Real-Time Crash Prediction Model
for Application to Crash Prevention in Freeway Traffic

Chris Lee
Department of Civil Engineering
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
Tel: (519) 888-4567 ext. 6596
Fax: (519) 888-6197
E-mail: chclee@uwaterloo.ca

Bruce Hellinga
Department of Civil Engineering
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
Tel: (519) 888-4567 ext. 2630
Fax: (519) 888-6197
E-mail: bhellinga@uwaterloo.ca

Frank Saccomanno
Department of Civil Engineering
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
Tel: (519) 888-4567 ext. 2631
Fax: (519) 888-6197
E-mail: saccoman@uwaterloo.ca

Words: 5,743 + 7 * 250 = 7,493 words

Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic

CHRIS LEE, BRUCE HELLINGA, AND FRANK SACCOMANNO

Department of Civil Engineering, University of Waterloo

Waterloo, Ontario, N2L 3G1, Canada.

The likelihood of a crash or crash potential is significantly affected by short-term turbulence of traffic flow. For this reason, crash potential must be estimated on a real-time basis by monitoring the current traffic condition. In this regard, a probabilistic real-time crash prediction model relating crash potential to various traffic flow characteristics which lead to crash occurrence, or "crash precursors", was developed. However, several assumptions were made in the development of this previous model that had not been clearly verified from either theoretical or empirical perspectives. Therefore, the objective of this study is to (1) suggest the rational methods by which crash precursors included in the model can be determined on the basis of experimental results; and (2) test the performance of the modified crash prediction model. The study found that crash precursors can be determined in an objective manner eliminating a characteristic of the previous model that the model results were dependent on analysts' subjective categorization of crash precursors.

In improving traffic safety on freeways, proactively preventing vehicle crashes may have much greater benefits than minimizing the consequences once a crash has occurred. In this paper, a crash is defined as an accident involving a vehicle collision. To implement crash prevention, it is necessary that the future occurrence of a crash can be anticipated on the basis of *hazardous* traffic flow conditions that are present prior to the occurrence of the crash. According to the National Academy of Engineering (*1*), precursors are "signals that illuminate system failure points with potential for future catastrophic loss". Precursors have been investigated to project future calamities and mitigate future risk exposure in many study areas – e.g. prediction of stock market crash in finance, prediction of the occurrence of earthquakes in geology, etc. In a similar manner, this study refers to the traffic conditions that exist prior to the occurrence of vehicle crashes as "crash precursors".

The identification of crash precursors from current traffic flow conditions is very important to predict the variation of crash potential over time and to establish real-time crash countermeasures to avoid the hazardous traffic condition leading to crashes. In this study, the term "crash potential" refers to the long-term likelihood that a crash will occur for given traffic, environment, and roadway conditions. Since crash potential is affected by many time-

dependent factors such as the variation of traffic flow, crash potential varies over time and therefore should be estimated in real time.

To reduce time-varying crash potential, most researchers have focused on timely detection of incidents. However, incident detection algorithms are unable to prevent the occurrence of primary crashes although they may help in reducing secondary crashes. Despite this inherent limitation of incident detection algorithm, a great deal of effort has been invested in developing these algorithms and much less effort invested methods in real-time crash prevention.

In real-time crash prevention, crash precursors based on real-time traffic measures are used to quantify crash potential. However, due to lack of real-time data in the past, most existing crash prediction models were not able to account for crash precursors in the prediction of crash occurrence. Instead, these models have used non-real-time and capacity-driven measures of traffic flow such as Average Annual Daily Traffic (AADT). Consequently, these models may be valuable for examining static, infrastructure based crash reduction measures such as paved shoulders, median barriers, etc. However, they are not helpful for evaluating the effect of real-time intervention measures such as those associated with Intelligent Transportation Systems (ITS) Advanced Traffic Management System (ATMS) concepts and services (*2*). Therefore, there is a need to develop a crash prediction model that estimates the variation of crash potential and enables us to evaluate the safety benefits of real-time crash prevention.

In this regard, we have identified a number of important crash precursors and developed a probabilistic real-time crash prediction model in our preliminary study (*3*). While this previous work demonstrated that a statistically significant real-time crash prediction model was possible, a number of assumptions were made in the development of this model. In particular, the assumptions were made with respect to the time duration over which the precursors were calculated and the categorization of the precursor variables. Thus, this study has the following objectives: 1) to suggest the rational methods by which observation time period duration and precursor categorization can be determined; and 2) to test the performance of the crash prediction model with the parameters modified using these rational methods.

This paper is organized into five sections. The second section reviews the past studies on crash precursors and real-time crash prediction model. The third section explains crash precursors and the structure of the proposed model. The fourth section suggests the methods to determine crash precursors in the model using real traffic flow

data and evaluates the performance of the model. Finally, the fifth section discusses the findings from the results and recommends future work.

REVIEWS OF PREVIOUS STUDIES

By far, most studies of crash precursors have focused on the behavior of individual drivers/vehicles. For example, Krishnan et al. (*4*) claimed that braking capability of cars, response time of drivers, speed of cars, type of cars and mass of cars are important factors affecting crashes. They used these factors as criteria for designing their rear-end collision-warning system. Smith et al. (*5*) suggested that the headway of two successive cars and the variation of headway have major impact on crash potential. They classified the crash risk into four levels according to these two factors. However, their results are based on the experiments for the selected driver group and it is uncertain that the defined risk levels are generally applicable to different driver groups. Furthermore, it appears that as a result of many other factors which cannot be easily measured – e.g. driver's characteristics, driving state, vehicle characteristics, etc, developing a general relationship using this approach is likely very difficult.

It may be advantageous to identify more aggregated relationship between crash potential and the "collective" behavior of individual drivers – i.e. traffic flow characteristics. There are a few studies that have presented statistical links between real-time traffic flow conditions prior to crash occurrence and crash potential.

Oh et al. (*6*) found that the standard deviation of speed 5 minutes prior to crash occurrence is the best indicator that distinguishes disruptive conditions (conditions leading to crash occurrence) from normal conditions in their analysis using loop detector data of a freeway section in California. Using this indicator, they developed probability density functions to estimate whether the current traffic condition belongs to either normal or disruptive traffic conditions. They concluded that reducing the variation in speed generally reduces the likelihood of freeway crashes.

Despite their innovative approach, the study displays some limitations. First, only a single measure of traffic performance (standard deviation of speed) was used to predict the crash likelihood. Since crashes normally occur as a result of complex interaction of many traffic and environmental factors, it is questionable whether the single variable can sufficiently explain a broad spectrum of pre-crash conditions. Second, the measure of crash likelihood estimated from probability density function overlooked such exposures as volume, distance of travel and so on. To

control for these external conditions, the variation of exposures over space and time must be taken into account in the probability density function.

Similar to this study, Kirchsteiger (*7*) described the distribution of accident precursors in generalized probability function such as the Gamma distribution. Although his study used industrial accident database instead of traffic accident data, his approach is very similar to the analysis of traffic accidents. In particular, he suggested that frequency of accidents in the observation time period is described as the product of two frequencies: 1) frequency of precursor and 2) conditional frequency of accident, given a precursor.

In a recent approach, Lee et al. (*3*) proposed a probabilistic crash prediction model using 13 months of loop detector data from an urban freeway in Toronto. The details of this model are explained in the next section. However, the model also displays some limitations. First, the determination of precursor variables is subjective. The model made use of the traffic factors 5 minutes prior to crash occurrence but it was not verified whether 5 minutes are the most desirable observation time period. Second, the model used a number of categorical variables but the study did not clearly explain how to choose the optimal number of categories and the boundary values of each category. Finally, the study failed to show the sensitivity of different boundary values that are determined subjectively to the model performance. These issues will be addressed in the following sections.

STRUCTURE OF PROPOSED MODEL

This study uses real-time traffic flow characteristics to explain the effect of traffic performance on crash occurrence. These characteristics are reflected by crash precursors. However, to explain the exclusive effect of crash precursors, crash frequency should be controlled for external factors. These external factors include road geometry and time of day (or level of congestion) which have been commonly used in the past crash prediction models. It has been logically and empirically proven that these factors have significant impacts on crash occurrence in the past studies. Also, exposure measures should be combined with crash data so that the effects of various freeway and traffic elements on crash potential can be explicitly compared within or between classifications of interest (*8*). Similar to most other crash prediction models, the proposed model expresses crash frequency as a function of a variety of traffic and environmental characteristics as follows:

Crash frequency = f (crash precursors, external control factors, exposure)

Using this functional relationship, the model is calibrated using actual crash data and the effects of crash precursors on crash potential can be examined. In the next subsections, the calculation of crash precursors in the above function and the model specification are described.

Specification of Crash Precursors

In our previous study (*3*), we identified three crash precursors representing the traffic flow conditions prior to the crash occurrence: (1) the average variation of speed on each lane (CVS_1); (2) the average variation of speed difference across adjacent lanes (CVS_2); and (3) traffic density (*D*). Variation of speed is measured by the coefficient of variation of speed (CVS) (= standard deviation of speed / average speed) computed over the given observation time slice duration. The mathematical expression of these three precursors is described in Lee et al. (*3*).

CVS_2 was formulated as a surrogate measure of lane change behavior in the assumption that lane changing tends to increase crash potential. However, in spite of its statistical significance in previous study, this current study found that CVS_2 does not have a direct impact on crash potential because there was no significant difference in its values calculated for crash cases and non-crash cases. The details of the comparison between crash and non-crash cases are explained in the next section. Therefore, CVS_2 was eliminated from the model. Since only one variation of speed is left in the model, CVS_1 is re-named as *CVS*.

In addition to the existing crash precursors, *CVS* and *D*, we considered a new crash precursor to reflect the impact of a traffic queue on crash occurrence. This impact can be reflected by the difference of speeds at upstream and downstream ends of road sections. The underlying principle of this variable is that as the difference in speed increases, there is more abrupt change in traffic condition within the road section – i.e. queue formation or dissipation. For example, if the speed at the downstream end is significantly lower than the speed at the upstream end for a prolonged time period, a tail of a queue is likely to exist somewhere within the section. Conversely, relatively high speed at the downstream end and low speed at the upstream end indicates the dissipation of a queue. In either case, drivers are required to react promptly to adjust their speed and these conditions are likely to increase crash potential. This additional precursor (*Q*) can be expressed by Equation 1.

$$Q = \left| \bar{s}_1 - \bar{s}_2 \right| = \left| \frac{t_p}{\Delta t} \sum_{t=t^*-\Delta t}^{t^*} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} s_{1i}(t) \right) - \frac{t_p}{\Delta t} \sum_{t=t^*-\Delta t}^{t^*} \left(\frac{1}{n_2} \sum_{i=1}^{n_2} s_{2i}(t) \right) \right| \qquad (1)$$

where,

Q : average speed difference between upstream and downstream ends of road section (km/hr);

$\bar{s}_1, \bar{s}_2$: average speed computed over period of Δt at upstream and downstream ends of road section,

      respectively (km/hour);

t_p : time interval of observation of speed (seconds);

Δt : observation time slice duration (seconds);

t^* : actual time of crash occurrence;

$s_{1i}(t)$: speed on lane i at time t at upstream end of road section (km/hour);

$s_{2i}(t)$: speed on lane i at time t at downstream end of road section (km/hour);

n_1, n_2 : number of lanes at upstream and downstream ends of road section, respectively.

To illustrate the property of this new crash precursor, Equation 1 was applied to a section of the Gardiner Expressway in Toronto, Canada. Q was calculated at every 20-second interval with Δt assumed to be 2 minutes. As shown in Figure 1, Q clearly shows the patterns of typical queue formation and dissipation in daily traffic. Particularly, Q characterizes high chances of queue propagation during afternoon peak period. On the other hand, Q is generally small and constant during non-peak period when a queue is highly unlikely to form.

Exposure

As explained in the model structure, exposure is included to reflect frequency of traffic events that create chances of crash occurrence. In this study, exposure is described as the product of daily traffic volume and the length of each road section. These are split into the volume-kilometers according to the probabilities of occurrence of crash precursor values and external control factors in daily traffic. However, the weather was excluded in external control factors since the available weather data do not adequately reflect the actual weather condition at the time of crashes.

By definition, crash rate is crash frequency divided by exposure. Thus, exposure must be determined for a given crash frequency. This means that individual crashes should be classified into generalized crash types based on their common characteristics. In this study, the crash types are characterized by typical traffic conditions when crashes occurred, such as (1) crash precursor values prior to crash occurrence; (2) congestion level (peak/off-peak period) at the time of crashes; and (3) road section type (straight section or merge/diverge section), traffic volume, and length of road section type where crashes occurred. For the ease of determining the exposure for given crash precursor values, precursors should be categorized into a number of discrete levels. For this reason, precursors are expressed in categorical variables instead of continuous variables in the model. For example, the exposure for crash type A can be estimated based on the levels of crash precursors, road geometry and congestion level corresponding to crash type A as shown in Equation 2.

$$EXP_A = p(CVS_A) \cdot p(D_A) \cdot p(Q_A) \cdot p(P_A) \cdot V_A \cdot L_A \cdot T \tag{2}$$

where,

EXP_A	: exposure for crash type A (vehicle-kilometers of travel);
CVS_A, D_A, Q_A:	: the levels of CVS, D and Q for crash type A, respectively;
$p(CVS_A), p(D_A), p(Q_A)$:	probabilities that the levels of CVS_A, D_A and Q_A occur in daily traffic, respectively;
$p(P_A)$	: proportion of volume during peak or off-peak periods in daily traffic for crash type A;
V_A	: average annual daily traffic of the road section (vehicles/day) for crash type A;
L_A	: length of road section type (km) for crash type A;
T	: total observation time period (number of days).

Model Specification

To analyze the effects of crash precursors and external control factors on crash potential, a probabilistic model of crash prediction was developed. The model estimates the relationship between crash frequency and the variables discussed in previous sections.

In this study, an aggregate log-linear model is developed since it allows us to investigate the nature of the relationship between selected precursors and frequency of crashes adjusted by the appropriate level of exposure. The

log-linear model in this study has only a limited number of parameters instead of all possible parameters to achieve a parsimonious representation of the data. In this analysis, a first-order log-linear model of crash prediction was derived as follows.

First, the multiplicative models are more preferable to additive (linear) models in describing the nature of crash occurrence because 1) they can describe a non-linear relationship between mutually independent variables and, 2) they can avoid the problem of negative-value prediction (*9*). Thus, the crash rate can be described in the following functional form:

$$\text{Crash Rate} = \frac{F}{EXP^{\beta}} = f\left(\theta \cdot \lambda_{CVS(i)} \cdot \lambda_{D(j)} \cdot \lambda_{Q(k)} \cdot \lambda_{R(l)} \cdot \lambda_{P(m)}\right) \qquad (3)$$

where,

F : the expected number of crashes over the analysis time frame;

EXP : the exposure in vehicle-kilometers of travel;

β : the parameter for the exposure;

θ : constant;

$\lambda_{CVS(i)}$: effect of the crash precursor variable *CVS* having *i* levels;

$\lambda_{D(j)}$: effect of the crash precursor variable *D* having *j* levels;

$\lambda_{Q(k)}$: effect of the crash precursor variable *Q* having *k* levels;

$\lambda_{R(l)}$: effect of road geometry (control factor) having *l* levels;

$\lambda_{P(m)}$: effect of time of day (control factor) having *m* levels.

It should be noted that crash precursors and external control factors are categorical variables whereas exposure is continuous variable. To express the above equation in a linear function of independent variables, the factors are converted to logarithmic terms as follows:

$$\frac{F}{EXP^{\beta}} = \exp\!\left(\theta + \lambda_{CVS(i)} + \lambda_{D(j)} + \lambda_{Q(k)} + \lambda_{R(l)} + \lambda_{P(m)}\right)$$

$$\ln\!\left(\frac{F}{EXP^{\beta}}\right) = \theta + \lambda_{CVS(i)} + \lambda_{D(j)} + \lambda_{Q(k)} + \lambda_{R(l)} + \lambda_{P(m)}$$

$$\ln(F) - \beta\ln(EXP) = \theta + \lambda_{CVS(i)} + \lambda_{D(j)} + \lambda_{Q(k)} + \lambda_{R(l)} + \lambda_{P(m)}$$

$$\ln(F) = \theta + \lambda_{CVS(i)} + \lambda_{D(j)} + \lambda_{Q(k)} + \lambda_{R(l)} + \lambda_{P(m)} + \beta\ln(EXP) \qquad (4)$$

To estimate the parameters in Equation 4, the model is calibrated for actual crash data using the maximum likelihood estimate (MLE) method. The MLE method runs an iterative process to fit the estimated data to the observed data. This fitting process continues until the difference between the current and previous estimates converges to the pre-specified error level.

MODEL CALIBRATION

This section describes the data used for the calibration of the model and suggests the methods to determine several important parameters to calculate crash precursors such as 1) actual time of crash, 2) observation time slice duration, and 3) number of categories and boundary values for each category of crash precursors.

Description of Data

To calibrate the proposed model, this study used incident logs and traffic flow data extracted from loop detectors along a 10-km stretch of the Gardiner Expressway in Toronto, Canada. A total of 38 loop detector stations are located along this stretch of freeway as shown in Figure 2. The data were collected for weekdays over a 13-month period from January 1998 to January 1999. A total of 234 crashes were confirmed by operators at the traffic control center on this section of the roadway during the study period. The crashes on ramps were not considered because they are likely to be more influenced by site-specific characteristics such as geometric design and traffic operation (*9*).

Determination of Actual Time of Crash

In this study, the actual time of crash (t^*) was estimated from the analysis of changes in detector speed profiles. To illustrate the method of estimation, the change of speed on the road section between two loop detectors upstream and downstream of the crash site is described on a time-space diagram as shown in Figure 3.

The figure shows that before the crash occurs at the location marked by "X" between a pair of loop detectors, all individual cars are assumed to travel at the speed in normal traffic condition (s_n) as indicated by arrows. After the crash occurs at "X" (separated by d^* from upstream detector) at time t^*, the vehicles upstream of the crash site experience delays and their speed suddenly drops to speed in the congested queue (s_q). On the contrary, the speed of vehicles downstream of the crash site increases to free-flow speed (s_f) due to a decrease in volume. As a forward-moving shock wave (u) passes over the downstream detector at time t_d and a backward-moving shock wave (ω) passes over the upstream detector at time t_u, the speed change can be observed at both detectors as shown in the two figures below the time-space diagram.

From these figures, we can see that it is normally easier to detect t_u than t_d since there are more noticeable changes in speed at upstream detector than downstream detector. This is because the impact of a queue formed by lane blockages on traffic flow disruption is more severe than the impact of the reduction in volume downstream of the crash site. Also since precursors represent the traffic condition before vehicles reach the crash site, we need to observe the condition immediately upstream of the crash site. Thus, t_u was used as a surrogate measure of t^*.

To evaluate the accuracy of t_u in estimating t^*, the expected error caused by the travel time of the backward-moving shock wave ($= t_u - t^*$) can be calculated as follows. If the crashes are likely to occur equally at any point of a road section with the length of n unit (i.e. the probability of crash occurrence at any point is the same), the expected location of crashes is estimated by Equation 5.

$$E[d^*] = \frac{1}{n} \times 1 \text{ unit} + \frac{1}{n} \times 2 \text{ unit} + \cdots + \frac{1}{n} \times n \text{ unit} = \frac{1+2+\ldots+n}{n}$$

$$= \frac{\dfrac{n(n+1)}{2}}{n} = \frac{n+1}{2} \cong \frac{n}{2} \text{ if } n \text{ is very large.}$$

(5)

Also, the past studies observed that the speed of the backward-moving shock wave ranges from 10 km/hr to 30 km/hr on urban freeways (*10-13*). In this study, the average speed of the backward-moving shock wave ($\overline{\omega}$) is assumed to be 20 km/hr. For example, if the length of the section (*n*) is 500 meters which is a typical spacing of detectors on urban freeways, the expected error (ε) can be calculated as follows:

$$\varepsilon = E[t_u - t*] = \frac{E[d*]}{E[\omega]} = \frac{n/2}{\overline{\omega}} = \frac{(0.5/2)\,\text{km}}{20\,\text{km/hour}} = 45 \ \text{seconds} \tag{6}$$

Considering the fact that typical polling intervals of loop detectors are 20~30 seconds, the expected errors are only 2~3 polling intervals. Due to relatively low expected errors, this study assumes that the errors do not significantly affect the calculation of crash precursors and t_u can be used as a good estimate of $t*$.

Determination of Observation Time Slice Duration

In our previous study (*3*), observation time slice duration (Δt) prior to crash occurrence was assumed to be 5 minutes. However, this duration was arbitrarily chosen on the basis of subjective judgment rather than empirical results. In this study, a more objective method was developed to determine Δt. The premise of the method is that Δt should be chosen to maximize the difference between precursor values calculated for crash cases and non-crash cases. For the analysis, crash precursors (*CVS*, *D* and *Q*) were calculated for the sample of 234 crash cases and 234 non-crash cases. Non-crash cases contain the data collected at the same road sections for the same time periods under the same weather condition as crash cases, but in different days when crashes did not occur. In this way, other traffic environmental factors such as road geometry, weather and typical traffic pattern are assumed to be controlled.

The objective of this analysis is to identify Δt which maximizes the difference in crash precursor values between crash and non-crash cases. Total difference can be directly obtained from the sum of absolute differences in precursor values between the two cases. However, since the variation of speed (*CVS*) tends to increase with Δt, the "standardized" differences must be used rather than absolute differences. Consequently, the distributions of crash and non-crash cases were compared based on the frequency of precursors for given ranges of precursor values. The number of intervals within the given ranges were determined such that the frequency of precursors is reasonably dispersed. The frequency differences were calculated using the following expression:

$$\delta = \sum_{i=1}^{I}\left(f_i - f'_i\right)^2 \tag{7}$$

where,

δ : difference in frequency of precursor values between crash and non-crash cases;

I : total number of intervals within a given range;

f_i, f'_i : frequency of precursor values for an interval i in crash and non-crash cases, respectively.

As a result, it was found that the frequency differences in the three precursor variables (*CVS*, *D* and *Q*) between crash and non-crash cases were maximum at Δt = 8, 3 and 2 minutes, respectively, as shown in Figure 4. This implies that the variation of speed (*CVS*) has relatively a longer-term effect on crash potential than *D* and *Q*. Although it is uncertain whether the maximum value identified within the range (1~20 minutes) actually represents a global maximum, it is hard to believe that the traffic condition more than 20 minutes prior to crash occurrence has a significant impact on crashes. Thus, the above-mentioned values were chosen as optimal observation time slice durations.

Categorization of Crash Precursors

In categorizing crash precursors, we need to define the level of crash precursors based on the distribution of *normal traffic flow condition* in daily traffic. For this purpose, 24-hour traffic data on two typical weekdays in clear weather condition when no crash has occurred was extracted from loop detectors. As shown in Figure 5, the distribution of precursor values from the 24-hour data is substantially different from that of the crash data. The figure shows that the frequency of high crash precursor values was relatively higher in the crash data than the 24-hour data. This implies that crash precursor values were relatively higher when crashes occurred compared to when crashes did not occur. Also, it should be noted that the disparity between the distribution of the 24-hour data and that of the crash data is much larger than the disparity between the two distributions of the 24-hour data. This indicates that crash precursors are good indicators of discriminating between crash and non-crash conditions.

In the categorization, the number of categories and the boundary values (or proportion of each category) must be determined. However, it is difficult to determine these factors objectively from the distribution of crash precursor

values. Instead, the proposed log-linear model was calibrated for different cases of categorization. Then, the performance of the calibrated log-linear models was evaluated in terms of 1) overall model fit, 2) the statistical significance of coefficients and 3) consistency of coefficients with the order of levels of crash precursors (i.e. high-level precursors have higher impact on crash potential than low-level precursors, vice versa). The categorization that produced the best model performance was chosen as the most suitable categorization.

In this study, the number of categories was considered varying from 2 to 4 inclusive. To insure statistical stability, total number of cells in a log-linear model should be less than the number of samples used to calibrate the model. Therefore, given that the calibration sample consisted of 234 observations, more than 4 categories were not considered. For each number of categories (2, 3 or 4), boundary values for the precursor variables were selected to achieve the specified proportions of crash precursors for each level in the 24-hour data under normal traffic condition as shown in Table 1. In most cases, the proportion of the highest level was set lower than all other lower levels in order to reflect rare occurrence of high-level precursors in normal traffic conditions. These proportions will affect the calculation of both crash frequency and exposure for each category as defined in Equation 2 and 3.

As a result of the calibration, as the number of categories increases, both log-likelihood ratio and Pearson chi-square statistics decrease, which means the model fits to the observed data better. However, it should be noted that this result stems from the fact that as the number of categories increases, the increased number of zero cells in a contingency table has a more dominant effect on overall model fit. But it is unclear whether these zero cells actually reflect that no crash occurred under the given circumstances or they are the result of missing data. With this uncertainty, increasing the number of categories does not necessarily improve the model performance.

For this reason, the statistical significance and consistency of coefficients also need to be checked. As a result, for 2 and 4 categories, some of the coefficients were insignificant and also they were inconsistent with the order of levels. Only for 3 categories, were all the coefficients statistically significant at a 95% confidence level and consistent with the order of levels in all 9 cases as shown in Table 2. Also, log-likelihood ratio and Pearson chi-square statistics were low, and p-values at a 95% confidence level were large in all cases. This means that there is no significant difference between observed and predicted crash frequency and therefore, the model fits the observed data well. This result implies that the performance of the log-linear model is not sensitive to our subjective categorization for 3 categories. Among 9 different cases, it was found that the proportions of 50%(low)-30%(intermediate)-20%(high) produced the best model fit (i.e. the lowest log-likelihood ratio and Pearson chi-

square). The boundary values for each crash precursor in this case are as follows as shown in Table 1: 0.056 and 0.074 for *CVS*; 16.4 and 25.8 veh/km for *D*; and 2.7 and 8.3 km/hr for *Q*.

In Table 2, lower coefficients indicate less impact on crash potential with respect to aliased cells. This result suggests that high-level crash precursors contribute to higher crash potential than low-level crash precursors. The control factors such as road geometry and time of day also have significant effect on crash potential. The result indicates that crash is more likely to occur on the road sections with on-ramp or off-ramp and during peak period. Finally, as expected, crash potential increases with exposure.

CONCLUSIONS AND RECOMMENDATIONS

This paper suggests the rational methods by which crash precursors are determined from experimental results and also evaluates the performance of the crash prediction model for different assumptions of categorized crash precursors. The findings from this study are summarized as follows:

1. The main criterion for selecting crash precursors is that the distribution of precursor values for traffic conditions when crashes occurred should be significantly different from the distribution of precursor values for normal traffic condition.

2. The difference between the speed at the upstream detector and the speed at the downstream detector was significantly higher when crashes occurred. This implies that the abrupt transition of speed within the road section, i.e. the formation and dissipation of a traffic queue, has positive effects on crash occurrence.

3. The time when the speed abruptly drops at the detector station immediately upstream of crash site is considered to be a good estimate of actual time of crashes. This speed drop occurs when a queue forms after the crash occurrence and the backward-moving shock wave passes over the nearest upstream detector station.

4. The observation time slice duration (Δt) prior to crash occurrence is determined such that the difference in distribution of crash precursor values for given Δt between crash and non-crash cases is maximized. It was found that the optimal observation time slice durations were different for each selected crash precursor – i.e. some precursors need to be observed for longer time period than other precursors to investigate their impact on crash occurrence.

5. The categorization of crash precursors is determined based on overall fit of crash prediction model, the statistical significance of coefficients and consistency of coefficients with the order of levels of crash precursors. In the analysis, three categories appear to be the most suitable to explain differential impact of precursors in different levels on crash potential as the model estimates showed consistent results for any combination of boundary values. Thus the performance of the model was reliable, not being affected by subjective categorization of crash precursors.

Having calibrated the model for historical crash data, the proposed model can be used to predict crash potential in real-time on the basis of the current traffic flow data. For the next step, we need to apply this model to actual traffic condition and examine how the crash potential estimated by the model can help reduce the crash potential and improve the safety of freeway traffic.

Since we can predict the crash potential on a real-time basis using the proposed model, we can also implement the automated real-time countermeasures to reduce crash potential such as variable speed limit. This certainly saves manual human intervention and effectively controls the traffic flow to prevent crashes. For the systematic implementation of real-time countermeasures, it is necessary to classify crash potential into different levels of risk tolerance. For example, variable speed limit is in operation only when the estimated crash potential exceeds the specified threshold value of risk tolerance.

Furthermore, we can also assess the safety benefit of an automated traffic control using the proposed crash prediction model. For one thing, the simulation can be performed before and after implementing the automated traffic control. From the simulation results, the impact of a change in driver behavior caused by this external control on the variation of traffic flow and real-time crash potential can be examined. This before-and-after study evaluates whether this automated traffic control can effectively reduce the overall crash potential on freeways for given conditions.

REFERENCES

1. *Management of Catastrophic Precursors: A Cross-Industry Analysis.* National Academy of Engineering, 2001.

2. Hughes, R. and F. Council. On Establishing the Relationship(s) between Freeway Safety and Peak Period Operations: Performance Measurement and Methodological Considerations. Presented at the 78th Annual Meeting of Transportation Research Board, Washington, D.C., 1999.

3. Lee, C., F. Saccomanno, and B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways. In *Transportation Research Record 1784*, TRB, National Research Council, Washington, D.C., 2002, pp. 1-8.

4. Krishnan, H., S. Gibb, A. Steinfeld, and S. Shladover. Rear-End Collision-Warning System. In *Transportation Research Record 1759*, TRB, National Research Council, Washington, D.C., 2002, pp. 52-60.

5. Smith, D. L., W. G. Najm, and R. A. Glassco. The Feasibility of Driver Judgment as a Basis for Crash Avoidance Database. In *Transportation Research Record 1784*, TRB, National Research Council, Washington, D.C., 2002, pp. 9-16.

6. Oh, C., J. Oh, S. Ritchie, and M. Chang. Real-Time Estimation of Freeway Accident Likelihood. Presented at the 80th Annual Meeting of Transportation Research Board, Washington, D.C., 2001.

7. Kirchsteiger, C. Impact of Accident Precursors on Risk Estimates from Accident Databases. *Journal of Loss Prev. Process Ind.*, Vol. 10, No. 3, 1997, pp. 159-167.

8. Chang, M. Conceptual Development of Exposure Measures for Evaluating Highway Safety. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1982, pp. 37-42.

9. Jovanis, P. P. and H. Chang. Modeling the Relationship of Accidents to Miles Traveled. In *Transportation Research Record 1068*, TRB, National Research Council, Washington, D.C., 1986, pp. 42-51.

10. Messer, C. J., C. L. Dudek, and J. D. Friebele. Method for Predicting Travel Time and Other Operational Measures in Real-Time During Freeway Incident Conditions. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1976, pp. 1-16.

11. Koshi, M., M. Iwasaki, and I. Ohkura. Some Findings and an Overview on Vehicular Flow Characteristics. Proceedings of the 8th International Symposium on Transportation and Traffic Theory, 1983, pp. 403-426.

12. Hurdle, V.F. and B. Son. Road Test of a Freeway Model. *Transportation Research A 34*, 2000, pp. 537-564.

13. Cassidy, M. and M. Mauch. An Observed Traffic Pattern in Long Freeway Queues. *Transportation Research A 35*, 2001, pp. 143-156.

LIST OF TABLES AND FIGURES

TABLE 1 Cases of Categorization and Boundary Values

No. of Categories	Assumed Proportions[*]	*CVS*			*D* (veh/km)			*Q* (km/hr)		
		B1[**]	B2[**]	B3[**]	B1	B2	B3	B1	B2	B3
2 Categories	50/50	0.056	-	-	16.4	-	-	2.7	-	-
	60/40	0.060	-	-	18.0	-	-	3.5	-	-
	70/30	0.065	-	-	20.3	-	-	4.7	-	-
	80/20	0.074	-	-	25.8	-	-	8.3	-	-
3 Categories	20/60/20	0.046	0.074	-	13.2	25.8	-	1.1	8.3	-
	20/50/30	0.046	0.064	-	13.2	20.3	-	1.1	4.7	-
	30/50/20	0.048	0.074	-	14.2	25.8	-	1.6	8.3	-
	33/33/33	0.049	0.063	-	14.5	19.4	-	1.7	4.2	-
	40/40/20	0.052	0.074	-	15.2	25.8	-	2.1	8.2	-
	40/30/30	0.052	0.064	-	15.2	20.3	-	2.1	4.7	-
	50/30/20	0.056	0.074	-	16.4	25.8	-	2.7	8.3	-
	50/20/30	0.056	0.064	-	16.4	20.3	-	2.7	4.7	-
	60/20/20	0.060	0.074	-	18.0	25.8	-	3.5	8.3	-
4 Categories	40/20/20/20	0.052	0.060	0.074	15.2	18.0	25.8	2.1	3.5	8.3
	30/30/20/20	0.048	0.060	0.074	14.2	18.0	25.8	1.6	3.5	8.3
	30/20/30/20	0.048	0.056	0.074	14.2	16.4	25.8	1.6	2.7	8.3
	25/25/25/25	0.047	0.056	0.068	13.6	16.4	22.3	1.3	2.7	5.9
	20/20/40/20	0.046	0.052	0.074	13.2	15.2	25.8	1.1	2.1	8.3
	20/30/30/20	0.046	0.056	0.074	13.2	16.4	25.8	1.1	2.7	8.3
	20/40/20/20	0.046	0.060	0.074	13.2	18.0	25.8	1.1	3.5	8.3

[*]The first number denotes the percentage of the lowest level, the second number the percentage of the second lowest level, and so on.
[**]B1 denotes the boundary value between the lowest level and the second lowest level for given crash precursor, B2 the boundary value between the second lowest level and the third lowest level, and so on.

Sample Illustration:

For example, in the case of 3 categories, 20/60/20 indicates that when the crash precursor values are ordered from the smallest to the largest, the first 20 percentile of crash precursor values represents "low" level, the next 60 percentile "intermediate" level and the remaining 20 percentile "high" level. The boundary values between "low" and "intermediate" levels for *CVS*, *D* and *Q* are 0.046, 13.2 and 1.1, respectively. The boundary values between "intermediate" and "high" levels for *CVS*, *D* and *Q* are 0.074, 25.8 and 8.3, respectively. Thus, the criteria of categorizing crash precursor values are as follows:

Low level:	$CVS \leq 0.046$	$D \leq 13.2$	$Q \leq 1.1$
Intermediate level:	$0.046 < CVS \leq 0.074$	$13.2 < D \leq 25.8$	$1.1 < Q \leq 8.3$
High level:	$CVS > 0.074$	$D > 25.8$	$Q > 8.3$

TABLE 2 Estimated Parameters of Log-linear Model (Three Categories)

Parameter	Assumed Proportions (Low/Intermediate/High)								
	20/60/20	20/50/30	30/50/20	33/33/33	40/40/20	40/30/30	50/30/20	50/20/30	60/20/20
θ	2.8920	3.0353	2.9050	0.5366	2.8272	1.0898	2.6569	2.5347	2.8191
$\lambda_{CVS=1}$	-4.9018	-4.4238	-3.2432	-2.8002	-3.0481	-3.6619	-3.3065	-3.2087	-2.3741
$\lambda_{CVS=2}$	-1.4274	-2.3140	-1.7207	-1.7216	-2.0652	-1.9155	-1.8415	-1.9522	-1.7157
$\lambda_{CVS=3}$*	0	0	0	0	0	0	0	0	0
$\lambda_{D=1}$	-1.3901	-1.4572	-1.4367	-1.4831	-1.7390	-2.7063	-2.3797	-2.3627	-0.9045
$\lambda_{D=2}$	-0.3733	-1.7751	-0.5753	-1.1818	-0.9418	-1.1090	-0.7088	-0.8791	-0.5833
$\lambda_{D=3}$*	0	0	0	0	0	0	0	0	0
$\lambda_{Q=1}$	-2.7247	-2.5041	-2.1406	-2.0095	-2.1778	-3.0160	-2.6859	-2.6107	-1.6682
$\lambda_{Q=2}$	-1.0554	-2.2371	-1.3869	-1.8993	-1.7864	-1.9362	-1.4794	-1.9960	-1.4076
$\lambda_{Q=3}$*	0	0	0	0	0	0	0	0	0
$\lambda_{R=0}$	-0.4171	-0.8405	-0.5613	-2.7014	-0.9062	-2.5029	-0.9916	-1.1961	-0.4728
$\lambda_{R=1}$*	0	0	0	0	0	0	0	0	0
$\lambda_{P=0}$	-0.4604	-0.5777	-0.4877	-1.0353	-0.5468	-0.9743	-0.4929	-0.6406	-0.4368
$\lambda_{P=1}$*	0	0	0	0	0	0	0	0	0
β (Exposure in 10^9 veh-km)	0.0075	0.0367	0.0228	0.1502	0.0710	0.1715	0.0964	0.0682	0.0171
Likelihood Ratio	87.4	62.9	85.7	57.2	62.8	54.6	44.5	66.6	106.27
p-value (α=0.05)	0.77	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.27

* This cell serves as the basis against which log-linear parameters are applied to obtain crash frequency for any combination of crash precursors. The cell is called the "aliased" cell.

Description of Parameters:

θ: Constant;

$\lambda_{CVS=1}$, $\lambda_{CVS=2}$, $\lambda_{CVS=3}$: Effect of *CVS* (=1 (low), =2 (intermediate), =3 (high));

$\lambda_{D=1}$, $\lambda_{D=2}$, $\lambda_{D=3}$,: Effect of *D* (=1 (low), =2 (intermediate), =3 (high));

$\lambda_{Q=1}$, $\lambda_{Q=2}$, $\lambda_{Q=3}$: Effect of *Q* (=1 (low), =2 (intermediate), =3 (high));

$\lambda_{R=0}$, $\lambda_{R=1}$: Effect of road geometry (=0 (straight section), =1 (merge/diverge section));

$\lambda_{P=0}$, $\lambda_{P=1}$: Effect of time of day (=0 (off-peak), =1 (peak));

β: Coefficient for exposure.

<u>Sample Calculation:</u>

At current time t^* (during peak period), if $CVS(t^*) = 0.04$, $D(t^*) = 10$ veh/km, and $Q(t^*) = 1$ km/hr, then the crash potential on the merge road section with exposure of 1×10^9 vehicles-km of travels over a 13-month period can be estimated as follows:

If the proportions of crash precursors are assumed to be 20/60/20, categories for *CVS*, *D*, and *Q* are 1 (low level) as the above values are lower than the boundary values between low level and intermediate level.

$$F(t^*) = \exp(\theta + \lambda_{CVS=1} + \lambda_{D=1} + \lambda_{Q=1} + \lambda_{R=1} + \lambda_{P=1} + \beta \cdot \ln(EXP))$$
$$= \exp(2.8920 - 4.9018 - 1.3901 - 2.7247 + 0 + 0 + 0.0075 \cdot \ln(1)) = 2.2 \times 10^{-3} \text{ crashes}$$

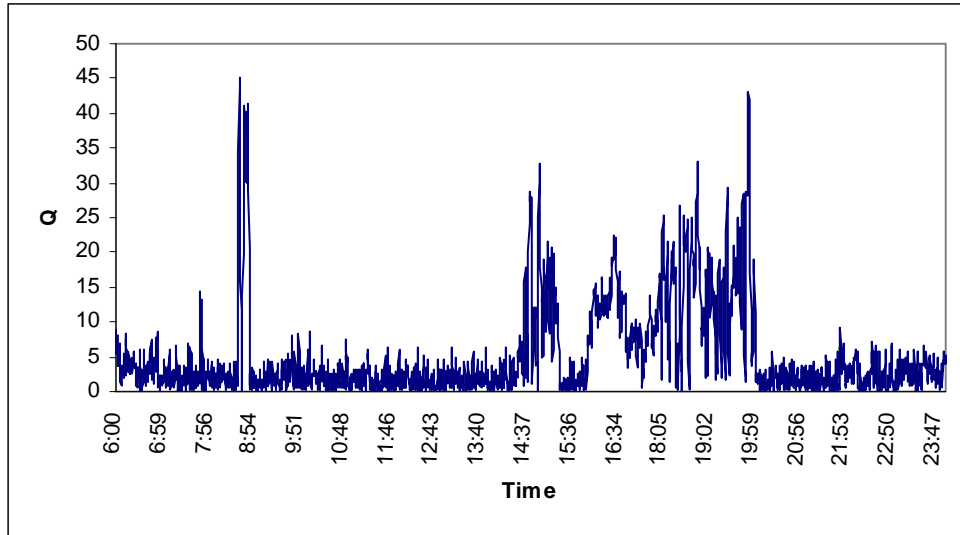Crash Potential $(t^*) = F(t^*) / EXP = 2.2 \times 10^{-3} / 10^9 = 2.2 \times 10^{-12}$ crashes/veh-km.
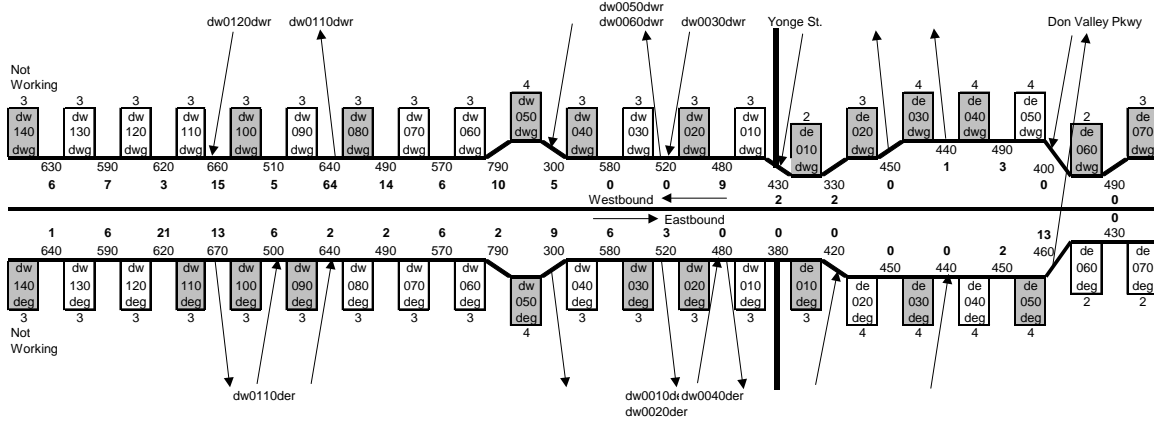
**FIGURE 1 Profile of queue formation precursor (*Q*) in typical daily traffic.**

Note:
1. The arrows pointing outward indicate off-ramps and the arrows pointing inward indicate on-ramps.
2. The letters inside the squares denote the station ID.
3. The numbers shown above or below the station ID are the number of lanes.
4. The numbers shown between two successive detectors are the distance in meter.
5. Shaded detector stations are the stations where traffic is influenced by merging or diverging vehicles.
6. The bold numbers shown above or below the distance are total number of crashes in 13 months.

FIGURE 2 Location of loop detector stations on Gardiner Expressway.
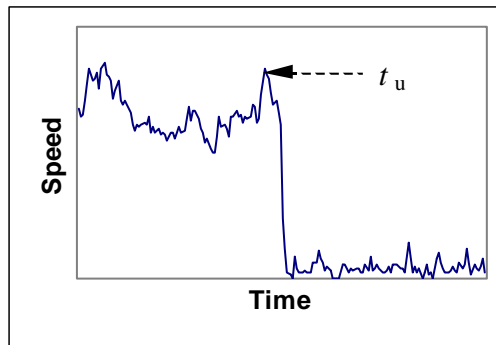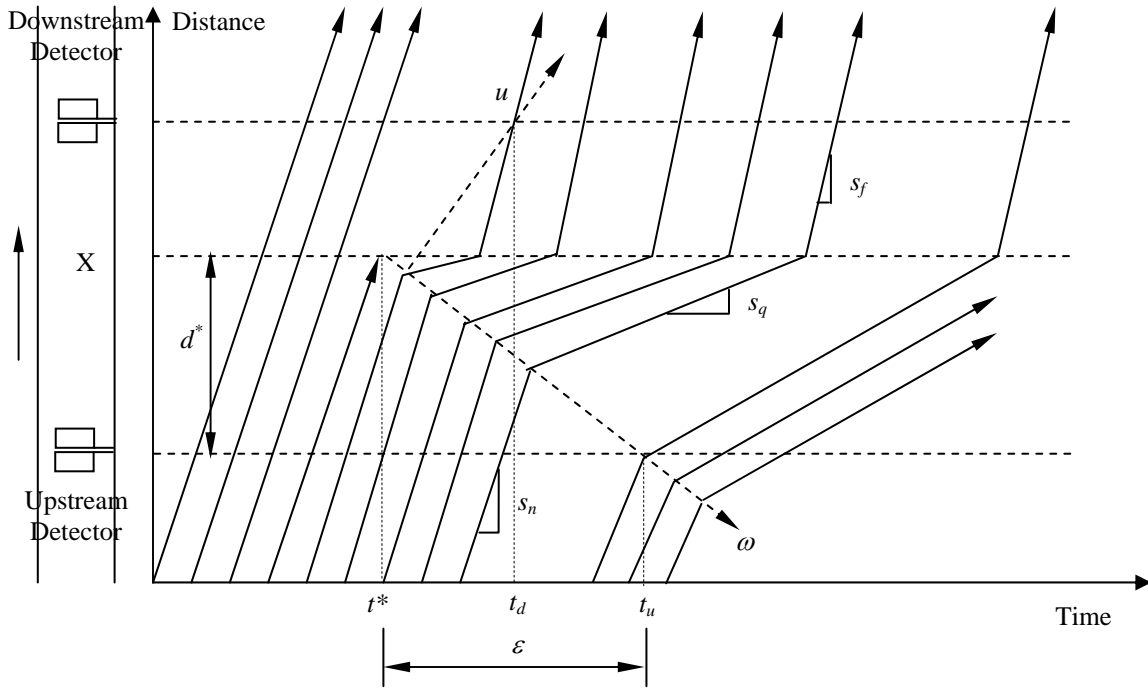
a) Speed change at upstream detector
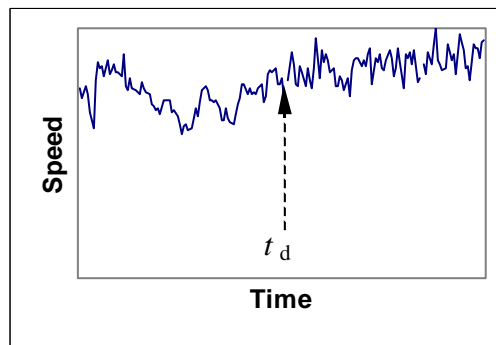
b) Speed change at downstream detector

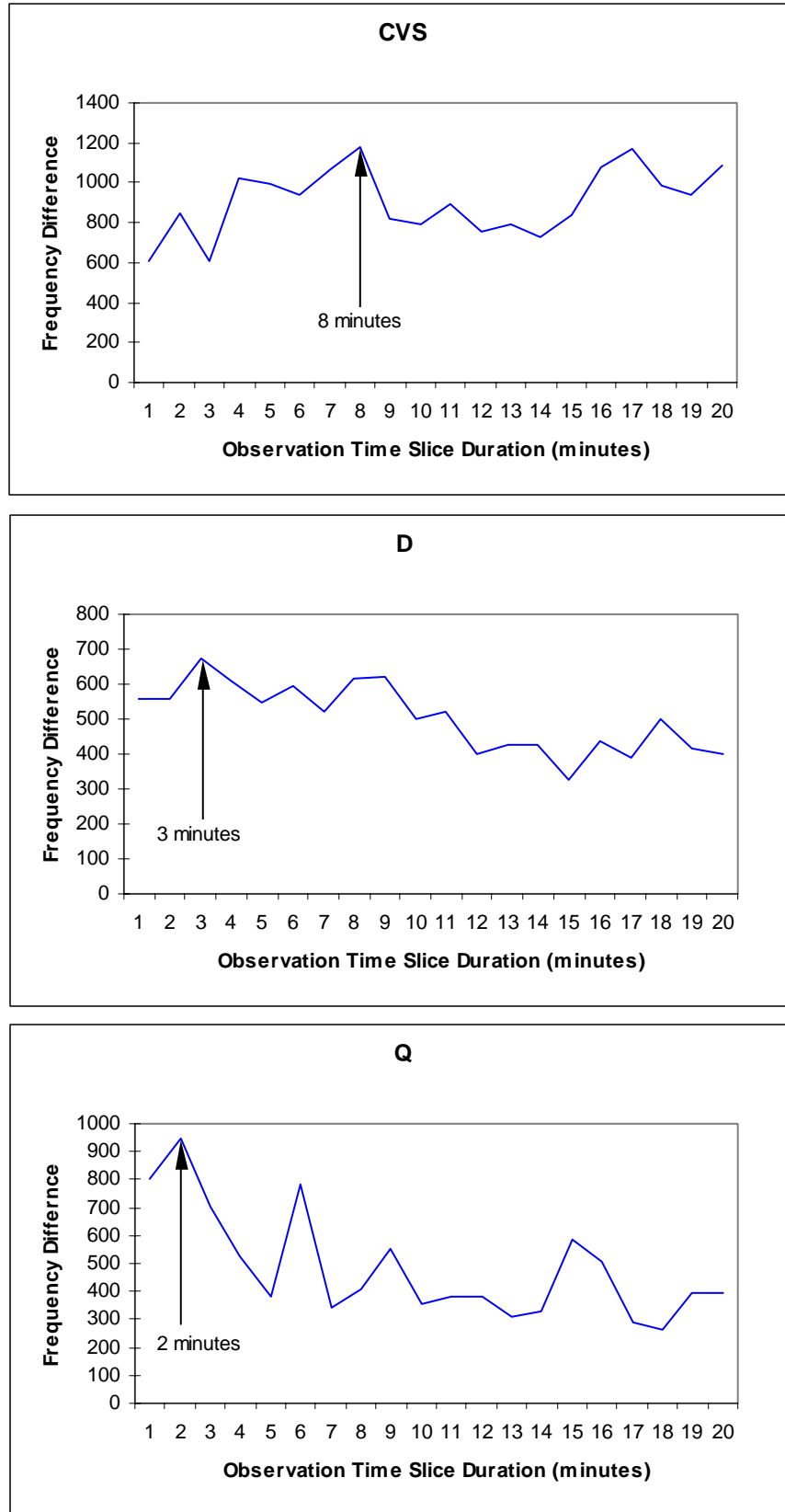**FIGURE 3 Illustration of speed changes after crash occurrence.**

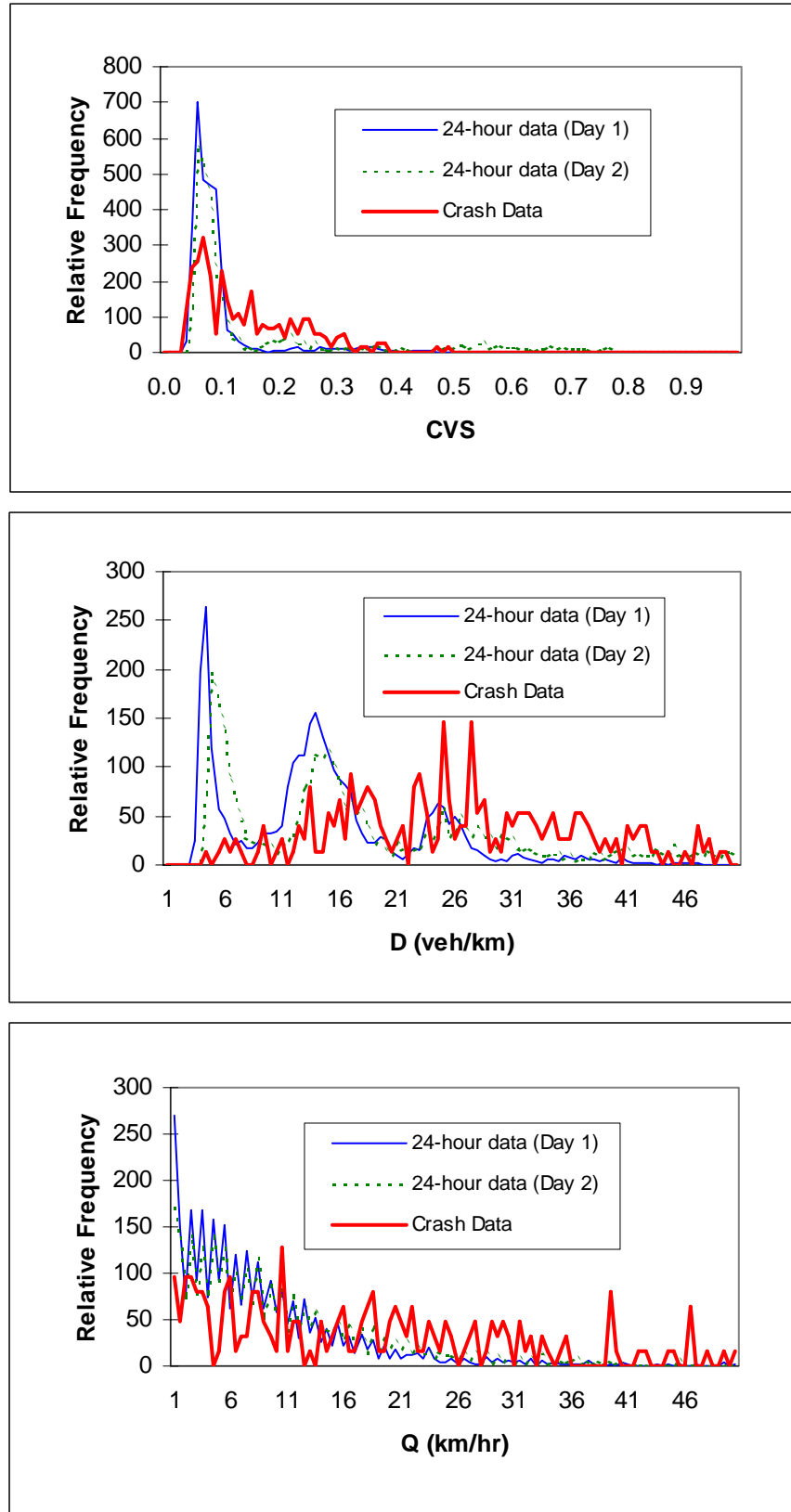**FIGURE 4 Determination of observation time slice duration.**

FIGURE 5 Distribution of crash precursors in two 24-hour data and crash data.