Transportation Research Part C 54 (2015) 176–186

Contents lists available at ScienceDirect

Transportation Research Part C

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

A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data

Jie Sun, Jian Sun *

Department of Traffic Engineering & Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form 4 March 2015
Accepted 4 March 2015
Available online 1 April 2015

Keywords:
Urban expressway
Real-time crash prediction
Dynamic Bayesian network
Traffic states
Speed conditions data

ABSTRACT

Traffic crashes occurring on freeways/expressways are considered to relate closely to previous traffic conditions, which are time-varying. Meanwhile, most studies use volume/occupancy/speed parameters to predict the likelihood of crashes, which are invalid for roads where the traffic conditions are estimated using speed data extracted from sampled floating cars or smart phones. Therefore, a dynamic Bayesian network (DBN) model of time sequence traffic data has been proposed to investigate the relationship between crash occurrence and dynamic speed condition data. Moreover, the traffic conditions near the crash site were identified as several state combinations according to the level of congestion and included in the DBN model. Based on 551 crashes and corresponding speed information collected on expressways in Shanghai, China, DBN models were built with time series speed condition data and different state combinations. A comparative analysis of the DBN model using flow detector data and a static Bayesian network model was also conducted. The results show that, with only speed condition data and nine traffic state combinations, the DBN model can achieve a crash prediction accuracy of 76.4% with a false alarm rate of 23.7%. In addition, the results of transferability testing imply that the DBN models are applicable to other similar expressways with 67.0% crash prediction accuracy.

1. Introduction

Traffic data collection has become more convenient and efficient with the development of advanced transportation information systems (ATIS). Thus, a number of studies assessing the real-time risk of traffic flow operation on freeways and urban expressways have been conducted using the traffic data collected from the ATIS, primarily the fixed-point data (e.g., loop detector, microwave radar and automatic vehicle identification (AVI) data). Consequently, numerous proactive crash prediction models have been developed (Oh et al., 2001; Lee et al., 2003; Abdel-Aty et al., 2004, 2005, 2012; Abdel-Aty and Pande, 2005; Pande et al., 2005, 2011, 2012; Pande and Abdel-Aty, 2006; Hossain and Muromachi, 2012, 2013b; Ahmed and Abdel-Aty, 2012; Ahmed et al., 2012; Golob et al., 2008; Zheng et al., 2010; Xu et al., 2013, 2014a, 2014b; Sun et al., 2014a). These models can distinguish crash-prone traffic conditions from normal conditions, which can be applied in safety promotion strategies such as proactive warning with in-car devices or variable message signs (VMS) and other traffic flow smoothing management strategies such as speed harmonization (Lee et al., 2004; Abdel-Aty et al., 2006; Allaby et al., 2007) to avoid crashes or decrease the likelihood of crashes.

* Corresponding author. Tel.: +86 21 69583650.
E-mail address: sunjian@tongji.edu.cn (J. Sun).

http://dx.doi.org/10.1016/j.trc.2015.03.006
0968-090X/© 2015 Elsevier Ltd. All rights reserved.

In general, numerous variables including traffic flow data (i.e., volume, speed and occupancy) and their combinations collected from detectors, road geometry alignment and environmental parameters were used in the development of a real-time crash prediction model. In this paper, regarding roads that lack of fixed point traffic flow detector data, the feasibility of predicting real-time crash risk utilizing only traffic speed condition data, which may be obtained from sampled floating cars or smart phones, was investigated. However, due to the lack of floating car data or smart phone data, the speed data collected from the dual-loop detectors were used as speed condition data instead. If this approach is successful, it can be deduced that the speed data collected via other methods are also effective for crash prediction. Moreover, computational complexity and over-fitting issues caused by redundant variables can be avoided by using relatively fewer speed condition variables.

Although several previous studies have adopted speed-only data collected from AVI sensors or loop detectors for crash prediction (Ahmed and Abdel-Aty, 2012; Li et al., 2013), it still makes sense to consider speed condition data for crash prediction with other concerns, as the traffic state before a crash is an essential factor in developing crash prediction models. It has been indicated that for different combinations of upstream and downstream traffic states, the crash involvement rates and crash risk ratios (ratio of crash cases and non-crash cases) are inconsistent (Yeo et al., 2013; Hossain and Muromachi, 2013a). When different traffic states were considered in previous studies, separated models were usually built under different levels of traffic conditions (Abdel-Aty et al., 2005; Xu et al., 2013, 2014b; Li et al., 2013). However, models employing the value of different traffic state combinations as input parameters might be more efficient and less complex. Thus, two types of state determining approaches were used and compared in this study to identify the better approach.

Regarding the correspondence of traffic data to crash data, there are several time interval data computed with 5-min aggregation that related to the crash occurrence (Pande et al., 2012; Sun et al., 2014a). Generally, only one time interval of traffic data was used in one model for real-time crash prediction, and the time interval (5–10 min before crash) traffic data show the most significant relationship with crashes (Abdel-Aty et al., 2004; Xu et al., 2013, 2014a). However, considering that a crash can be induced by the disturbance of traffic flow before the crash occurs, time series traffic data consisting of several time intervals should be used to illustrate the dynamic process of traffic flow before crash occurrence. Thus, it is essential to establish a single model that can address such time series data and the evolving process of traffic flow.

To address the above issues, a dynamic Bayesian network (DBN) model that can handle time series data was proposed in this study. First, with speed condition data collected in several time intervals, the congestion levels upstream and downstream of the crash location were determined and considered as the explanatory variables of the prediction model. Then, with a matched case-control dataset that includes speed data corresponding to 551 crash cases and 2755 matched non-crash cases collected from two expressways in Shanghai, the DBN models were calibrated and evaluated. Meanwhile, both the DBN models with traffic flow detector data (i.e., volume, speed and occupancy) and a static Bayesian model with speed data were built for comparison purposes. Finally, the transferability of DBN models was also tested to examine the ability to implement it directly on other expressways.

2. Data sources

2.1. Study area and crash data

The real-time crash prediction model established in this study aims to investigate the relationship of traffic flow characteristics and crash risk on urban expressways. Thus, to establish the crash prediction model, crash data and corresponding traffic data should both be collected. In this study, crash data and traffic data were collected from 3 segments on the Yan-an expressway and 3 segments on the North–South expressway in Shanghai, China, which have similar road geometry and on/off-ramp arrangement. Thus, the road geometries are not additional influencing factors on the crash prediction model. All of these sites are three-lane expressway segments, with detectors spaced at approximately 300–500 m. Considering the different quantities of crashes on the two expressways, the 411 crash cases collected on the Yan-an expressway were used for the training and testing of the crash prediction model because of the larger number of crashes which occurred on this road, whereas the 140 crash cases collected on the North–South expressway were used to examine the transferability of the crash prediction model. The sites and crash statistics of segments are presented in Table 1.

Table 1
Summary of study sites and crash data.

Expressway	Segment on expressway	Length/km	Number of detectors	Number of crashes	% of total crashes
Westbound Yan-an expressway	Yandong Interchange to Maoming Road	0.8	4	103	18.7
Eastbound Yan-an expressway	Hongxu Road to Loushanguan Road	1.3	5	202	36.7
	Jiangsu Road to Huashan Road	1.5	6	106	19.2
Northbound North–South expressway	Yanchang Road to Guangzhong Road	1.8	6	28	5.1
Southbound North–South expressway	Guangzhong Road to Luochuang Road	2.0	7	52	9.4
	Gongjiang Road to Changzhong Road	2.2	7	60	10.9
Total		n/a	n/a	551	100

The crash dataset was manually extracted from Shanghai expressway video surveillance systems that recorded accidents between April 2010 and December 2010. We do not use the records from police reports, as there are many missing records. Thus, crash severity information cannot be recorded via the surveillance system. The data on each crash contain information including the occurrence data and time, location, type of crash and weather information. According to the number of involved vehicles, crash data were classified into vehicle break-downs, single-vehicle crashes, two-vehicle crashes and multi-vehicle crashes. Only two-vehicle and multi-vehicle crashes were included in the dataset, while vehicle break-downs and single-vehicle crashes were excluded from this study. Crashes that occurred within one hour of the previous crash at the same location were discarded to avoid the impact of secondary crashes (Yang et al., 2014). A final total of 551 crashes were extracted and used in the study.

2.2. Traffic data

As one purpose of this study is to examine the accuracy of speed condition data for real-time crash prediction, corresponding speed condition data were extracted for each crash case. Regarding the speed condition data, as shown in Fig. 1a, two sections were considered, the upstream section and the downstream section. The average speed data in the upstream and downstream sections of the crash location were collected and aggregated from dual-loop detectors in 5-min intervals. Then, congestion levels of the two sections and speed difference between the two sections were determined and adopted as input variables of the crash prediction models (three variables for model).

For comparison, traffic flow data (i.e., volume, speed and occupancy) were also extracted from the detectors near the crash. With reference to the detector installation interval and crash position, the traffic data were collected from the nearest four detectors with two detectors upstream and two downstream. The four detectors were named D1, D2, D3 and D4, respectively, in order from upstream to downstream, as shown in Fig. 1b. The models built with traffic flow detector data in this study used only the data from detectors D2 and D3 (six variables per model) because it was demonstrated in our previous study that the crash classification results with two-detector data are better than those with four-detector data (Sun et al., 2014a).

As previously mentioned, the evolution of traffic flow resulting in a crash is a dynamic process. Thus, the combination of several time-intervals of traffic data before the crash should be applied in crash prediction. These intervals are as follows: 0–5 min before the crash (time interval 1); 5–10 min prior to the crash (time interval 2); 10–15 min before the crash (time interval 3); and 15–20 min prior to the crash (time interval 4). It should be noted that the data collected in time interval 1 were used as a reference value without any practical use because the traffic management center needs some time for the recognition of the crash and the execution of corresponding measures. Speed condition data and traffic flow detector data were collected within 5–20 min before the crash, which includes three 5-min intervals.

2.3. Non-crash data

When developing the real-time crash prediction model, non-crash cases were also collected to present normal conditions to address the classification issue of distinguishing crash-prone traffic conditions from normal conditions. Similar to crash data, non-crash data were collected under different traffic conditions such as peak time and off-peak time according to the matched case-control design (Abdel-Aty et al., 2004). Thus, for each crash (case) in the dataset, five corresponding non-crash cases (controls) were randomly determined for the same segment and time in the same month, where no crashes occurred within one hour of the original crash time. Then, traffic data for non-crash cases were extracted from the raw dataset, and 2,755 non-crash data were generated. With traffic data supplemented, two case-control datasets including speed condition data and traffic flow detector data corresponding separately to every crash and five randomly selected matched non-crash cases were finally created. Moreover, according to the study by Abdel-Aty et al. (2012) and experiments conducted by the authors, there exist no significant differences about the model performance when changing the ratio of cases and controls.

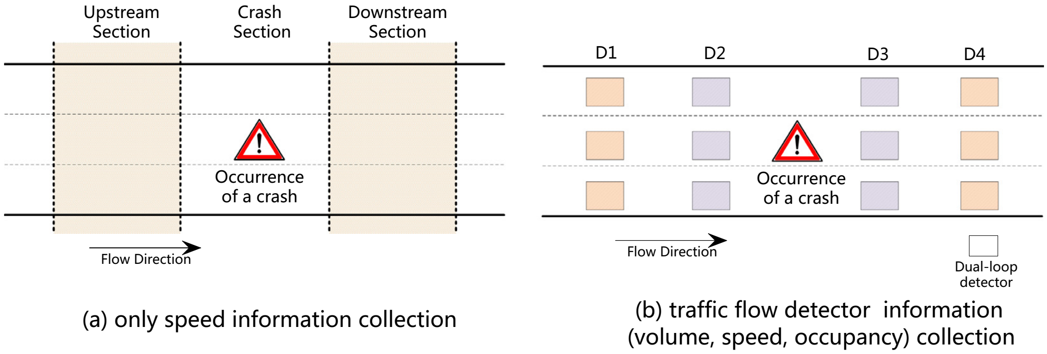


Fig. 1. Scheme of data collection at a crash segment.

3. Relationship of crashes vs. traffic states

3.1. Identification of traffic states

As presented in Section 1, crash occurrence has a close relationship with traffic states, while the occurrence of crashes may be influenced by the interaction of different combinations of traffic states upstream and downstream of the crash site. It is necessary to take the traffic states near the crash location into account when developing the crash prediction model. Then, the first issue that should be addressed is the identification of the upstream and downstream traffic states of the crash site. The upstream and downstream traffic state of the crash site could be identified via the speed data extracted from the corresponding sections. In this study, two methods of determining traffic states were applied according to different congestion levels.

The first one is that two traffic states (i.e., free flow and congestion flow) were identified for both the upstream and the downstream section with a predetermined threshold (Yeo et al., 2013; Li et al., 2013). According to earlier studies by the authors (Sun et al., 2014b; Zheng et al., 2015), 45 km/h is identified as the critical speed between free flow and congestion on expressways in Shanghai based on the analysis of the fundamental diagram as shown in Fig. 2. Furthermore, when analyzing consecutive speed data, 45 km/h is also determined to be the threshold speed associated with abrupt speed drop and breakdown identification at on-ramp bottleneck (Hu et al., 2014). Thus, 45 km/h is considered to be the speed threshold between free flow and congestion in this study. After the two traffic states of the upstream and downstream sections are identified, there exist four outcomes produced by the combinations of the traffic states in the two sections. These states are defined as free flow (FF), congested traffic (CT), bottleneck front (BN), and back of queue (BQ), according to Yeo et al. (2013). Detailed descriptions of the four states are given in Table 2.

In Shanghai, there are three traffic states distributed through the VMS. The three states correspond to the state information displayed on the VMS: green stands for free flow, yellow represents congested flow, and jam flow is labeled in red. The three states can be divided by two speed thresholds. The traffic state is identified as free flow (FF) when the average speed exceeds 45 km/h and jam flow (JF) when the speed is below 20 km/h, while it is congestion (CT) when the speed is between the two thresholds. Thus, nine traffic state combinations were defined with three possible states in the upstream or downstream section, which could be regarded as subdivisions of the above-mentioned four states. The nine states take the following form: AB-XY, where AB/XY takes the value of FF, CT, or JF for the upstream and downstream traffic states, respectively. For instance, the state FF–CT indicates that the traffic upstream is free flow, and downstream traffic is congested. The detailed descriptions of determined traffic states with two different methods list in Table 2.

3.2. Analysis of crash vs. different state combinations

As it was indicated that crash risk ratios (ratio of crashes and non-crashes) differ for different combinations of upstream and downstream traffic states, the crashes and corresponding non-crashes used in this paper related to different traffic states before crash occurrence were analyzed. First, the distributions of the crashes in four states were summarized in Table 3. Most of the crashes (64.2%) in the dataset occurred under the CT state, followed by BQ state crashes of approximately 16.5%. Crashes under the FF state and the BN state were 13.1% and 6.1%, respectively. Additionally, the crash risk ratio of different states are different, as expected: the crash risk ratios in the BN, BQ, and CT states, which include congested regions, are much higher than in the FF traffic state. It is worth noting that the BN state poses the highest risk ratio even though there is the lowest percentage of crash under this state. This result is similar to the results in Hossain and Muromachi (2013a).

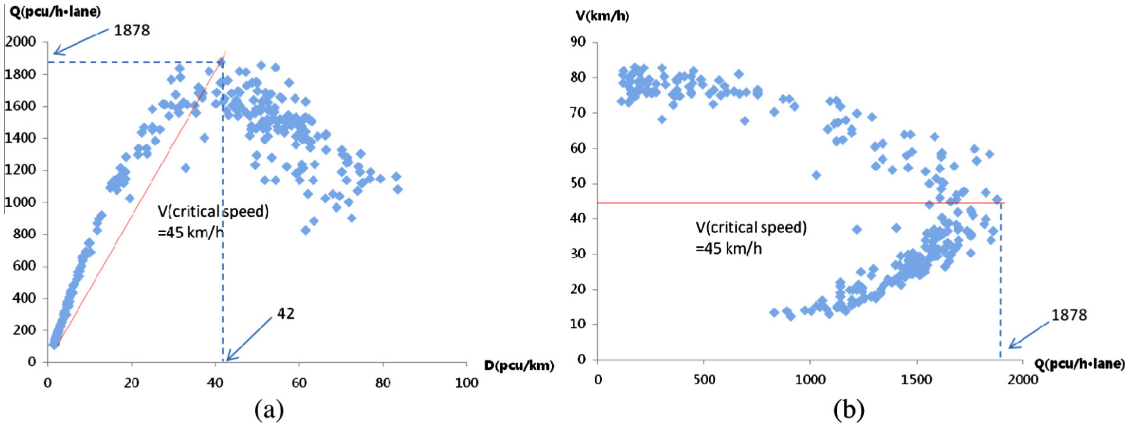


Fig. 2. Diagram of (a) flow-density; (b) speed-flow (Day: July/16/2010; Segment: Hongxu Rd. to Loushanguan Rd. EB, Yan-an Expressway).

Table 2
Description of various traffic states.

	Traffic state combination	Upstream state	Downstream state
Four states	FF	Free flow	Free flow
	BN	Congestion	Free flow
	BQ	Free flow	Congestion
	CT	Congestion	Congestion
Nine states	FF–FF	Free flow	Free flow
	FF–CT	Free flow	Congestion
	FF–JF	Free flow	Jam flow
	CT–FF	Congestion	Free flow
	CT–CT	Congestion	Congested
	CT–JF	Congestion	Jam flow
	JF–FF	Jam flow	Free flow
	JF–CT	Jam flow	Congestion
	JF–JF	Jam flow	Jam flow

Table 3
Distribution of crashes in four states.

Traffic state combination	Crashes	Non-crashes	Crash risk ratio (ratio of crashes and non-crashes)	Percentage of crashes
FF	54	1451	0.04	13.1
BN	25	17	1.47	6.1
BQ	68	223	0.30	16.5
CT	264	415	0.64	64.2

Table 4
Distribution of crashes in nine states.

Traffic state combination	Crashes	Non-crashes	Crash risk ratio(ratio of crashes and non-crashes)	Percentage of crashes
FF–FF	54	1451	0.04	13.1
FF–CT	68	223	0.31	16.5
FF–JF	0	0	0	0
CT–FF	20	14	1.43	4.9
CT–CT	176	203	0.87	42.8
CT–JF	2	58	0.03	0.5
JF–FF	5	3	1.67	1.2
JF–CT	77	109	0.71	18.7
JF–JF	9	74	0.12	2.2

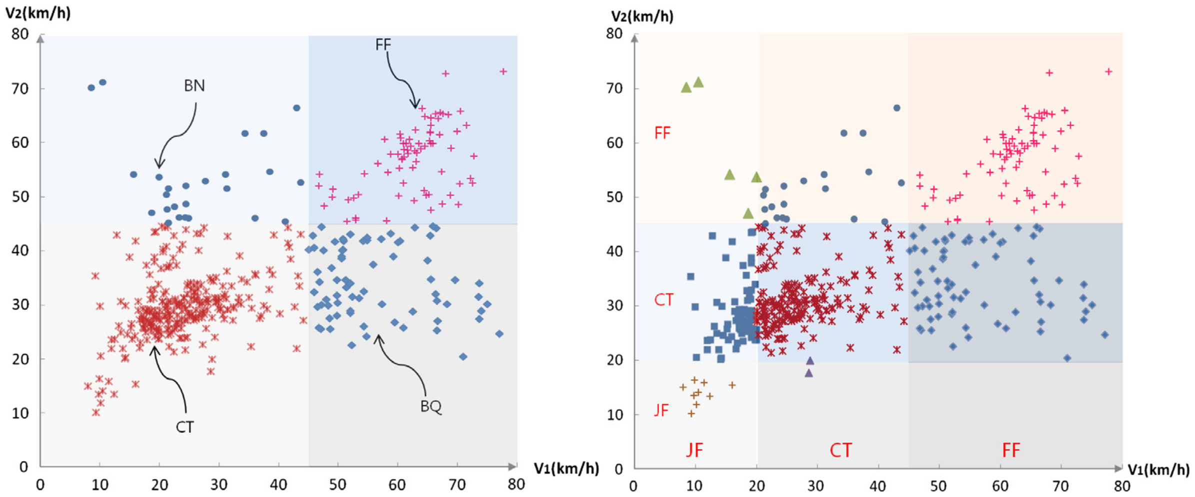
Using the alternative state identification method, the distributions of the crashes in nine states are summarized in Table 4. As indicated in Table 4, the major crash cases occurred under the CT–CT state, with few crashes in the JF–JF state. Meanwhile, the top two high risk states are subdivisions of the BN state and are followed by the CT–CT and JF–CT state. The difference observed in the results in Table 4 shows the necessity of the division of free flow, congested flow and jam flow.

Scatter plots of collected crash cases with average speed extracted from the upstream and downstream sections of crash locations before crash occurrence are shown in Fig. 3, illustrating the relationship of crashes and traffic states. Four states are shown in Fig. 3a and nine in Fig. 3b. In these figures, V_1 stands for upstream speed, shown on the horizontal axis, and V_2 is downstream speed on the vertical axis.

The above analysis was based on the speed condition data aggregated in the period of 5–10 min before the crash, while results are similar when the data within 10–15 min before the crash were analyzed. However, it was found that the preceding traffic states vary intensively during the two 5-min intervals. In the analysis using four states, there are 78 crash cases (19.0%) that experience state transition, whereas there are 115 crash cases (28.0%) in the analysis using nine states. Therefore, it is very important to capture the state transition process using several time intervals of traffic data when developing crash prediction models.

4. Methodology

The Bayesian Network (BN) model, an efficient statistical inference approach for uncertainty issues in artificial intelligence, has been used in accident severity analysis, incident detection and other traffic studies (De Oña et al., 2011; Zhang and Taylor, 2006). Regarding the real-time crash prediction issue, a static BN model has been used in studies by Hossain and Muromachi (2012, 2013b). Nevertheless, the multi-time-interval time-series traffic data adopted in this paper should be handled using dynamic models. Thus, the dynamic Bayesian network (DBN) model, which has been used previously for handling dynamic traffic issues (Hofleitner et al., 2012; Liang and Lee, 2014), was adopted for crash prediction.



(a) four states

(b) nine states

Fig. 3. Relationship of crashes vs. different traffic states.

4.1. Dynamic Bayesian network

A Bayesian network is a directed acyclic graph model annotated with probability that can express a joint probability distribution (physical or Bayesian) of a large set of variables. It is advantageous for data analysis when the graphical model is used in conjunction with statistical approaches. First, a Bayesian network can be used to learn causal relationships and hence can be used to predict the consequences of intervention. Second, because the model has both causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data (Heckerman, 1998). Dynamic Bayesian network is a type of BN that can model time-series data to capture the fact that time flows forward (Murphy, 2002).

With respect to BN, the key generalization is to conduct probabilistic inference for the hidden state in terms of a set of random variables (observations). Moreover, random variables in DBN that represent the hidden state are in time sequence. The observations can be represented in a factorized or distributed manner. Then, the graphical model can be used to represent conditional independence between these variables, both within and across positions in the sequence. In addition, because DBN is a type of directed graphical model, the conditional probability distribution (CPD) of each node in DBN can be estimated independently, making DBN easy to interpret and learn (Murphy, 2002).

In a DBN, the hidden state in time slice t is generally represented in terms of a set of N_h random variables, $H_t^{(i)}$, $i \in \{1, \ldots, N_h\}$, each of which can be discrete or continuous. Similarly, the observation can be represented in terms of N_o random variables, $E_t^{(j)}$, $j \in \{1, \ldots, N_o\}$, each of which can be discrete or continuous. In a hidden Markov models or state-space models, there is a transition model, $P(H_t|H_{t-1})$, an observation model, $P(E_t|H_t)$, and an initial state distribution, $P(H_1)$. However, as a more general model, a DBN is defined as a pair $(B_0, B_{\rightarrow})$ where B_0 defines the prior $P(Z_1)$, and $B_{\rightarrow}$ is a two-slice temporal Bayes net (2TBN) that defines the transition and observation models as a product of the CPDs in the 2TBN:

$$P(Z_t|Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^{(i)}|\pi(Z_t^{(i)})) \tag{1}$$

where $Z_t^{(i)}$ is the i'th node in time slice t (which can be a hidden or observation node, $N = N_h + N_o$), and $\pi(Z_t^{(i)})$ are the parents of $Z_t^{(i)}$, which may be at either time slice t or $t-1$. The nodes in the first slice of a 2TBN do not have parameters associated with them, while each node in the second slice has an associated CPD. Then, for a DBN with T slices, the joint distribution can be obtained by unrolling the 2TBN until the network has T slices and multiplying together all of the CPDs:

$$P(Z_{1:T}^{(1:N)}) = \prod_{i=1}^{N} P_{B_0}(Z_1^{(i)}|\pi(Z_1^{(i)})) \times \prod_{t=2}^{T}\prod_{i=1}^{N} P_{B_{\rightarrow}}(Z_t^{(i)}|\pi(Z_t^{(i)})) \tag{2}$$

As example structures of BN models shown in Fig. 4, the nodes in the graph model can present the hidden state or observed evidence. The hidden state is represented in terms of a single discrete random variable, expressed as the probability of each possible value. The observed evidence nodes represent random variables used for inference of the dependent variable. For the models used in this study, the hidden state is a binary variable H_t, which means “crash-prone” or “not crash-prone.” The traffic data related to crashes were recognized as the N_o observed evidence nodes $E_t^{(j)}, j \in \{1, \ldots, N_o\}$. Regarding

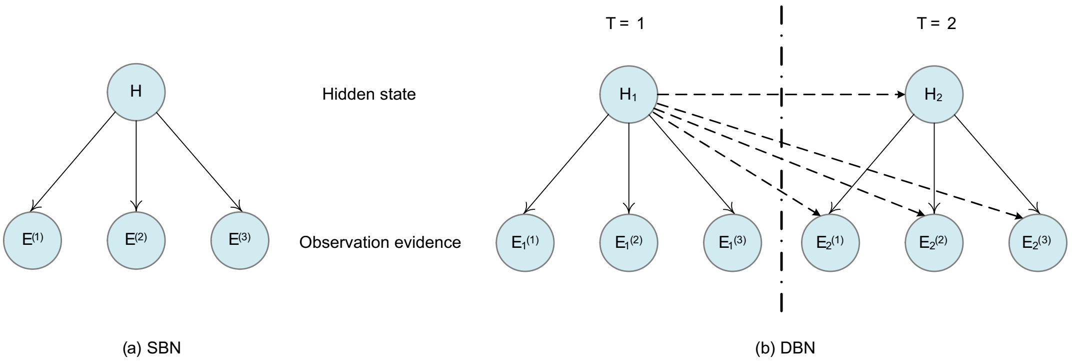


Fig. 4. Example structures of BNs; solid arcs represent intra-links, and dotted arcs represent inter-links.

the different input variables in this study, the model built with only speed condition data has three observation evidence nodes, while the model built with traffic flow detector data has six. Thus, it can be inferred whether a crash occurred or not by the observation of traffic data. Moreover, other than the static BN (SBN) in Fig. 4a, in the graph structure of a DBN in Fig. 4b, several consistent SBNs are connected in each time slice. The arcs in the graph represent conditional dependencies between variables, while the arcs within a time slice model instantaneous correlation and arcs across time slices depict transitions between the crash-prone states at two consecutive time slices.

When building DBN models, the time slices were another relatively important influencing factor. Because of the effectiveness of 5 min of aggregation of the data for predicting real-time crashes, the size of the time slice was fixed to 5 min. Additionally, the number of time slices may affect the model performance. In this paper, two two-time-slice models and one three-time-slice model were built separately and compared because at most three 5-min intervals of traffic data before the crash were collected.

There are usually three aspects of building a Bayesian network in a designated domain. (1) Variables and their ranges to describe the issue in the domain. The variables in this study include the crash/non-crash state and the observed traffic data. (2) Structure learning, which graphically represents the dependencies between variables. For the DBN, the structure considers not only the dependencies between variables in one time slice but also the dependencies existing in several time slices. Thus, structure learning algorithms in SBNs, such as the hill climbing algorithm, simulated annealing algorithm and genetic algorithm, cannot be directly used for DBN (Murphy, 2001). However, due to the few variables used in this study, the BN structures were pre-defined as shown in Fig. 4. (3) Parameter estimation to learn the conditional probability distribution of the observation variables. Parameter learning in DBN is very similar to learning in static networks, except that parameters must be tied across time-slices. To address situations in which some values of variables are missing that may appear in the traffic data collection, the gradient descent algorithm or expectation maximization (EM) algorithm are efficient approaches to compute the maximum likelihood estimates (Murphy, 2001). Thus, the EM algorithm was adopted in this study for parameter estimation in both DBN and SBN.

The EM algorithm learns the dependence among the observation evidences as an iterative process of parameters estimation (Dempster et al., 1977). Let θ be the set of unknown parameters of the model. We need to find the optimization parameters to maximize the log likelihood of the data, which is the log of the marginal probability of the observations given the following parameters:

$$\ell(\theta; E) = \log p(E|\theta) = \log \sum_H p(H, E|\theta) \tag{3}$$

Introducing a distribution $Q(H)$, we define an initialization distribution of θ. Using Jensen's inequality,

$$\ell(\theta; E) = \log \sum_H Q(H) \frac{p(H, E|\theta)}{Q(H)} \geqslant \sum_H Q(H) \log \frac{p(H, E|\theta)}{Q(H)} \tag{4}$$

- Then, the *E step* computes the distribution $Q(H) = P(H|E; \theta)$, which is in fact the joint probability distribution of the hidden state variables given the observed variables and the current values of the parameters. In a DBN, at each time slice t, we estimate the joint probability distribution of the hidden state variables.
- The *M step* optimizes the parameters based on the estimation of the joint probability distribution of the state variables:

$$\theta^{new} = \arg\max_\theta \sum_H Q(H) \log \frac{p(H, E|\theta)}{Q(H)} \tag{5}$$

Then, the original parameters are replaced by the new optimization parameters.

The *EM* algorithm repeats the *E* step and *M* step until reaching the specified number of iterations or achieving a local optimum of the estimated parameters.

After the DBN parameters are estimated, a DBN model can now infer the hidden states $H_{1:T}$ given the observations evidences $E_{1:T}$ by computing $P(H_{1:T}|E_{1:T};\theta)$. Consequently, the crash-prone state can be identified from normal conditions.

4.2. The model training and evaluation

Two types of BN models were built for crash prediction: DBN models were the main focus, and SBN models were developed for comparison. The time dependency of traffic states and crash occurrence can be evaluated by the comparison of these two types of models.

Similar to other classification models, a training dataset and testing dataset were assigned randomly from a dataset for the training and evaluation of BN models. Both training and testing datasets contained roughly similar crashes and non-crashes with their corresponding traffic condition data. To reduce the error caused by the random division of the dataset, ten experiments with different training and testing datasets randomly generated from one set of traffic data were conducted. Then, the mean value of the model evaluation metrics was used to represent the model performance.

Because of the small ratio of crash data (the ratio of crash cases and non-crash cases is approximately 1:5) in the dataset, the overall accuracy metric in Eq. (6) is no longer sufficient to evaluate the performance of a classification model for handling imbalanced data. Hence, several metrics based on the confusion matrix were used to evaluate the effectiveness of the imbalanced classification, as shown in Table 5.

The sensitivity in Eq. (7) and specificity in Eq. (8) are usually used to separately investigate the classification performance on two classes. The sensitivity is the crash classification accuracy, and the specificity is the non-crash classification accuracy, which is a complement of the false alarm rate. Additionally, the precision in Eq. (9) and the recall in Eq. (10) are often adopted to evaluate the effective detection ability for only one class. The optimally balanced classification ability can be observed by *G*-means in Eq. (11), which is the geometric mean of the sensitivity and specificity, while the *F*-measure in Eq. (12) represents the ability of the model to detect crashes (Tang et al., 2009).

$$\text{Overall accuracy} = \frac{T_{crash} + T_{non\_crash}}{T_{crash} + F_{crash} + F_{non\_crash} + T_{non\_crash}} \tag{6}$$

$$\text{Sensitivity} = T_{crash}/(T_{crash} + F_{non\_crash}) \tag{7}$$

$$\text{Specificity} = T_{non\_crash}/(T_{non\_crash} + F_{crash}) \tag{8}$$

$$\text{Precision} = T_{crash}/(T_{crash} + F_{crash}) \tag{9}$$

$$\text{Recall} = T_{crash}/(T_{crash} + F_{non\_crash}) \tag{10}$$

$$G\text{-means} = \sqrt{\text{sensitivity} * \text{specificity}} \tag{11}$$

$$F\text{-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{12}$$

Therefore, we evaluated the algorithms in terms of the prediction effectiveness using the crash classification accuracy, false alarm rate, overall classification accuracy, *G*-means and *F*-measure. Moreover, the *F*-measure of the crash prediction model was calculated here because finding crash-prone conditions is what we are particularly interested in.

5. Results

5.1. Results of DBN models

Based on the 411 crashes and corresponding 2055 non-crashes in the dataset collected on the Yan-an expressway, various DBN models were developed in this study in terms of different input variables. The model built using traffic flow detector data and the models built using only speed condition data were compared. Due to the two types of state classification methods, there are two speed condition data models: the model with four state combinations and the model with nine state combinations. In addition, with regard to the separated training dataset and testing dataset, independent evaluation of models was conducted. The results of the testing dataset are the focus of analysis. The analysis result is shown in Table 6, below.

As we can see from Table 6, all the DBN models obtained desirable classification accuracy. The majority of DBN models in Table 6 can predict approximately 75% of crashes and non-crashes. With different input variables, DBN models show different results. For the model with the training and testing dataset of traffic flow detector data, the crash prediction accuracy increases slightly, while the false alarm rate increases more. The overall result shows a degree of overfitting of the model. However, with regard to the speed condition data models, the results of the testing dataset appear steady or even better than

Table 5
Confusion matrix.

	Predicted crashes	Predicted non-crashes
Real crashes	T_{crash}	F_{non_crash}
Real non-crashes	F_{crash}	T_{non_crash}

the results of the training dataset. Therefore, these results of models with fewer variables can perform better with regard to overfitting. Furthermore, for all the results with the testing dataset, the models with speed condition data show better crash prediction ability than the model with detector data, as represented by the *F*-measure. It can thus be seen that with only speed data collected, the model is still reliable for predicting real-time crash occurrence.

With respect to traffic state identification, the model with nine state combinations shows better accuracy than the model with four state combinations for both the crash prediction and false alarm rates. It is also demonstrated that the traffic condition can be better interpreted by three levels of congestion. This approach is also more convenient for crash prediction because the state classification method is compatible with the current available method of determining congestion level on expressways in Shanghai.

Regarding different time slices, the prediction accuracy of DBN models with different input variables presents similar characteristics as expected. It is shown that the two-time-slice model with data collected in time intervals 2 and 3 always performed better than the model using collected in time intervals 3 and 4, while the three-time-slice model shows similar accuracy to the two-time-slice model with data collected in time intervals 2 and 3. Thus, it is suggested that the two-slice model with data collected in time intervals 2 and 3 is reasonable for crash prediction. For the purpose of clarity and concise, the complete results are not included in this paper.

5.2. Comparison between DBN and SBN model

With the best-performed data in DBN models, SBN model was also developed for comparison. As the SBN model can only address one-time-slice data, traffic data collected in time interval 2 (5–10 min before crash) were used in the establishment of model. The results of the comparison are presented in Fig. 5.

It can be indicated from Fig. 5 that the DBN model performs better than the SBN model for both crash prediction accuracy and false alarm rate. The DBN model can predict 3% more crashes than the SBN model with a 0.6% reduction in false alarm rate. It was also proved that there is a relationship between consecutive time-slice traffic data before crashes that can be captured by the DBN model. The identification of traffic state transition is considered to be more closely related to the crash occurrences.

5.3. Transferability of DBN models

The spatial transferability of DBN models was also evaluated in this study. The transferability is the ability of the crash prediction model developed on one expressway as assessed when applied directly to other similar expressways. The data

Table 6
Results of various DBN models.

Model		Crash accuracy	False alarm rate	Overall accuracy	*G*-means	*F*-measure
Model with traffic flow detector data	Training dataset	0.751	0.232	0.766	0.760	0.511
	Testing dataset	0.757	0.26	0.743	0.749	0.490
Model with speed condition data (four traffic state combinations)	Training dataset	0.769	0.263	0.742	0.752	0.492
	Testing dataset	0.744	0.241	0.757	0.751	0.499
Model with speed condition data (nine traffic state combinations)	Training dataset	0.761	0.237	0.762	0.762	0.511
	Testing dataset	**0.764**	**0.237**	**0.763**	**0.763**	**0.512**

The bold values denote the results of the best fitted model.

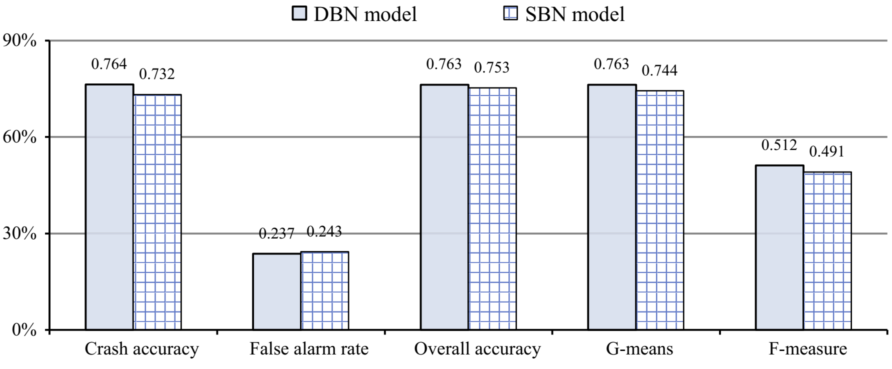


Fig. 5. Comparison of DBN model and SBN model.

Table 7
Transferability comparison of DBN models.

Model	Crash accuracy	False alarm rate	Overall accuracy	*G*-means	*F*-measure
Model with traffic flow detector data	0.700	0.271	0.724	0.714	0.458
Model with speed condition data (four traffic state combinations)	0.674	0.228	0.756	0.721	0.482
Model with speed condition data (nine traffic state combinations)	0.670	0.208	0.772	0.728	0.495

collected on the Shanghai North–South expressway were used to evaluate the transferability of DBN models built with the data on the Yan-an expressway. Here, we treated the combined 140 crash and 700 non-crash data as a testing dataset to predict the category of each case. Models with different input variables were also compared with regard to transferability. Only data collected in time intervals 2 and 3 were used to evaluate the transferability of DBN models.

As shown in Table 7, the DBN models can predict many crashes with a relatively low false alarm rate. When compared with the original model, the crash prediction accuracy decreases to a reasonable extent when the DBN model is applied on another expressway. Fortunately, the false alarm rate also decreases slightly. With the traffic state variables and speed condition data used, the transferability of DBN models is promoted, as the best-performing model shown by the highest *F*-measure value can predict 67% of crashes with a 20.8% false alarm rate. With regard to the traffic states, the transferability of the DBN model using four state combinations was worse than the nine state combination model, as shown by the *F*-measure. However, the higher crash prediction accuracy of the detector data model indicates that we should not be over-optimistic about the transferability of the speed condition data model.

For the spatial transferability test results in previous studies, 55.84% crash prediction was obtained using the logistic regression model in the study by Pande et al. (2012), and 59.3% of crashes can be identified with a 20% false alarm rate using the Bayesian updating approach by Xu et al. (2014a). Thus, it can be seen that the DBN models have good transferability for application to other similar expressways.

6. Conclusions

This study aimed to utilize traffic speed condition data for real-time crash prediction on expressways. Two datasets totaling 551 crashes and 2755 non-crashes coupled with the corresponding traffic data before the crashes (and non-crashes) were collected on two expressways in Shanghai. A real-time crash prediction model using the DBN model to capture the relationship of dynamic traffic state variation and crashes was proposed. The DBN models were developed and evaluated using data collected on the Yan-an expressway, and the transferability of the model was examined using data collected on the North–South expressway. The main conclusions are as follows:

(1) The traffic state near the crash site was abstracted into several state combinations with average traffic speed data collected in the upstream and downstream sections according to the congestion levels. The results indicated different characteristics associated with crash occurrences and different state combinations. By comparison within different DBN models, the results indicate that with nine state combinations included and speed condition data adopted, the best fitted DBN model achieved a crash prediction accuracy of 76.4% with a false alarm rate of 23.7%.
(2) Compared with the SBN model, the DBN model used in this study can handle time series traffic data before crash occurrences and identify the state transition. The results compared with SBN suggested that the DBN model is more suitable to the prediction of real-time crashes, considering the time dependency between different time slice data.
(3) From the results regarding the transferability of the three different models, it was found that traffic state variables and speed condition data are both advantageous for crash prediction, especially for reducing the false alarm rate. The best performing DBN model on the Yan-an expressway can predict 67% of crashes on the North–South expressway with a 20.8% false alarm rate.

Acknowledgements

The authors would like to thank the Natural Science Foundation of China (51278362, 51422812), the New Century Excellent Talents in University (NCET13-0425) and the Fundamental Research Funds for the Central Universities for supporting this research.

References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Safety Res. 36 (1), 97–108.
Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., Hsia, L., 2004. Predicting freeway crashes from loop detector data using matched case-control logistic regression. Transp. Res. Rec. 1897, 88–95.
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., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. Accid. Anal. Prev. 38 (2), 335–345.

Abdel-Aty, M., Hassan, H.A., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transport. Res. Part C: Emerg. Technol. 24, 288–298.

Ahmed, M., Abdel-Aty, M., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transp. Syst. 13 (2), 459–468.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012. A Bayesian updating approach for real-time safety evaluation using AVI data. Transp. Res. Rec. 2280, 60–67.

Allaby, P., Hellinga, B., Bullock, M., 2007. Variable speed limits: safety and operational impacts of a candidate control strategy for freeway applications. IEEE Trans. Intell. Transp. Syst. 8 (4), 671–680.

De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accid. Anal. Prev. 43 (1), 402–411.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.), 1–38.

Golob, T., Recker, W., Pavlis, Y., 2008. Probabilistic models of freeway safety performance using traffic flow data as predictors. Saf. Sci. 46 (9), 1306–1333.

Heckerman, D., 1998. A tutorial on Learning with Bayesian Networks. Springer, Netherlands, pp. 301–354.

Hofleitner, A., Herring, R., Abbeel, P., Bayen, A., 2012. Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network. IEEE Trans. Intell. Transp. Syst. 13 (4), 1679–1693.

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. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accid. Anal. Prev. 57, 17–29.

Hossain, M., Muromachi, Y., 2013b. A real-time crash prediction model for the ramp vicinities of urban expressways. IATSS Res. 37 (1), 68–79.

Hu, J., Sun, J., Zhao, L., 2014. Some Flow Features at Urban Expressway On-ramp Bottlenecks in Shanghai. In: Presented at the 93rd Annual Meeting of the Transportation Research Board, Washington, D.C.

Lee, C., Saccomanno, F., Hellinga, B., 2003. Real-time crash prediction model for the application to crash prevention in freeway traffic. Transp. Res. Rec. 1840, 67–77.

Lee, C., Hellinga, B., Saccomanno, F., 2004. Assessing safety benefits of variable speed limits. Transp. Res. Rec. 1897, 183–190.

Li, Z., Wang, W., Chen, R., Liu, P., Xu, C., 2013. Evaluation of the impacts of speed variation on freeway traffic collisions in various traffic states. Traffic Injury Prevent. 14 (8), 861–866.

Liang, Y., Lee, J.D., 2014. A hybrid Bayesian Network approach to detect driver cognitive distraction. Transport. Res. Part C: Emerg. Technol. 38, 146–155.

Murphy, K.P., 2001. The Bayes net toolbox for matlab. Comput. Sci. Stat. 33 (2), 1024–1034.

Murphy, K.P., 2002. Dynamic Bayesian networks: representation, inference and learning (Doctoral dissertation, University of California, Berkeley).

Oh, C., Oh, J., Ritchie, S., Chang, M., 2001. Real-time estimation of freeway accident likelihood. In: Presented at the 80th Annual Meeting of the Transportation Research Board, Washington, D.C.

Pande, A., Abdel-Aty, M., 2006. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transport. Res. Rec.: J. Transport. Res. Board 1953, 31–40.

Pande, A., Abdel-Aty, M., Hsia, L., 2005. Spatiotemporal variation of risk preceding crashes on freeways. Transp. Res. Rec. 1908, 26–36.

Pande, A., Das, A., Abdel-Aty, M., Hassan, H., 2011. Real-time crash risk estimation are all freeways created equal? Transp. Res. Rec. 2237, 60–66.

Pande, A., Nuworsoo, C., Shew, C., 2012. Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways (No. MTI Report 11–15).

Sun, J., Sun, J., Chen, P., 2014a. Use of support vector machine models for real-time crash risk prediction on urban expressways. Transp. Res. Rec. 2432, 91–98.

Sun, J., Zhao, L., Zhang, H.M., 2014b. The mechanism of early-onset breakdown at Shanghai's expressway on-ramp bottlenecks. Transp. Res. Rec. 2421, 64–73.

Tang, Y., Zhang, Y.Q., Chawla, N.V., Krasser, S., 2009. SVMs modeling for highly imbalanced classification. IEEE Trans. Syst. Man Cybern. B Cybern. 39 (1), 281–288.

Xu, C., Wang, W., Liu, P., 2013. A genetic programming model for real-time crash prediction on freeways. IEEE Trans. Intell. Transp. Syst. 14 (2), 574–586.

Xu, C., Wang, W., Liu, P., Guo, R., Li, Z., 2014a. Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models. Transport. Res. Part C: Emerg. Technol. 38, 167–176.

Xu, C., Wang, W., Liu, P., Zhang, F., 2014b. Development of a real-time crash risk prediction model incorporating the various crash mechanisms across different traffic states. Traffic Injury Prevent. 16 (1), 28–35.

Yang, H., Bartin, B., Ozbay, K., 2014. Mining the characteristics of secondary crashes on highways. J. Transport. Eng. 140 (4), 04013024.

Yeo, H., Jang, K., Skabardonis, A., Kang, S., 2013. Impact of traffic states on freeway crash involvement rates. Accid. Anal. Prev. 50, 713–723.

Zhang, K., Taylor, M.A.P., 2006. Effective arterial road incident detection: a Bayesian network based algorithm. Transport. Res. Part C: Emerg. Technol. 14 (6), 403–417.

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

Zheng, J., Sun, J., Yang, J. 2015. Relationship of lane width to capacity for urban expressways. Transp. Res. Rec., in press.