the model fitting and complexity. In Monte Carlo simulations, the use of DIC is more convenient than using the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The DIC is calculated as follows:

$$DIC = D\bar{(}\beta) + p_D \quad (15)$$

where $D(\beta)$ is the Bayesian deviance of estimated parameter β, $D\bar{(}\beta)$ is the posterior mean of $D(\beta)$, and p_D is the effective number of parameters and can be viewed as an indicator to model complexity.

The proposed hierarchical Poisson-gamma model with ME and the non-hierarchical model (as a comparison) were developed within a full Bayesian approach and all the parameters were estimated using the Monte Carlo Markov chain (MCMC). The MCMC algorithm is a classic method to use independent and identically distributed simulations of a random process to approximate a desired distribution. Once the distribution of the simulated chains was observed to converge to the target posterior distribution, the complete Bayesian estimates of the model parameters were obtained from the remaining iterations. The multivariate potential scale reduction factor (MPSRF) proposed by Brooks and Gelman (1998) was used to test the convergence of the chains.

In this paper, WinBUGS software and two R packages (R2WinBUGS package, CODA package) were used for the Bayesian model estimation. The settings of both the hierarchical and non-hierarchical models were identical. With no credible prior information, uninformative priors were assumed. For the regression coefficients for all the parameters, a normal distribution with zero mean and a relatively large variance (10^5) was assumed, which has been frequently used in prior distributions for the regression parameters. Gamma (0.001,0.001) was applied to parameters $\sigma_{\varepsilon}^{-2}$, σ_{ρ}^{-2}, and σ_{q}^{-2}. Two independent chains with different initial values were generated, and 150,000 iterations were set up. The first 50,000 samples were discarded as burn-ins, and 100,000 iterations were used to obtain the model results. The thin is 2. After examining the MPSRF values, which were 1.05 (hierarchical model) and 1.21 (non-hierarchical model), the chains were identified as converged in both models.

6. Results and discussion

The results of the hierarchical and non-hierarchical Poisson-gamma models with ME are presented in Tables 2 and 3, respectively. For the Bayesian model, the 95% Bayesian credible interval (BCI) was used to test the significance of a variable in the model. If the signs of the upper bound and lower bound were the same, the variable was considered to be significant. In this study, 95% BCI (2.5% BCI–97.5% BCI) was used to examine the significance of the estimates. After testing all the variables in the hierarchical model, only the significant variables were retained. To conduct effective comparisons, all the explanatory variables included in the non-hierarchical model were also used in the hierarchical model.

Table 2
Estimation results of hierarchical Poisson-gamma model with ME.

Variables	Mean	S.D.	2.50% BCI	97.50% BCI
Constant	−12.5063	2.6944	−18.2900	−7.8320
Log(SL)	0.4877	0.1790	0.1407	0.8331
SDCSM	0.0843	0.0295	0.0267	0.1421
MCSSD	0.0364	0.0187	0.0008	0.0490
Log(TV)	1.0430	0.2855	0.5657	1.6110
$\sigma_{q\ SDCSM}$	4.3879	1.6084	3.5610	9.1450
$\sigma_{q\ MCSSD}$	8.0752	6.1894	5.4440	23.2103
σ_{ρ}	0.4234	0.1317	0.1497	0.6802
DIC	554.4	–	–	–

Table 3
Estimation results of non-hierarchical Poisson-gamma model with ME.

Variables	Mean	S.D.	2.50% BCI	97.50% BCI
Constant	−10.9838	2.1894	−14.7700	−7.0550
Log(SL)	0.3717	0.1696	0.0572	0.7040
SDCSM	0.0996	0.0294	0.0456	0.1601
MCSSD	0.0461	0.0186	0.0095	0.0820
Log(TV)	0.8775	0.2534	0.3559	1.3630
$\sigma_{q\ SDCSM}$	8.6017	4.5828	3.5720	19.2400
$\sigma_{q\ MCSSD}$	11.9893	11.7607	5.4520	40.2303
DIC	591.8			

The results show that in both the models, the Log(SL), SDCSM, MCSSD, and Log(TV) are positively related to the PDO crash frequency. The DIC of the hierarchical model (554.4) is smaller than that of the non-hierarchical model (591.8), which suggests that the hierarchical model outperforms the non-hierarchical model by including random effect that varies among different expressways. In the hierarchical model, the estimated SD of random effects is significantly positive.

The segment length and traffic volume are often recognized as important factors for the link-based crash frequency. In the Highway Safety Manual (American Association of State Highway and Transportation Officials, 2010), the safety performance functions (SPFs) of most of the segments are developed only based on these two factors, which are relatively easy to measure. For example, the SPF for divided roadway segments on rural multilane highways is expressed in Eq. (16),

$$N_{spf\ rd} = e^{(a+b\times\ln(AADT)+\ln(L))} \quad (16)$$

where N_{spfrd} is the base of the total roadway segment crash frequency per year, AADT is the annual average daily traffic (vehicle/day) on the roadway segment, L is the length of the roadway segment, and a and b (usually positive) are the regression coefficients. Based on Eq. (16), the segment length and traffic volume are positively related to the crash frequency on divided rural multilane highways. In this study, the same statistics inference can be made based on the modeling results. Longer segments may increase the exposure to crash risk, and a high traffic volume may not only directly provide additional exposure but also indirectly be associated with more conflicts. Therefore, as the most important control variables, the effects of the segment length and traffic volume are expected to be correctly estimated.

SDCSM is the key indicator developed to measure the spatial speed variance. In the hierarchical model, the estimated SDCSM is 0.0843, which implies that one unit increase in the SDCSM predicts a 8.80% ($e^{0.0843}$-1) increase in the PDO crash frequency. SDCSM is defined as the SD of the cross-sectional speed mean, and it measures the average speed fluctuation along a road segment. Considering an ideal situation, if the average speed everywhere on a road segment is constant, the SDCSM will be equal to 0. If the average speed changes along the road, the SDCSM is positive. The Solomon curve (Solomon, 1964) shows the safest traveling speed on a road is the average speed. Maintaining a constant average speed curve (SDCSM = 0) is less difficult than adjusting to a varying average speed curve (SDCSM > 0). Moreover, intuitively, such a difficulty increases with the extent of the spatial speed variance. Failing to follow the spatial variance of the traffic speed may lead to crashes, and this could be the reason that the SDCSM is positively related to the crash frequency.

The MCSSD extracted from the FCD is introduced as an explanatory factor to measure the temporal speed variance. The estimated coefficient of the MCSSD in the hierarchical model is 0.0364, which implies

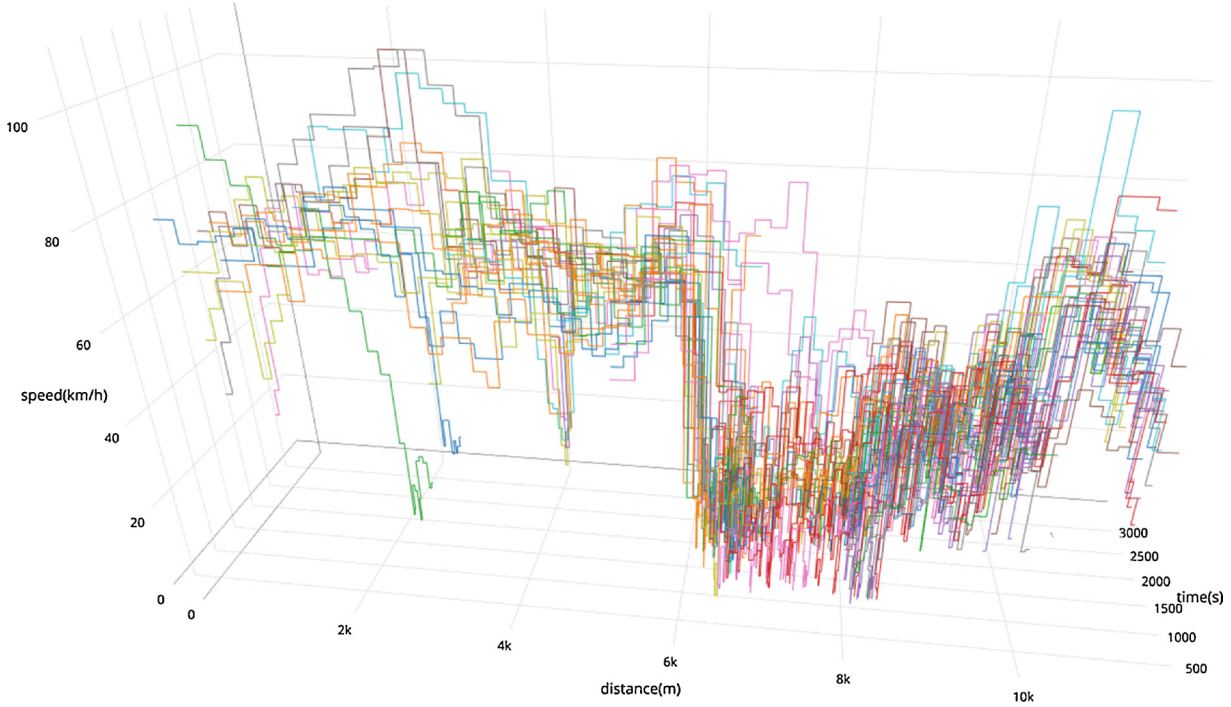

Fig. 6. Time-space-speed 3D graph example of FCD for successive road segments.

that one unit increase in the MCSSD leads to 3.71% ($e^{0.0364}$-1) increase in the PDO crash frequency. The temporal speed variance produced by two main sources, including the driver's individual speed differences and the traffic condition shifting. When the speed selection difference increased, especially for adjacent vehicles, the number of potential conflicts may increase. Meanwhile, during the traffic condition shifting, the speed of the traffic flow may change and the potential conflicts may increase as well. Thus, when the MCSSD increases, the temporal speed variance for each cross-section increases and the crash risk tends to be higher. In order to observe temporal speed variance, the time-space-speed 3D graph was drawn (see Fig. 6). Each line represented a certain vehicle passed the road, and different vehicles were distinguished by different colors. It can be observed that at the same location (a certain value in distance axle), the speeds of vehicles on different time are not similar. At some location, for example, distance equals to 2,000, the temporal variation of speed is large. While at some other location, for example, distance equals to 10,000, the temporal variation of speed is small.

To address the measurement errors, the normal distribution error terms for both the SDCSM and MCSSD were included in the model. In the hierarchical model, the variance parameters $\sigma_{q\,SDCSM}$ and $\sigma_{q\,MCSSD}$ were estimated to be 4.3879 and 8.0752, respectively, and their 95% BCIs both did not cover zero, which implies that they are both significant. This suggests the existence of measurement errors in the SDCSM and MCSSD, and the measurement error models can help account for a biased estimation.

7. Conclusion

In this paper, we provided a feasible approach to capture the speed variance on urban elevated expressways and then examine its effect on safety. The taxi FCD was used to extract the speed variance. To avoid incorrectly selecting the GPS points on the surface road under the expressways, a semi-automatic process for identifying on-expressway GPS points was introduced. The FCD speed with temporal and spatial information was collected. The SDCSM and MCSSD were developed and obtained to indicate the speed variance. For 199 urban expressway segments in Shanghai, the PDO crash frequency and factors such as the road parameters (number of lanes, ramp type, and segment length) and traffic flow parameters (traffic volume and speed) were collected.

To account for the consideration of potential measurement errors associated with SDCSM and MCSSD, and the potential heterogeneity across homogeneous groups, and the random effect hierarchical Poisson-gamma model was used, and the non-hierarchical model was also established for comparison. The hierarchical Poisson-gamma model was found to outperform the non-hierarchical model. The SDCSM was found to be significant and positively related to the PDO crash frequency, which implies that a larger spatial and temporal speed variance increase the probability of crashes on an urban expressway. Segment length and traffic volume were also determined to be significant and positively related to the PDO crash frequency.

The FCD provides new opportunities for obtaining a detailed speed profile of urban expressways. For future study, enhancement of the manual identification of the urban expressway data by an automatic process is highly suggested for processing large-scale probe data. It also deserves to incorporate spatial speed variance into modeling real-time crash risk. Some related data such as lane width, roadside features data, if available, should also be considered in future endeavors.

Acknowledgements

This study was jointly sponsored by the Chinese National Science Foundation (51138003, 61703352) and Fundamental Research Funds for the Central Universities (2682017CX021).

References

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. Trans. Res. Rec.: J. Trans. Res. Board 1897, 88–95.

American Association of State Highway and Transportation Officials, 2010. Highway Safety Manual. American Association of State Highway and Transportation Officials, USA.

Basso, F., Basso, L.J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transp. Res. Part C Emerg. Technol. 86, 202–219.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7 (4), 434–455.

Dendrinos, D.S., 1994. Urban traffic flows and fourier transforms. Geogr. Anal. 26 (3), 261–281.
Hossain, M., Muromachi, Y., 2012. A bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.
Hossain, M., Muromachi, Y., 2013a. A real-time crash prediction model for the ramp vicinities of urban expressways. IATSS Res. 37 (1), 68–79.
Hossain, M., Muromachi, Y., 2013b. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accid. Anal. Prev. 57, 17–29.
Kockelman, K.K., Ma, J., 2010. Year. Freeway speeds and speed variations preceding crashes, within and across lanes. Proceedings of the Journal of the Transportation Research Forum.
Kononov, J., Bailey, B., Allery, B., 2008. Relationships between safety and both congestion and number of lanes on urban freeways. Trans. Res. Rec.: J. Trans. Res. Board 2083, 26–39.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44 (5), 291–305.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Mei, Y., Tang, K., Li, K., 2015. Real-time identification of probe vehicle trajectories in the mixed traffic corridor. Transp. Res. Part C Emerg. Technol. 57, 55–67.
Qu, X., Meng, Q., 2014. A note on hotspot identification for urban expressways. Saf. Sci. 66, 87–91.
Qu, X., Yang, Y., Liu, Z., Jin, S., Weng, J., 2014. Potential crash risks of expressway on-ramps and off-ramps: A case study in beijing, china. Saf. Sci. 70, 58–62.
Quddus, M., 2013. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and gis. J. Transp. Saf. Secur. 5 (1), 27–45.
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 Emerg. Technol. 58 (Part B), 380–394.
Shi, Q., Abdel-Aty, M., Lee, J., 2016a. A bayesian ridge regression analysis of congestion's impact on urban expressway safety. Accid. Anal. Prev. 88, 124–137.
Shi, Q., Abdel-Aty, M., Yu, R., 2016b. Multi-level bayesian safety analysis with unprocessed automatic vehicle identification data for an urban expressway. Accid. Anal. Prev. 88, 68–76.
Solomon, D., 1964. Accidents on Main Rural Highways Related to Speed, Driver, and Vehicle.
Sun, J., Sun, J., 2015. A dynamic bayesian network model for real-time crash prediction using traffic speed conditions data. Transp. Res. Part C Emerg. Technol. 54, 176–186.
Sun, J., Li, T., Li, F., Chen, F., 2016. Analysis of safety factors for urban expressways considering the effect of congestion in shanghai, china. Accid. Anal. Prev. 95 (Part B), 503–511.
Suzuki, K., Imada, K., Matsumura, Y., 2016. A study of collision risk estimation and users evaluation at merging section of urban expressway in japan. Transp. Res. Procedia 15, 783–793.
Transportation Research Board, 2010. Highway Capacity Manual. USA: Washington, DC. .
Wang, X., Liu, H., Yu, R., Deng, B., Chen, X., Wu, B., 2014. Exploring operating speeds on urban arterials using floating car data: case study in shanghai. J. Transp. Eng. 140 (9), 04014044.
Wang, L., Abdel-Aty, M., Shi, Q., Park, J., 2015a. Real-time crash prediction for expressway weaving segments. Transp. Res. Part C Emerg. Technol. 61, 1–10.
Wang, X., Fan, T., Chen, M., Deng, B., Wu, B., Tremont, P., 2015b. Safety modeling of urban arterials in shanghai, china. Accid. Anal. Prev. 83, 57–66.
Wang, X., Fan, T., Li, W., Yu, R., Bullock, D., Wu, B., Tremont, P., 2016. Speed variation during peak and off-peak hours on urban arterials in shanghai. Transp. Res. Part C Emerg. Technol. 67, 84–94.
Wu, L., Sun, J., 2015. Relationship of Lane Width to Safety for Urban Expressways.
Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in shanghai, china using bayesian hierarchical models. Accid. Anal. Prev. 50, 25–33.
Xie, K., Ozbay, K., Yang, H., Holguín-Veras, J., Morgul, E.F., 2015. Modeling safety impacts of off-hour delivery programs in urban areas. Trans. Res. Rec.: J. Trans. Res. Board 2478, 19–27.
Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47, 162–171.
Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39.
Yang, H., Ozbay, K., Ozturk, O., Yildirimoglu, M., 2013. Modeling work zone crash frequency by quantifying measurement errors in work zone length. Accid. Anal. Prev. 55, 192–201.
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 (Part B), 495–502.
Yu, R., Wang, X., Abdel-Aty, M., 2017. A hybrid latent class analysis modeling approach to analyze urban expressway crash risk. Accid. Anal. Prev. 101, 37–43.
Yun, M., Zhao, J., Zhao, J., Weng, X., Yang, X., 2017. Impact of in-vehicle navigation information on lane-change behavior in urban expressway diverge segments. Accid. Anal. Prev. 106, 53–66.
Zhang, J., Chen, X., Tu, Y., 2013. Environmental and traffic effects on incident frequency occurred on urban expressways. Procedia - Soc. Behav. Sci. 96, 1366–1377.