


Predicting Imminent Crash Risk with Simulated Traffic from Distant Sensors

Transportation Research Record
2018, Vol. 2672(38) 12–21
© National Academy of Sciences:
Transportation Research Board 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0361198118791379
journals.sagepub.com/home/trr


Zhi Chen¹, Xiao Qin¹, Renxin Zhong², Pan Liu³, and Yang Cheng⁴

Abstract

The aim of this research was to investigate the performance of simulated traffic data for real-time crash prediction when loop detector stations are distant from the actual crash location. Nearly all contemporary real-time crash prediction models use traffic data from physical detector stations; however, the distance between a crash location and its nearest detector station can vary considerably from site to site, creating inconsistency in detector data retrieval and subsequent crash prediction. Moreover, large distances between crash locations and detector stations imply that traffic data from these stations may not truly reflect crash-prone conditions. Crash and noncrash events were identified for a freeway section on I-94 EB in Wisconsin. The cell transmission model (CTM), a macroscopic simulation model, was applied in this study to instrument segments with virtual detector stations when physical stations were not available near the crash location. Traffic data produced from the virtual stations were used to develop crash prediction models. A comparison revealed that the predictive accuracy of models developed with virtual station data was comparable to those developed with physical station data. The finding demonstrates that simulated traffic data are a viable option for real-time crash prediction given distant detector stations. The proposed approach can be used in the real-time crash detection system or in a connected vehicle environment with different settings.

A driver must constantly respond to changes in traffic and other roadway conditions by changing speed, switching lanes, or even changing directions. The inevitable and frequent changes in driving conditions can result in driver errors. The early detection of crash-prone traffic conditions can alert the driver to make necessary evasive maneuvers, and it can also lead to appropriate traffic control strategies to mitigate imminent crash risk. The factors contributing to a crash are directly related to or result from the prevailing traffic conditions before the event. The wide deployment of Advanced Traveler Information Systems (ATIS) has made the collection, storage, and processing of real-time traffic data readily available. Researchers can now gather real-time information related to crash occurrence. Among different types of traffic sensors, inductive loop detectors have been a popular data source for real-time crash prediction. Numerous studies have investigated the relationship between crash risk and prevailing traffic conditions using real-time traffic information collected from loop detectors; therefore, it would be beneficial to develop a real-time crash prediction model that can detect crash-prone traffic patterns.

The lead time before a crash has been thoroughly studied in cases in which freeway crashes were predicted

using loop detector data at different time slices (*I*). However, studies regarding how the space between detector stations and crash locations affects crash prediction accuracy are rare. It is expected that the traffic conditions near a crash location would better reflect the hazardous conditions leading to the crash. In nearly all previous studies, crash prediction models were developed using the traffic data from immediate upstream and downstream loop detector stations; however, crashes may take place anywhere between two detector stations, and the distance between a crash location and the nearest detector station can vary considerably from site to site. The variation creates inconsistency in detector data retrieval and subsequent crash prediction, and also casts

¹Department of Civil and Environmental Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI

²School of Engineering, Sun Yat-Sen University, Guangzhou, China

³Jiangsu Key Laboratory of Urban ITS, Southeast University, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Nanjing, China

⁴Department of Civil & Environmental Engineering, University of Wisconsin-Madison, Madison, WI

Corresponding Author:

Address correspondence to Xiao Qin: qinx@uwm.edu

doubt on how well the traffic data from these stations reflects crash-prone conditions.

The goal of this study was to examine the performance of simulated traffic data in detecting crash-prone traffic patterns given large distances between loop detector stations. The cell transmission model (CTM) was applied to simulate traffic conditions close to a crash location to accommodate the challenges of spatial disparities in loop detectors and discrepancies in crash prediction.

Literature Review

Ideally, traffic conditions at the location and time of a crash should be used in capturing and recording what exactly happened before the crash. The reality, however, is that traffic data often are observed from physical detectors which are not likely to be installed anywhere near the crash. Alternatively, virtual detectors may be exploited at any location through the development of traffic simulation models. CTM is a representative macroscopic traffic flow simulation model proposed by Daganzo (2). In a CTM, a highway is partitioned into a series of cells, and the cell length can be user-defined. The traffic density and other characteristics in each cell can be calculated following the traffic flow–density relationship defined by the fundamental diagram (FD). The methodology is detailed in the next section. CTM is very compatible with traffic flow data collected from loop detectors, which is a clear advantage for real-time crash prediction. CTM and its extensions, including switching-mode model (SMM) and stochastic CTM, have shown promising results in accurately predicting traffic flow characteristics using loop detector data as inputs (3–5). Studies show that simulated traffic densities and flow rates may achieve an error rate as low as 7.9% (5) and 4% (4).

Although CTM is a powerful model for simulating traffic flow characteristics, the existence of an underlying relationship between traffic flow and crash risk warrants the validity of a real-time crash prediction model. Many efforts have been made to define and quantify such a relationship. Some studies used data mining techniques including the Kohonen clustering algorithm, neural networks, and the Bayesian network (6–8). Although data mining methods are capable of accounting for correlation within speed, flow, and occupancy (6), they cannot identify explicit relationships between crash risk and traffic flow variables. Other studies have used statistical regression models because they provide a clear connection between crash probability and traffic flow variables. Among all methods, the case-control design has been the most popular in real-time crash prediction studies (1, 6–13) because the design controls for exogenous factors such as locations and roadway geometries, and provides

more accurate estimates by using both crash and non-crash traffic information (1). However, the case-control method assumes that each stratum of a crash event and its matched noncrash events have different constant terms, meaning that only the odds ratio of crash probability is predicted. In other words, the crash probability cannot be directly predicted given the explanatory variables. An alternative method is the binary logistic regression. The binary logistic model is used to estimate the probability of a binary response, such as a crash event based on one or more predictors, meaning the crash probability can be directly predicted given the explanatory variables (14–16).

The relationship between crash probability and traffic flow variables (e.g., mean, standard deviation, and coefficient of variation of traffic flow, speed, and occupancy) is under intense scrutiny (1, 8, 9, 17–19). Roshandel et al. conducted a comprehensive review of the relationship between real-time traffic conditions and freeway crashes using the meta-analysis of past literature (20). The authors identified some statistically significant crash-contributing factors such as average speed and speed variation which have been consistently reported in past studies. The authors also pointed out a limitation of using loop detector data, which is that researchers must use data from loop detectors that could be far from crash locations (20). The spacings between loop detector stations vary significantly within and across studies, rendering findings that are unreliable within the same study and inconsistent between studies.

Methodology

This section covers the methodology for simulating spatial and temporal traffic during the period before a crash, and also details the regression method for predicting crash occurrence. The traffic variables are simulated with CTM, and the binary logistic regression model is used to estimate the probability of a crash based on simulated traffic conditions.

Cell Transmission Model (CTM)

CTM is a macroscopic traffic simulation model that predicts macroscopic traffic behavior such as flow and density at a finite number of cells at different time steps on a given highway corridor (2). It is a powerful simulation tool that can capture many important traffic phenomena such as queue formation and dissipation as well as shock-wave propagation (2). Compared with microscopic simulation models, CTM is computationally efficient and easier to configure and calibrate. CTM also operates sufficiently with aggregated traffic flow data from detector stations. The core component of CTM is the FD that

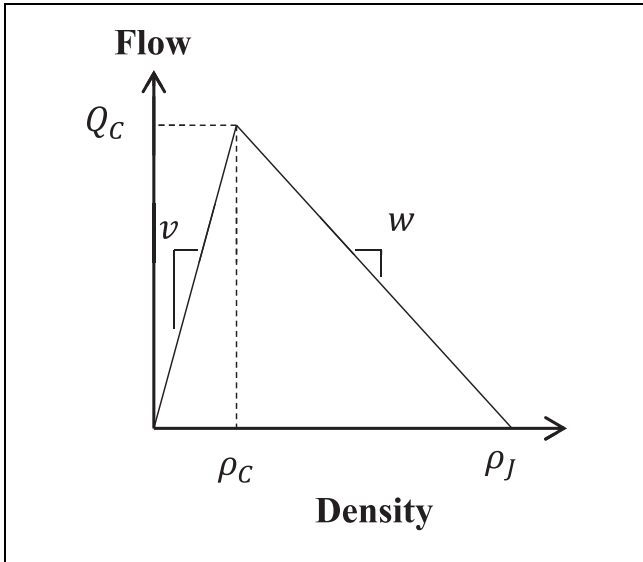


Figure 1. Triangular fundamental diagram.

defines the flow–density relationship and governs how the traffic state in each cell evolves over time. Figure 1 shows a triangular FD for developing CTM.

In CTM, a highway segment is divided into a series of cells. The density of each cell evolves following the conservation law of vehicles. Assume that Cell i is characterized by the triangular FD in Figure 1, where Q_C is the capacity flow, ρ_C is the critical density, ρ_J is the jam density, v is the free-flow speed, and w is the shockwave speed. The density for Cell i without on- or off-ramps is determined by

$$\rho_i(k+1) = \rho_i(k) + \frac{T}{l_i}(q_i(k+1) - q_i(k)) \quad (1)$$

where

k is the time step index,

$\rho_i(k)$ is the density of Cell i during the k th time step,

T is the length of the time step,

l_i is the length of Cell i , and

$q_i(k)$ is the flow rate into Cell i during the k th time step. The flow rate is determined by the sending and receiving functions. For Cell i , the sending function $S_i(k)$ represents the maximum flow that may be supplied during the k th time step, and the receiving function $R_i(k)$ represents the maximum flow that may be received. The two functions are determined in Equations 2 and 3, respectively, as

$$S_i(k) = \min(v_i \rho_i(k), Q_{C,i}) \quad (2)$$

$$R_i(k) = \min(Q_{C,i}, w_i(\rho_{J,i} - \rho_i(k))) \quad (3)$$

The flow rate, $q_i(k)$, is determined by

$$q_i(k) = \min(S_{i-1}(k), R_i(k)) \quad (4)$$

Binary Logistic Regression Model

In a binary logistic regression model, the probability of a crash event can be formulated as

$$p(\mathbf{X}_i) = \frac{1}{1 + e^{-g(\mathbf{X}_i)}} \quad (5)$$

where $p(\mathbf{X}_i)$ represents the crash probability given $\mathbf{X}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,k})$, a set of k explanatory variables for sample i , and $g(\mathbf{X}_i)$ is a linear combination of the following variable set

$$g(\mathbf{X}_i) = \beta_0 + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \dots + \beta_k * x_{i,k} \quad (6)$$

where $(\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ are the corresponding coefficients for $(x_{i,1}, x_{i,2}, \dots, x_{i,k})$.

The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ can be estimated by maximizing the following log-likelihood function:

$$\ln L(\boldsymbol{\beta}, \mathbf{X}_i) = \sum_{i=1}^n \left[(\beta_0 + \beta_1 * x_{i,1} + \dots + \beta_k * x_{i,k}) - \ln \left(1 + e^{-(\beta_0 + \beta_1 * x_{i,1} + \dots + \beta_k * x_{i,k})} \right) \right] \quad (7)$$

Data Description

The dataset used in this study consists of 3 years (2012–2014) of crash data, roadway characteristics, and traffic data for a 4.15-mi I-94 EB corridor in Wisconsin. The corridor has three lanes with one on-ramp and one off-ramp. There are seven mainline loop detector stations and one off-ramp loop detector station on the EB and WB corridor, and there is no detector station on the on-ramp. The layout of the corridor and detector stations is shown in Figure 2.

The 3rd and 6th detector stations were assumed to be absent and were not used in this study in order to test the concept of using simulated traffic data for crash prediction in the setting of large detector spacing. The remaining stations were named as S1, S2, S3, S4, and S5 as shown in Figure 2. The CTM model would provide inaccurate simulated traffic data given missing on-ramp flow, so one segment between S1 and S3 and one segment between S4 and S5 were included in place of the missing on-ramp detector station between stations S3 and S4.

Spacings between S1 and S2, S2 and S3, and S4 and S5 are 1.00 mi, 1.06 mi, and 1.38 mi, respectively. Segments of interest include one 2.06 mi-long segment between S1 and S3, and a 1.38 mi-long segment between S4 and S5. The two segments are divided into 13 cells for CTM simulation. Cells have a uniform length of 0.25 mi with a few exceptions that range from 0.26 mi to 0.29 mi.

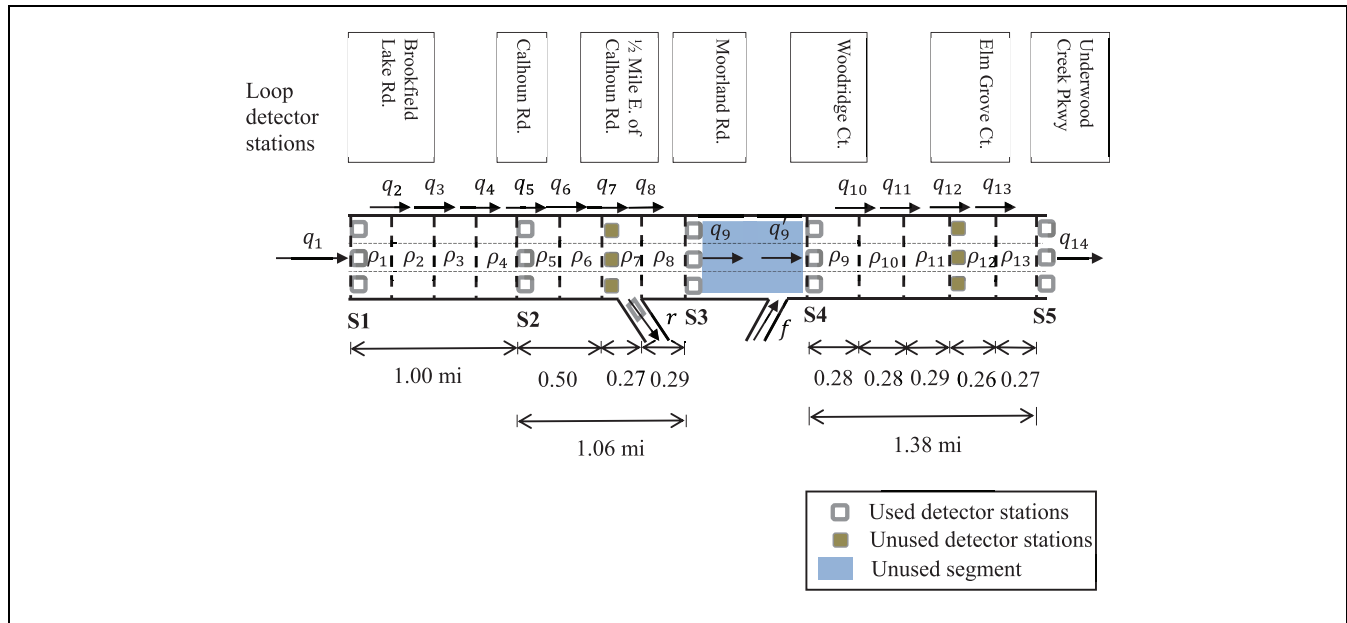


Figure 2. Layout of physical loop detector stations and CTM cells.

In Figure 2, the default cell length is 0.25 mi if no other length is specified. The average cell length is 0.264 mi with a standard deviation of 0.016 mi.

Crash data were retrieved from the web-based query and retrieval facility for Wisconsin Department of Transportation (DOT) crash data as well as from reports archived in the WisTransPortal data management system. Crashes occurring at the study site from 2012 to 2014 were included in the study. Any other crash that happened within 2 h after a crash occurrence and within 2 mi of the crash location was considered a secondary crash and was subsequently removed (21). Crash time is required to build up the stratum, so crashes with missing times were excluded. Weather information such as snow or rain was collected from the Wisconsin DOT's Road Weather Information System (RWIS), and traffic information in a 1-min time interval was extracted from the V-SPOC (Volume, Speed, and Occupancy) application suite (22). All inaccurate data were eliminated from the raw 1-min data based on the following criteria: 1) occupancy < 0 or > 100 ; 2) speed < 0 or > 100 ; 3) volume < 0 or > 50 in 1 min; 4) volume > 0 with speed = 0 or speed > 0 with volume = 0 (23).

For each crash in the dataset, 1-min traffic data from the 0–5 min or 5–10 min intervals before the crash occurred were collected from immediately upstream and downstream physical loop detector stations. For example, if a crash occurred at 10:00 a.m., the traffic data were extracted from 9:55 to 10:00 a.m. if the 0–5 min interval was of interest, or from 9:50 to 9:55 a.m. if the 5–10 min interval was of interest. Each crash has a geolocation, which determines the cell in which the crash occurred.

Traffic data that are not affected by or associated with crash occurrence, also called “noncrash events,” are also required to develop the crash prediction model. Ten noncrash events were selected for each crash by randomly selecting a time among 1,578,240 1-min intervals in 2012–2014 ($60 \text{ min} \times 24 \text{ h} \times 1096 \text{ days}$ in 2012–2014) and a cell among 13 cells. Noncrash events were selected in a way that ensured none of the times was within 2 h of any crash. The 5-min traffic data consisting of data from five 1-min intervals was retrieved from physical detector stations for noncrash events in the same way it was for crash events.

CTM Setup and Calibration

An FD is required to operate the CTM simulation. Differences in roadway characteristics such as distances to on-/off-ramps could lead to cells having different traffic patterns and therefore different FDs. The calibration algorithm proposed by Zhong et al. (24) was applied for calibrating FDs. The algorithm found the optimal FD parameters to minimize the discrepancy between CTM simulated traffic data and observed traffic data. That study can be referenced for more details.

Traffic data must be collected from detector stations in order to calibrate FDs. Traffic data from five detector stations were collected between 4:00 a.m. and 12:00 p.m. on May 6, 2013. Free flow, which is the onset and offset of congestion, was observed at all five stations during this time, which is ideal for FD calibration. It is not computer-efficient to solve for optimal FDs for each cell when the cell number is relatively large and the station

Table 1. FD Parameters

Cell*	v	ρ_c	ρ_j	Q_c	w
1, 2	58.74	108.04	398.65	6347	21.84
3, 4	58.22	112.96	545.34	6577	15.21
5, 6	66.38	100.16	409	6649	21.53
7	60.52	107.76	569.44	6522	14.13
8	56.5	99.26	544.03	5608	12.61
9	63.42	92.6	370.42	5873	21.14
10, 11, 12	73.32	79.77	644.5	5848	10.36
13	67.64	86.6	359.95	5858	21.43

Note: * denotes the cells sharing the same FD.

number is small. Therefore, it was determined that several adjacent cells might share one FD. Table 1 presents the FD parameters for different cells. The first column shows the cells with the same FD. For example, Cell 1 and 2 have the same FD. Note that Cell 7 has its own FD as it has an off-ramp and could have a very different FD.

The validity of proposed FD parameters was examined by comparing simulated traffic variables with observed traffic variables at the two physical stations that were not used for calibrating CTM. Traffic data from physical detector stations in Cells 7 and 12 were collected from 4:00 a.m. to 12:00 p.m. on May 6, 2013. During the same time period, traffic data were simulated from virtual stations by CTM in the same cells. The mean absolute percentage error (MAPE) is used to measure the difference:

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \frac{|\bar{y}(i) - y(i)|}{y(i)} \quad (8)$$

where N is the count of 1-min traffic records, $\bar{y}(i)$ is the i th simulated traffic volume or density, and $y(i)$ is the i th observed traffic flow rate or traffic density. In Cell 7, MAPEs are 5.76% and 7.03% for the flow rate and density at the first physical station. In Cell 12, MAPEs are 5.95% and 11.46% for the flow and density at the second physical station. The small MAPE values show that CTM with proposed FD parameters can provide reliable measurements of traffic flow behavior at these virtual stations.

CTM Simulation

In the simulation, the Courant–Friedrichs–Lewy (CFL) condition (25) needs to be fulfilled to guarantee a feasible CTM. The CFL condition occurs when a vehicle cannot travel across more than one cell during one simulation step, that is, $v_i * T \leq l_i$ where v_i is the free-flow speed, T is the simulation time step, and l_i is the cell length. The CFL constraint prohibits the use of a time step as large

as 1 min based on the constructed cell length. Therefore, 5 s was used as the time step ($T = 5$ s) to guarantee the feasibility of the CTM.

The cell and segment within which the crash occurred, as well as its upstream and downstream detector stations, were identified to simulate crash location traffic data. S1 and S3 are considered as the upstream and downstream stations for the first segment, and S4 and S5 are the two stations for the second segment. The flow data from both stations during the 0–5 min or 5–10 min period before a crash were used as the in-flow and out-flow of the segment consisting of the cells in between. A 0th-order interpolation was applied to generate the in-flow/out-flow data in 5 s. A CTM was then run to simulate how traffic data in those cells evolve at each time step within the 5-min time interval.

Virtual detector stations were set up at the beginning of each cell for the CTM simulation. The spacing between virtual stations was equal to the cell length, and was therefore consistent in distance. The virtual stations were expected to capture traffic conditions at locations closer to the crash site where traffic conditions should be more related to the crash than those collected from more distant physical detector stations. Similar to physical detector stations, virtual stations were set up to measure flow, speed, and density. Because the time step was set to be 5 s for the CTM simulation, virtual stations would detect 5-s traffic data. To be comparable to the 1-min traffic data from physical stations, 5-s simulated traffic data were aggregated to 1-min traffic data.

A virtual station k is located at the beginning of Cell k , and measures flow, q_k , density, ρ_k and speed, s_k . The model developed with traffic data from the two virtual upstream and two virtual downstream stations is called a two-station setting. The model developed with traffic data from one virtual upstream and one virtual downstream station is called a one-station setting.

Figure 3 shows that for crashes occurring in Cell k , the two-station setting would include virtual station $k-1$ and k as two virtual upstream stations, and virtual

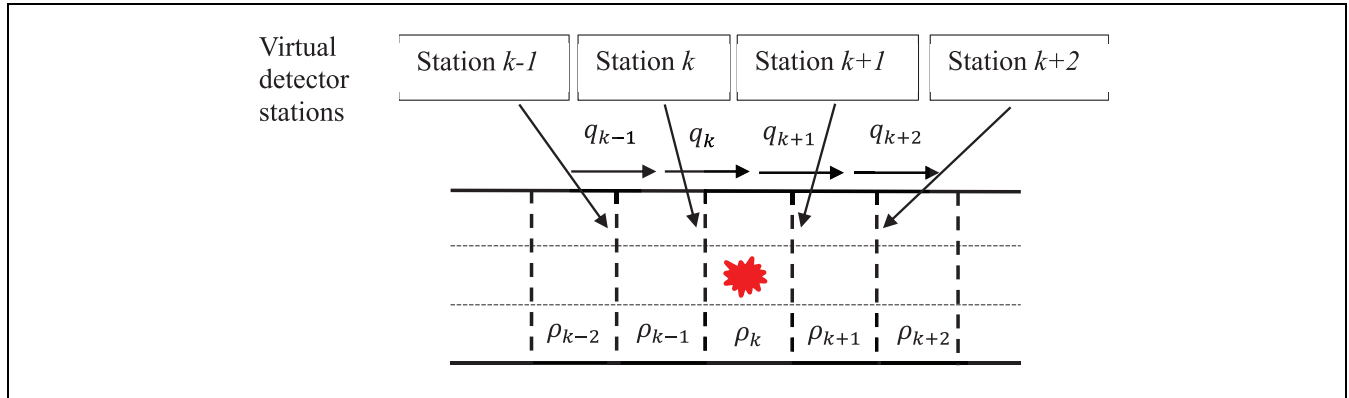


Figure 3. Layout of virtual loop detector stations.

station $k + 1$ and $k + 2$ as two virtual downstream stations; the one-station setting would only include virtual station k as one virtual upstream station, and virtual station $k + 1$ as one virtual downstream station. For each crash/noncrash that occurred in the remaining cells, the virtual upstream and downstream detector stations were identified and traffic data were collected from these virtual detector stations for developing crash prediction models.

Analysis and Discussion

The traffic variables such as mean, standard deviation, and coefficient of variation of flow, density, and speed were calculated for physical or virtual detector stations for each event. Traffic data from one immediate upstream and one immediate downstream station were used to generate variables for physical stations.

Several studies have considered the impact of various traffic states on safety. Abdel-Aty et al. assessed crash risk at two different traffic states categorized by speed (15). The authors discovered different crash-contributing traffic parameters in the high- and low-speed regimes. Li et al. (10) found that different sets of statistically significant traffic variables affect crash probability in distinct traffic states when expanding from two to four traffic states based on the speed from upstream and downstream detector stations (26). Traffic state was added to this study to improve model performance per the previous study's findings. The average density was used to measure the level of traffic congestion (26); traffic is congested only if the average density is greater than the critical density. Traffic state was determined based on the combination of the upstream and downstream traffic conditions:

- Free Flow (FF): when both upstream state and downstream state are free flow;

- Bottleneck front (BN): when upstream is congested and downstream is free flow;
- Back of queue (BQ): when upstream is free flow and downstream is congested; and
- Congested traffic (CT): when both upstream and downstream are congested.

Table 2 displays the candidate variables including traffic flow variables, weather condition and ramp presence. Three models were developed using different data sources to generate traffic variables: physical stations, virtual stations in a one-station setting, and virtual stations in a two-station setting. The three models are referred to as Model P, V1, and V2. Two time intervals were tested for all three models: 0–5 min before a crash and 5–10 min before a crash. A crash/noncrash dataset suitable for all models was identified so that all six models could be compared; Events with missing physical detector data were removed, and the final dataset consists of 531 events, including 66 crashes and 465 noncrashes.

Crash prediction models were developed using the binary logistic regression model to identify the relationship between crash risk and explanatory variables. The stepwise variable selection method was applied to identify the significant variables providing the best goodness-of-fit. Estimation results were obtained by fitting the model with those selected variables. The prediction accuracy of models was checked by conducting the leave-one-out cross-validation (LOOCV) with selected significant variables from each model. The LOOCV method uses one observation as the validation dataset and all the remaining observations as the training dataset. A model was fitted, given the training dataset, and was then used to predict the crash probability of that single observation in the validation dataset. This procedure was then repeated for all observations in the dataset. Based on the LOOCV results, ROC (receiver operating characteristic) curves for all six models are plotted in Figure 4 and the AUC (area under curve) values are presented in Table 3.

Table 2. Candidate Variables for Model Development

Variable	Description
UFA	Average flow at the upstream station(s) (vph)
UFS	SD of flow at the upstream station(s) (vph)
UFCVS	Coefficient of variation of flow at the upstream station(s) (vph)
UDA	Average density at the upstream station(s) (vpm)
UDS	SD of density at the upstream station(s) (vpm)
UDCVS	Coefficient of variation of density at the upstream station(s) (vpm)
USA	Average speed at the upstream station(s) (mph)
USS	SD of speed at the upstream station(s) (mph)
USCVS	Coefficient of variation of speed at the upstream station(s) (mph)
DFA	Average flow at the downstream station(s) (vph)
DFS	SD of flow at the downstream station(s) (vph)
DFCVS	Coefficient of variation of flow at the downstream station(s) (vph)
DDA	Average density at the downstream station(s) (vpm)
DDS	SD of density at the downstream station(s) (vpm)
DDCVS	Coefficient of variation of density at the downstream station(s) (vpm)
DSA	Average speed at the downstream station(s) (mph)
DSS	SD of speed at the downstream station(s) (mph)
DSCVS	Coefficient of variation of speed at the downstream station(s) (mph)
Diff_FA	Average absolute difference in flow between upstream and downstream stations
Diff_DA	Average absolute difference in density between upstream and downstream stations
Diff_SA	Average absolute difference in speed between upstream and downstream stations
Diff_FS	SD of absolute difference in flow between upstream and downstream stations
Diff_DS	SD of absolute difference in density between upstream and downstream stations
Diff_SS	SD of absolute difference in speed between upstream and downstream stations
State	Traffic state including four types: FF, BN, BQ, CT ^a
Ramp	1 = if there is a ramp between upstream and downstream stations; 0 = otherwise
Weather	Weather condition including three types: Normal, Rain, Snow

Note: vph = vehicles per hour; SD = standard deviation; vpm = vehicles per mile.
^aBQ state was not observed in the dataset.

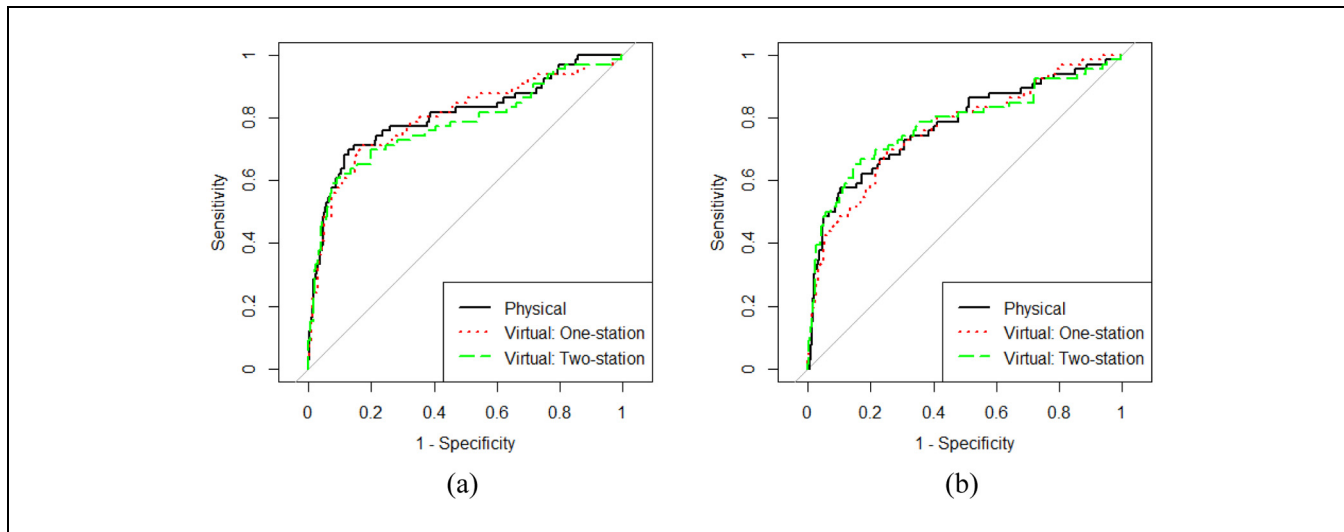


Figure 4. ROC curves for three models with different data sources from two time intervals: (a) 0–5 min before the crash occurrence and (b) 5–10 min before the crash occurrence.

ROC curves in Figure 4 do not show evident differences between three models at two time intervals as differences in AUC values are marginal. A larger AUC indicates better accuracy. Model P provides the best accuracy for the 0–5 min interval, but Model V1 provides

almost the same accuracy rate. For the 5–10 min interval, Model V2 outperformed Model P and V1. Generally, virtual station data provides comparable accuracy to the physical station data. The satisfactory performance of virtual station data indicates that the CTM with well-

Table 3. AUC for Three Models at Two Time Intervals

Time interval	Physical (P)	Virtual: one-station (V1)	Virtual: two-station (V2)
0–5 min	0.8068	0.8017	0.7798
5–10 min	0.7751	0.7643	0.7778

Table 4. Model Comparison for 5–10 min Interval

Physical (P)		Virtual: two-station (V2)	
Variable	Estimate	Variable	Estimate
Constant	–3.887	Constant	–2.765
DFA	0.000525	DFA	0.00235
Diff_FA	–0.000771	UFA	–0.00231
Diff_DA	0.0266	UDS	0.0716
DSCVS	4.869	Ramp	–1.129
		State: FF ^a	–
		State: BQ	2.946
		State: CT	1.666

^aFF state is the base level.

calibrated FDs can sufficiently capture the traffic conditions within a long segment. The comparison between AUCs also revealed that the accuracy of the 0–5 min interval is always better than that of the 5–10 min interval. It is expected that the crash occurrence should be more related to the traffic conditions as they become closer in time to the crash occurrence.

At the current stage, traffic agencies may need buffer time to detect crash risk, disseminate warnings, and implement traffic control strategies. Therefore, it is worthwhile to investigate the models for the 5–10 min interval. Because Model V2 provides the best performance in this interval, it was compared with Model P. Table 4 shows that the two models have very different sets of statistically significant variables. It should be noted that the two models were developed from different data sources. Both models include DFA as a significant variable, and both show consistent signs on the estimate coefficients. The results from Model P suggest that crash risk increases with an increase in the average flow and speed variation at the downstream station. However, crash risk decreases with an increase in the absolute flow difference between two stations. Model V2 suggests that crash risk increases with an increase in the average flow at the downstream station and the density variation at the upstream station. However, crash risk decreases with an increase in the average flow at the upstream station. Model V2 also identifies the impact of traffic states: the BQ and CT states are more crash-prone than the FF state. The small size of cells means that CTM may

capture local state transitions which cannot be captured by physical stations located far away. The negative sign of the ramp presence is counterintuitive, as it suggests that the presence of off-ramp would reduce crash risk. It is possible that the actual impact of the ramp is confounded by the traffic state.

Conclusion

Many real-time crash prediction studies have used traffic data collected from detector stations. The traffic data from these stations may not represent the actual traffic conditions contributing to crashes if the stations are located far from the actual crash location. Moreover, the spacing between stations could vary considerably from site to site, which compromises crash prediction accuracy when multiple sites are considered. This critical issue may render unreliable findings based on the traffic data collected directly from detector stations.

This paper attempted to exploit simulated traffic data and predict crash occurrence in a real-time fashion given distant stations. A macroscopic simulation model CTM was applied to instrument the freeway corridor with virtual detector stations that can be placed anywhere. It is expected that this simulated traffic data from virtual detector stations would yield more accurate prediction results than the traffic data from distant physical detector stations. Two time intervals were tested—0–5 min and 5–10 min before a crash occurrence—and both a one-station and a two-station setting were applied to determine which setting is more suitable in two different time interval choices. Models developed with physical station data were compared with those developed with virtual station data, leading to the conclusion that models developed with virtual station data provided a prediction accuracy rate similar to those developed with physical station data. The results suggest that simulated traffic data can be used when the spacing between detectors is large.

The proposed CTM approach can be applied in a real-time crash detection system. The application can vary according to the constraints and accuracy between one-station and two-station settings. The one-station setting can be used when a buffer time is needed for agencies to detect crash risk, take corresponding actions, and inform drivers. The two-station setting can be used when a buffer time is not necessary, such as in a connected and autonomous driving environment in which vehicles have the ability to make immediate maneuvers.

Future research should be directed toward furthering the development of the CTM approach. More importantly, the methodological advancement should be measured by improved flexibility to model traffic dynamics that correlate to crash risk, higher crash prediction accuracy, and reliability. One potential improvement could

be calibrating separate FDs under some safety-related circumstances such as inclement weather, low light conditions, or work zones. Another direction for future research is establishing the relationship between macroscopic traffic characteristics correlated with crash occurrence and microscopic traffic characteristics that correspond to safety surrogate measures (e.g. traffic conflicts). The empirical relationship may be explored and measured via simulation models, but theoretical development will require in-depth knowledge of traffic flow theory and its application in highway safety.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Xiao Qin, Zhi Chen; data collection: Zhi Chen, Yang Cheng; analysis and interpretation of results: Zhi Chen, Xiao Qin, Renxin Zhong; draft manuscript preparation: Zhi Chen, Xiao Qin, Renxin Zhong, Pan Liu. All authors reviewed the results and approved the final version of the manuscript.

References

- Abdel-Aty, M., N. Uddin, A. Pande, F. Abdalla, and L. Hsia. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record: Journal of the Transportation Research Board*, 2004. 1897: 88–95.
- Daganzo, C. F. The Cell Transmission Model: Network Traffic. *Transportation Research Part B: Methodological*, Vol. 29, No. 2, 1994, pp. 79–93.
- Muñoz, L., X. Sun, R. Horowitz, and L. Alvarez. Traffic Density Estimation with the Cell Transmission Model. *Proc., 2003 American Control Conference*, Vol. 5, IEEE, Denver, Colo., 2003, pp. 3750–3755.
- Muñoz, L., X. Sun, R. Horowitz, and L. Alvarez. Piecewise-Linearized Cell Transmission Model and Parameter Calibration Methodology. *Transportation Research Record: Journal of the Transportation Research Board*, 2006. 1965: pp. 183–191.
- Sumalee, A., R. X. Zhong, T. L. Pan, and W. Y. Szeto. Stochastic Cell Transmission Model (SCTM): A Stochastic Dynamic Traffic Model for Traffic State Surveillance and Assignment. *Transportation Research Part B: Methodological*, Vol. 45, No. 3, 2011, pp. 507–533.
- Hossain, M., and Y. Muromachi. A Bayesian Network Based Framework for Real-Time Crash Prediction on the Basic Freeway Segments of Urban Expressways. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 373–381.
- Sun, J., and J. Sun. A Dynamic Bayesian Network Model for Real-Time Crash Prediction Using Traffic Speed Conditions Data. *Transportation Research Part C: Emerging Technologies*, Vol. 54, 2015, pp. 176–186.
- Pande, A., and M. Abdel-Aty. Comprehensive Analysis of The Relationship Between Real-Time Traffic Surveillance Data and Rear-End Crashes on Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 2006. 1953: 31–40.
- Abdel-Aty, M. A., and R. Pemmanaboina. Calibrating a Real-Time Traffic Crash-Prediction Model Using Archived Weather and ITS Traffic Data. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 7, No. 2, 2006, pp. 167–174.
- Li, Z. B., W. Wang, R. Y. Chen, P. Liu, and C. C. Xu. Evaluation of the Impacts of Speed Variation on Freeway Traffic Collisions in Various Traffic States. *Traffic Injury Prevention*, Vol. 14, No. 8, 2013, pp. 861–866.
- Kwak, H.-C., and S. Kho. Predicting Crash Risk and Identifying Crash Precursors on Korean Expressways Using Loop Detector Data. *Accident Analysis & Prevention*, Vol. 88, 2016, pp. 9–19.
- Xu, C., P. Liu, and W. Wang. Evaluation of the Predictability of Real-Time Crash Risk Models. *Accident Analysis & Prevention*, Vol. 94, 2016, pp. 207–215.
- Chen, Z., X. Qin, and M. R. R. Shaon. Modeling Lane-Change Related Crashes with Lane-Specific Real-Time Traffic and Weather Data. *Journal of Intelligent Transportation Systems*, 2017, pp. 1–10.
- Xu, C., W. Wang, and P. Liu. A Genetic Programming Model for Real-Time Crash Prediction on Freeways. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 2, 2013, pp. 574–586.
- Abdel-Aty, M., N. Uddin, and A. Pande. Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. 1908: 51–58.
- Zheng, Z., S. Ahn, and C. M. Monsere. Impact of Traffic Oscillations on Freeway Crash Occurrences. *Accident Analysis & Prevention*, Vol. 42, No. 2, 2010, pp. 626–636.
- Hourdos, J., V. Garg, P. Michalopoulos, and G. Davis. Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations. *Transportation Research Record: Journal of the Transportation Research Board*, 2006. 1968: 83–91.
- Lee, C., M. Abdel-Aty, and L. Hsia. Potential Real-Time Indicators of Sideswipe Crashes on Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 2006. 1953: 41–49.
- Lee, C., B. Hellinga, and F. Saccomanno. Proactive Freeway Crash Prevention Using Real-Time Traffic Control. *Canadian Journal of Civil Engineering*, Vol. 30, No. 6, 2003, pp. 1034–1041.
- Roshandel, S., Z. Zheng, and S. Washington. Impact of Real-Time Traffic Characteristics on Freeway Crash Occurrence: Systematic Review and Meta-Analysis. *Accident Analysis & Prevention*, Vol. 79, 2015, pp. 198–211.
- Moore, J. E., G. Giuliano, and S. Cho. Secondary Accident Rates on Los Angeles Freeways. *Journal of Transportation Engineering*, Vol. 130, No. 3, 2004, pp. 280–285.
- Parker, S. T., and Y. Tao. WisTransPortal: A Wisconsin Traffic Operations Data Hub. *Proc., Ninth International Conference on Applications of Advanced Technology in Transportation (AATT)*, ASCE, New York, 2006, pp. 611–616.

23. Al-Deek, H. M., C. Venkata, and S. Ravi Chandra. New Algorithms for Filtering and Imputation of Real-Time and Archived Dual-Loop Detector Data in I-4 Data Warehouse. *Transportation Research Record: Journal of the Transportation Research Board*, 2004. 1867: 116–126.
 24. Zhong, R., C. Chen, A. H. Chow, T. Pan, F. Yuan, and Z. He. Automatic Calibration of Fundamental Diagram for First-Order Macroscopic Freeway Traffic Models. *Journal of Advanced Transportation*, Vol. 50, No. 3, 2016, pp. 363–385.
 25. Courant, R., K. Friedrichs, and H. Lewy. On the Partial Difference Equations of Mathematical Physics. *IBM journal of Research and Development*, Vol. 11, No. 2, 1967, pp. 215–234.
 26. Yeo, H., K. Jang, A. Skabardonis, and S. Kang. Impact of Traffic States on Freeway Crash Involvement Rates. *Accident Analysis and Prevention*, Vol. 50, 2013, pp. 713–723.
- The Safety Section (ANB00) peer-reviewed this paper (18-05725).*