Identifying crash propensity using specific traffic speed conditions

Mohamed Abdel-Aty , Anurag Pande

Abstract

Introduction: In spite of recent advances in traffic surveillance technology and ever growing concern over traffic safety, there have been very few research efforts establishing links between real time traffic flow parameters and crash occurrence. This study aims at identifying patterns in the freeway loop detector data that potentially precede traffic crashes. *Method*: The proposed solution essentially involves classification of traffic speed patterns emerging from the loop detector data. Historical crash and loop detector data from the Interstate 4 corridor in the Orlando metropolitan area were used for this study. Traffic speed data from sensors embedded in the pavement (i.e., loop detector stations) to measure characteristics of the traffic flow were collected for both crash and non crash conditions. Bayesian classifier based methodology, probabilistic neural network (PNN), was then used to classify these data as belonging to either crashes or non crashes. PNN is a neural network implementation of well known Bayesian Parzen classifier. With its superb mathematical credentials, the PNN trains much faster than multilayer feed forward networks. The inputs to final classification model, selected from various candidate models, were logarithms of the coefficient of variation in speed obtained from three stations, namely, station of the crash (i.e., station nearest to the crash location) and two stations immediately preceding it in the upstream direction (measured in 5 minute time slices of 10 15 minutes prior to the crash time). *Results*: The results showed that at least 70% of the crashes on the evaluation dataset could be identified using the classifiers developed in this paper.

1. Introduction

The conventional approach to traffic safety analysis has been to establish relationships between the traffic characteristics (e.g., flow, speed), roadway and environmental conditions (e.g., geometry of the freeway, weather conditions), driver characteristics (e.g., gender, age), and crash occurrence. The shortcoming of most of the models developed using this approach is that they rely upon aggregate measures of traffic speed (e.g., speed limit) and volume (e.g., AADT or hourly volumes) and hence are not sufficient to identify the real-time “black spots” (i.e., locations having a high probability of crashes), created due to the interaction of ambient traffic conditions with the geometric characteristics of freeway segments.

In this study, the problem of predicting crashes (i.e., identifying freeway locations with high real-time crash potential) has been approached as a classification problem in which the real-time traffic conditions are categorized as measured by underground sensors (i.e., loop detectors) into either leading or not leading to a crash.

The idea of applying loop data to predict crashes in real-time is still in preliminary stages. However, there have been some efforts in this area. Lee, Saccomanno, and Hellinga (2002) introduced the concept of “crash precursors” and hypothesized that the likelihood of a crash is significantly affected by short-term turbulence of traffic flow. They came up with factors like speed variation along the length of the roadway (i.e., difference between the speeds upstream and downstream of the crash location) and also across the three lanes at the crash location. Another important factor identified by them was traffic density at the instant of the crash. Weather, roadway geometry, and the time of the day were used as external controls. With these variables, a crash prediction model was developed using log-linear analysis. In a later study Lee, Saccomanno, and Hellinga (2003) continued their work along the same lines and modified the aforementioned model. They incorporated an algorithm to get a better estimate of the time of the crash and the length of time slice (prior to the crash) to be examined. It was found that the average variation of speed difference across adjacent lanes doesn't have direct impact on crashes and hence was eliminated from the model. They also concluded that variation in speed has a relatively longer-term effect on crash potential than do either traffic density or average speed difference between upstream and downstream ends of roadway sections.

A study by Oh, Oh, Ritchie, and Chang (2001) also showed that the standard deviation of speed in a 5-minute interval was the best indicator of "disruptive" traffic flow leading to a crash as opposed to "normal" traffic flow. They used the Bayesian classifier to categorize the two possible traffic flow conditions. Since Bayesian classifier requires probability distribution function for each class, they fitted their crash and non-crash speed standard deviation data to non-parametric distribution functions using kernel smoothing techniques. Due to lack of crash data (only 52 crashes), their model remains far from being implemented in the field. It is also important to note that in order for a crash prediction model to be useful in preventing crashes, it is necessary to identify the crash prone conditions far ahead of the crash occurrence time, not just 5-minutes prior; more lead time allows traffic management authorities sufficient time for analysis, prediction, and dissemination of information.

Although these studies do indicate the potential of applying real-time loop detector data for identification of "alarming" traffic patterns on freeways, the biggest shortcoming of their analysis is that the data used in these studies were coming from just one station downstream and/or upstream of the crash location. Alarming conditions leading to crashes on a freeway might actually originate far upstream and "travel" with traffic platoons until they culminate into a crash at a certain downstream location. To account for this possibility, we examined data from several stations upstream of the crash location at several time periods leading to the crash. This will also serve the purpose of identifying how far in advance (in terms of both time and distance) of a crash occurrence certain freeway segments may be flagged for the impending hazard.

1.1. Introduction to the study area

This study was conducted on the Interstate-4 (I-4) corridor in Orlando. The freeway stretch under consideration is about 13.25 miles long and has a total of 28 loop detector stations, spaced out at approximately 1/2 mile intervals. Through dual loop detectors (sensors located beneath the pavement) these stations collect and store the following measurements every 1/2 minute for three lanes in each direction:

a) Volume (number of vehicles passing each lane in 30 seconds)
b) Lane-occupancy (percentage of the 30-second interval the loop detector was occupied), and
c) Average speed (of all vehicles passing over the loop detector in the 30-second interval).

Also, this freeway stretch is under the jurisdiction of the Orlando police department (OPD) and hence OPD was the source of the crash data for this study.

First, the mile-post location was identified for each of the 670 crashes that occurred on the Interstate-4 corridor during the period of April 1999 through November 1999. The remaining months of the year 1999 had to be excluded, as no loop data were available for those months. For every crash, its time and location were identified using the new police reporting system (the time the police department receives a call reporting a crash, which is entered and checked with the drivers involved and witnesses; this time is usually very close to the actual time of crash occurrence), and verified based on inspecting the space- and time-volume, occupancy, and speed diagrams. For every crash, the loop detector station nearest to its location was determined. This station is referred to as the station of the crash from here on. The next step was to extract pre-crash loop detector data from the archived loop detector database. As mentioned earlier our focus is on comparison and classification of crash and non-crash traffic flow variables, therefore if a crash is reported to occur on April 12, 1999 (Monday) 6 p.m., I-4 Eastbound, and the nearest loop detector was at station 30, data were extracted from station 30, five loops upstream and one loop downstream of station 30 for half an hour period prior to the reported time of the crash for all the Mondays of the eight month period of analysis at the same time. This matched sample design was created in order to control for roadway and geometric factors and driver population on the freeway (e.g., more commuters on weekday peak hours, indicating more young to middle age drivers, etc.). Hence, this crash will have a loop data table consisting of the speed, volume, and occupancy values for all three lanes from the loop stations 25–31 (on eastbound direction) from 5:30 p.m. to 6 p.m. for all the Mondays of the aforementioned eight-month period of the year 1999, with one of them being the day of crash. The data were available for only 377 (out of 670) crashes. During the time of the remaining crashes none of the loops from which data were required were functioning.

The loop detectors suffer from intermittent hardware problems that result in unreasonable values of speed, volume, and occupancy. These values include Occupancy

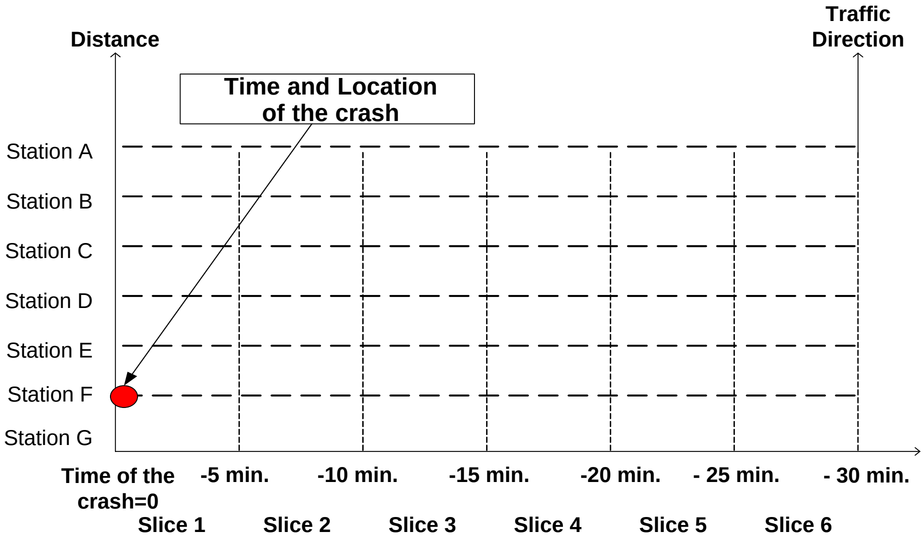

Fig. 1. The nomenclature for defining the station and time slice to which any "effect" (average or standard deviation) belongs.

>100 (percent), speed = 0 or >100 (MPH), flow >25 (vehicles per lane), and flow = 0 (vehicles per lane) with speed >0 (MPH), and were removed from raw 30-second data. From the "cleaned" data tables the average and standard deviation of speed were extracted over each lane for six, 5-minute intervals recorded prior to the crash on the station nearest to the crash location (referred to as station of the crash), five stations upstream and one station downstream of the station of the crash. It requires creation of 252 fields (7 stations*6 time slices*3 lanes*2 variables, i.e., average and standard deviation of speed) in the database for each crash. The same 252 fields were extracted for all "corresponding" non-crash days as well.

The nomenclature procedure adopted for defining the station and time slice to which the average and standard deviation belongs is shown in Fig. 1. All the stations were named as "A" to "G," with "A" being the farthest station upstream and so on. It should be noted that "F" is the station of the crash and "G" will be the station downstream of the crash location since we have collected data from 5 upstream stations, station of the crash itself, and one downstream station. Similarly the 5-minute intervals were also given "ID" from 1 to 6. The interval between time of the crash and 5 minutes prior to the crash was named as slice 1, interval between 5 to 10 minutes prior to the crash as slice 2, and interval between 10 to 15 minutes prior to the crash as slice 3, and so on.

2. Methodology: Theoretical background of the classification technique

The proposed solution to the research problem essentially involves classification of traffic speed patterns emerging from the loop detectors. This section provides a theoretical overview of the probabilistic neural network (PNN) classifiers used here. The PNN is a neural network implementation of the well-established multivariate Bayesian classifier, using Parzen estimators to construct the probability density functions of different classes (Specht, 1996).

2.1. Bayes classification

The PNN is strongly based on Bayesian method, which is arguably the single most popular classification paradigm. Suppose there is a collection of random samples from K populations *(*$k = 1, 2, \ldots\ldots, K$; *e.g., for crash vs. non crash* $K = 2$*)* and each of these samples is a vector $x = [x_1, x_2, \ldots\ldots\ldots x_m]$, then these samples may be used to devise a Bayes optimal decision rule in order to classify a pattern of unknown class. Essentially this rule favors a class (e.g., crash vs. non-crash) if it has high density in the vicinity of the pattern of unknown class. The probability density function $f_k(\boldsymbol{x})$ corresponds to the concentration of class k cases around the pattern of unknown class. The problem with this rule is that the probability density functions (PDFs) are generally unknown and they should be estimated from the random samples available from K populations (Masters, 1995).

2.2. Parzen Estimator

Parzen estimator uses the weight function $W(d)$ (frequently referred to as potential function or a kernel) having largest value at "distance $d = 0$" and it decreases rapidly as the absolute value of "d" increases (Masters, 1995). The weight functions are centered at each training sample point with the value of each sample's function at a given abscissa being determined by the distance "d" between x and that sample point. The PDF estimator is the scaled sum of that function for all the sample cases.

The method can be stated mathematically using the following equation:

$$g(x) = \frac{1}{n\sigma} \sum_{i=1}^{n} W\left(\frac{x - x_i}{\sigma}\right).$$

The scaling parameter σ (also known as spread value) defines the width of the bell curve that surrounds each sample point. The value of this parameter has a profound influence on the performance of a PNN. While too small values will cause individual training cases to have too much of an influence, thereby losing the benefit of aggregate information, extremely large values will cause so much blurring that the details of density will be lost, often distorting the density estimate badly (Masters, 1995). This idea will be clearer when the results from various PNN models are discussed later in the paper. Note that the above description for Parzen estimator corresponds to a univariate case. Oh et al. (2001) used a similar procedure to obtain the density function in their research work. An extension to multivariate setting is intuitive, the details of which may be found in Abdulhai and Ritchie (1999).

2.3. Multivariate bayesian discrimination and classical PNN

The accuracy of decision boundaries' estimation and the subsequent classification depends on the accuracy with which the underlying PDFs are estimated. A nice feature of this approach and the related PNN implementation is the estimation consistency. Consistency refers to the fact that the error in estimating the PDF from a limited sample gets smaller as the sample size increases, and therefore the estimated PDF (the class estimator) collapses on the unknown true PDF as more patterns in the sample become available.

Also, note that the estimated PDF for a given class, say $f_k(x)$, is the sum of small multivariate Gaussian distributions centered at each training sample. However, the sum is not necessarily Gaussian. It can, in fact, approximate any smooth density function. The smoothing factor σ can alter the resulting PDF. The optimal σ can be easily determined experimentally (Abdulhai & Ritchie, 1999).

The network in Fig. 2 shows p dimensional inputs to be classified into two classes. The pattern layer contains one neuron for each training case while the summation layer has one neuron for each class. In the creation (training) phase of the PNN each training case (patterns with known classification) is stored in a neuron of the pattern layer. To classify an unknown input pattern, the execution starts by simultaneously presenting this input vector to all pattern layer neurons. Each pattern neuron then computes a distance measure (Euclidean in the case of a classical PNN) between the input and the training case represented by that neuron. It then subjects the distance measure to neuron's potential function (*W(d)*). The following layer contains summation units that have a modest task. Each summation layer neuron is dedicated to a single class. It just sums up the pattern layer neurons corresponding to the members of that summation neuron's class. The attained activation of the summation neuron is the estimated density function value for that population class. The output neuron is merely a threshold discriminator and decides which of its inputs from the summation units is the maximum (Masters, 1995).

In other words, the PNN computes the potential function for the distances between unknown input pattern and the stored training patterns from two competing classes (i.e., crash vs. non-crash). Whichever class has higher potential function (i.e., has more density around the unknown input pattern) is chosen to be the class of the unknown vector.

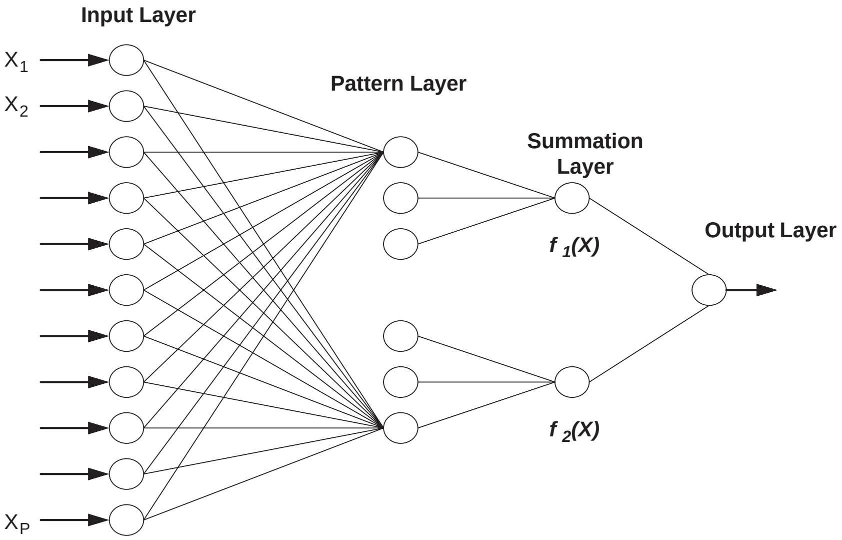

Fig. 2. The traditional PNN architecture for a two-class classification problem. Statistical distance and the modified PNN.

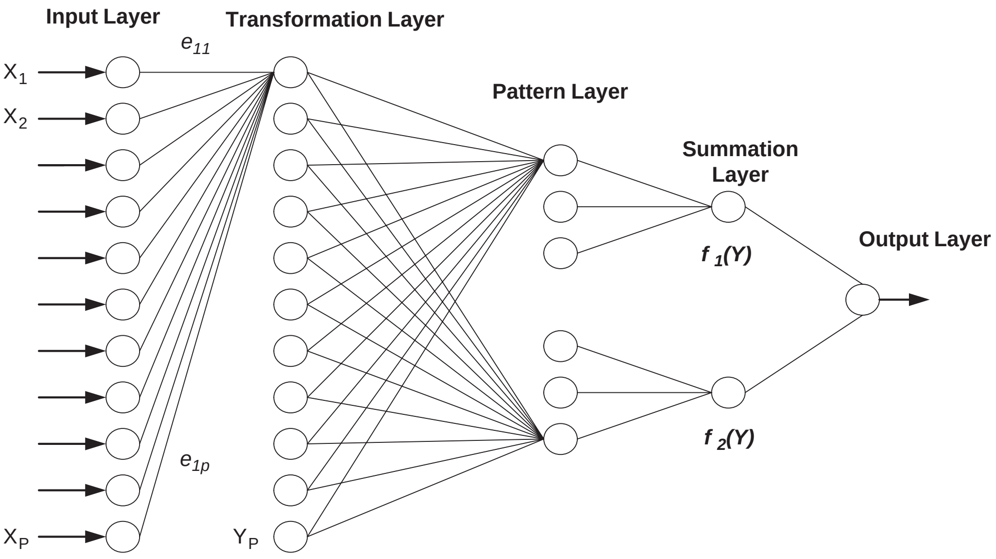

Fig. 3. The modified PNN for a two-class classification problem (Abdulhai and Ritchie, 1999).

The classical PNN uses Euclidean distance as a measure of nearness among different patterns. Euclidean distance is statistically unsatisfactory for some applications because it does not account for differences in variations along the axes (i.e., if some parameters in the input vector are more important than the others) nor the presence of correlation among the variables constituting the pattern vector. To overcome this deficiency, Abdulhai and Ritchie (1999) proposed a modification in the classical PNN algorithm.

To replace the employed Euclidean distance with the preferred statistical distance, principal components rather than the original variables should be used. Algebraically, principal components are particular linear combinations of the original set of random variables.

Fig. 3 shows the modified version of the PNN (referred to as PNN2) that takes the above transformation into account. Two layers, an input layer and a transformation layer, replace the previous input layer of the classical PNN. The original input vector X is transformed into a rotated vector Y using the eigenvectors (e_{ij}) of the covariance matrix Σ associated with random vector X. The component variables of the vector in terms of the rotated axes are then divided by their standard deviations $(\lambda_i)^{0.5}$ to equalize the variances and obtain a new set of inputs free of the effects of correlation and widely varying variances. Beyond this transformation, layer processing of PNN2 is identical to the original PNN described earlier (Abdulhai & Ritchie, 1999).

2.4. Exploration with the loop detector data

There are several studies associating crash occurrences with increasing variation in vehicle speeds (e.g., Shinar, 1999; Garber & Ehrhart, 2000). It has been argued that as individual vehicle speeds deviate more and more from the average speed of the traffic stream, the probability of having a crash increases. The data emanating from a series of consecutive loop detectors on a freeway section has been used here as a surrogate for the detailed vehicle movement data in order to capture the variance in vehicle speeds.

Table 1
Average values of 5-minute standard deviation of speeds observed at various time slice-station combinations

	Time Slice											
	1		2		3		4		5		6	
	Y		Y		Y		Y		Y		Y	
	0	1	0	1	0	1	0	1	0	1	0	1
Station												
A	5.38	5.63	5.39	5.39	5.36	5.46	5.36	5.61	5.31	5.41	5.30	5.21
B	5.33	5.67	5.37	5.69	5.29	5.65	5.31	5.59	5.34	5.60	5.31	5.51
C	5.38	5.58	5.36	5.71	5.36	5.71	5.34	5.73	5.34	5.44	5.33	5.42
D	5.23	6.00	5.27	5.70	5.26	5.69	5.26	6.05	5.25	5.59	5.24	5.47
E	5.27	6.00	5.30	5.55	5.22	5.63	5.24	5.51	5.26	5.86	5.23	5.63
F	5.33	5.89	5.33	5.79	5.34	5.89	5.33	5.89	5.30	5.85	5.27	5.42
G	5.14	5.50	5.20	5.67	5.15	5.26	5.20	5.60	5.17	5.55	5.17	5.56

Table 2
Average values of 5-minute average speeds observed at various time slice-station combinations

	Time Slice											
	1		2		3		4		5		6	
	Y		Y		Y		Y		Y		Y	
	0	1	0	1	0	1	0	1	0	1	0	1
Station												
A	49.24	46.81	49.15	46.30	49.11	45.82	49.11	46.63	49.13	47.01	49.24	46.72
B	46.96	43.80	46.95	43.55	47.06	43.57	47.06	43.94	47.08	44.20	47.15	44.65
C	46.62	42.59	46.62	42.76	46.79	42.34	46.86	42.57	46.97	42.94	47.12	42.86
D	47.23	41.56	47.20	42.27	47.41	42.73	47.66	42.81	47.78	43.57	47.89	43.43
E	46.23	40.18	46.27	41.00	46.39	41.47	46.50	41.30	46.62	42.20	46.79	42.80
F	45.71	40.08	45.71	39.93	45.92	39.88	46.01	39.38	46.21	40.18	46.38	41.02
G	48.09	42.60	48.10	42.43	48.21	41.69	48.38	41.67	48.49	41.61	48.66	42.89

Due to malfunctioning of loops, the speed, volume, and occupancy data were rarely available simultaneously over the three lanes. Moreover, there were a lot of data missing from all three lanes (such data was not used in the analysis). To overcome the problems due to missing data, it was decided to replace the values on three lanes with one value that was the average over the three lanes. Averaging was preferred over imputation of missing values because imputation procedures would have been very time consuming and beyond the scope of this study. The averaging over three lanes is acceptable for the analysis carried out in this paper, because about 76% of crashes in the database were rear-end. Therefore, the longitudinal variation of traffic parameters was deemed to be more critical than the variation across lanes.

To detect the trends in 5-minute averages and standard deviations of speed at various time slices and stations, their averages over all the crash and non-crash cases were obtained. Table 1 provides the average of 5-minute standard deviation of speeds over all the crash cases (Columns with $Y = 1$) and non-crash cases (Columns with $Y = 0$). Similarly, Table 2 provides the average of 5-minute averages of speeds. The values are obtained at all 42 (7 stations * 6 slices) time-slice and station combinations. It should be noted that the average over non-crash cases is observed over much more data points than the crash cases.

Observing Table 1 closely shows that the crash case variance ($Y = 1$) is higher than the non-crash ($Y = 0$) counterpart at all the stations during every time slice except for station A (that is 5 stations upstream of the station of the crash) during time slice 6 and time slice 2 (where they are equal). Another interesting aspect is that as we “approach” the time and location of the crash the difference in standard deviation tends to increase. Also, the differences during all time slices at station A are relatively smaller than the other entries in the Table.

These two parameters (5-minute average and standard deviation) may be chosen as crash precursors as they represent turbulent traffic conditions ahead of the crash occurrence. Lower average speed signifies congestion and queuing conditions on freeways while high variability associated with it depicts frequent formation and dissipation of such queues. In such a scenario, the drivers on the freeway might have to slow down and speed up quite often while traversing through small distances. These conditions can potentially lead to rear-end crashes.

In view of the above argument, 5-minute coefficient of variation (standard deviation/average) in speed may be used to account for the trends observed in Tables 1 and 2. Table 3

Table 3
Aggregated 5-minute coefficient of variation in speed expressed in term of percentages at various time slice-station combinations

	Time Slice											
	1		2		3		4		5		6	
	Y		Y		Y		Y		Y		Y	
	0	1	0	1	0	1	0	1	0	1	0	1
Station												
A	10.93	12.03	10.97	11.64	10.91	11.92	10.91	12.03	10.81	11.51	10.76	11.15
B	11.35	12.95	11.44	13.07	11.24	12.97	11.28	12.72	11.34	12.67	11.26	12.34
C	11.54	13.10	11.50	13.35	11.46	13.49	11.40	13.46	11.37	12.67	11.31	12.65
D	11.07	14.44	11.17	13.48	11.09	13.32	11.04	14.13	10.99	12.83	10.94	12.59
E	11.40	14.93	11.45	13.54	11.25	13.58	11.27	13.34	11.28	13.89	11.18	13.15
F	11.66	14.70	11.66	14.50	11.63	14.77	11.58	14.96	11.47	14.56	11.36	13.21
G	10.69	12.91	10.81	13.36	10.68	12.62	10.75	13.44	10.66	13.34	10.62	12.96

consists of the values for the coefficient of variation in speed expressed in terms of percentage for every time slice-station combination. This variable also magnifies the difference between crash and non-crash cases, which would help the distance based classifiers to correctly identify certain patterns.

2.5. Preliminary matched case control logistic regression

A basic matched case-control analysis, where the crashes act as cases and all corresponding non-crash data are used as controls, was performed. In this analysis the value of "Hazard ratio" (ratio of odds for crash occurrence versus not, i.e., odds ratio) for the data combined over three lanes was derived.

In a logistic regression setting the function of dependent variables yielding a linear function of the independent variables would be the logit transformation.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_{1,x},$$

where $\pi(x) = E\ (Y|x)$ is the conditional mean of Y (dummy variable representing crash occurrence) given x when the logistic distribution is used. Under the assumption that the logit is linear in the continuous covariate x the equation for the logit would be $g(x) = \beta_0 + \beta_{1,x}$. It follows that the slope coefficient β_1, gives the change in the log odds for an increase of 1 unit in x, that is $\beta_1 = g\ (x + 1) - g\ (x)$ for any

Table 4
Hazard ratio for average volume, average occupancy, and log coefficient of variation of speed

Variable		5-minute average of Volume		5-minute average of Occupancy		Log (5-minute coefficient of variation in speed)	
Station	Time slice	Pr>Chisq	Hazard ratio	Pr>Chisq	Hazard ratio	Pr>Chisq	Hazard ratio
A	1	0.197	0.967	0.008	1.018	0.007	1.643
A	2	0.282	0.969	0.008	1.022	0.024	1.499
A	3	0.366	0.971	0.005	1.021	0.015	1.642
A	4	0.694	0.992	0.039	1.019	0.017	1.660
A	5	0.947	1.004	0.083	1.015	0.069	1.409
A	6	0.582	1.011	0.034	1.018	0.186	1.289
B	1	0.138	0.962	0.031	1.016	0.003	1.631
B	2	0.060	0.945	0.012	1.019	0.002	1.680
B	3	0.080	0.939	0.029	1.013	0.001	1.764
B	4	0.231	0.972	0.044	1.012	0.004	1.593
B	5	0.200	0.969	0.116	1.011	0.007	1.502
B	6	0.960	1.002	0.130	1.010	0.019	1.446
C	1	0.192	0.971	0.005	1.013	0.000	1.722
C	2	0.640	0.991	0.036	1.012	0.000	1.805
C	3	0.180	0.964	0.005	1.018	0.000	1.852
C	4	0.546	0.985	0.011	1.012	0.000	1.807
C	5	0.925	0.991	0.003	1.025	0.001	1.616
C	6	0.754	1.003	0.017	1.016	0.001	1.706
D	1	0.174	0.965	<.0001	1.023	<.0001	2.789
D	2	0.210	0.969	0.001	1.024	0.000	2.212
D	3	0.329	0.974	0.001	1.027	0.001	2.054
D	4	0.443	0.988	0.000	1.027	<.0001	2.607
D	5	0.712	1.006	0.001	1.023	0.002	1.770
D	6	0.861	1.002	0.001	1.021	0.002	1.771
E	1	0.550	0.985	<.0001	1.035	<.0001	2.981
E	2	0.535	0.983	<.0001	1.032	<.0001	2.256
E	3	0.902	0.998	<.0001	1.031	<.0001	2.284
E	4	0.344	0.975	<.0001	1.031	<.0001	2.320
E	5	0.947	1.001	<.0001	1.033	<.0001	2.541
E	6	0.953	1.002	<.0001	1.031	<.0001	2.100
F	1	0.016	0.945	<.0001	1.037	<.0001	2.335
F	2	0.013	0.981	<.0001	1.034	<.0001	2.568
F	3	0.031	0.967	<.0001	1.031	<.0001	2.603
F	4	0.031	0.982	<.0001	1.029	<.0001	2.857
F	5	0.110	0.998	<.0001	1.030	<.0001	2.630
F	6	0.347	0.969	<.0001	1.026	0.000	2.022
G	1	0.201	0.984	<.0001	1.032	<.0001	2.272
G	2	0.152	0.987	<.0001	1.029	<.0001	2.571
G	3	0.170	0.972	<.0001	1.035	<.0001	2.221
G	4	0.140	0.963	<.0001	1.033	<.0001	2.486
G	5	0.113	0.962	<.0001	1.031	<.0001	2.704
G	6	0.562	0.989	<.0001	1.031	<.0001	2.411

value of x. Hazard ratio is defined as e raised to the power of this coefficient (Agresti, 2002).

Table 4 shows the values of hazard ratio for average volume, average occupancy and log of coefficient of variation in speed (logcvs) at all time slice-station combinations, when used one at a time as the risk factor (i.e., independent variable) in the matched case-control logistic regression analysis. It could be seen that the values of hazard ratio are much less for volume and occupancy when compared to those of logcvs. Fig. 4 depicts the trends shown by the values of "hazard ratio" when logcvs are used one at a time as independent variable. Note that the crashes are treated as cases while all available corresponding non-crash cases act as controls.

The "hazard ratio" essentially represents the factor by which the risk of observing a crash in the vicinity of "station of the crash" will increase when the corresponding "risk factor" (i.e., the covariate used as independent variable) is increased by one unit. This means that the time slice-station combination with a higher value of "hazard ratio" will affect the probability of crash occurrence to a greater degree. It may be seen that the values observed for stations "D" "E" "F" and "G" are higher than those observed for stations "A" "B" and "C" during all the time slices. The higher value of the hazard ratio is an important consideration when selecting which of the traffic parameters will become inputs to the PNN models.

2.6. Development of classification models

The variable (Logcv) with highest hazard ratios was chosen as input to the PNN models, but this was not the only consideration. If the Logcvs during time slice 1 and 2 (i.e., 0–5 and 5–10 minutes prior to the crash), despite having maximum hazard ratios, were to become inputs to the model the prediction will come out too late to predict a crash and warn the drivers about it, once the model is applied on-line. It was therefore decided to work with variables that are observed at least 10–15 minutes prior to the crash. Also, all the Logcvs to be fed into a model for training and testing should belong either to the same time slice duration or to the same station. This was required from a field application standpoint because if a model uses data from different detectors at different time slices and classifies a real-time pattern as "alarming," it would be difficult to determine exactly which section should be flagged as a potential crash location.

The curves on the segment of Interstate-4 under consideration are not of widely varying radii, therefore the roadway alignments along the corridor were divided into two categories (i.e., straight and curved). To incorporate this into the PNN architecture, the population classes were increased to four (i.e., crash on curved section, crash on straight section, non-crash on curved section, and non-crash on straight section), instead of just two (i.e., crash vs. non-crash).

2.7. Preparation of training and evaluation datasets

As described in the previous section, loop detector data were obtained for 377 crashes. The data were then used to calculate Logcvs for various (42 in all) time slice-station combinations. To classify these data through a PNN classifier, all the Logcvs to be fed in the model should be simultaneously available. Based on this consideration, due to poor availability of data we were left with 148 (out of 377) crash and 2,857 non-crash data points. From both categories (crash and non-crash) approximately two-third (66%) of the data points were used for creation of the networks with the remaining one-third used for evaluation. The data belonging to crash category were heavily underrepresented, hence it was necessary to balance the dataset in order to have equal crash and non-crash data points used for the creation of the networks.

First, 100 crash data points (66% of the total 148) were randomly selected from the available crashes. Subtractive clustering procedure was then used in order to reduce 1,883 non-crash data points (66% of the total 2,857; to be used for creation of PNN) into 100 cluster centers. The procedure essentially involved identifying an appropriate cluster radius such that 100 points (out of 1,883) are selected as cluster centers representing all the points lying within that particular radius. With randomly selected 100 crash data

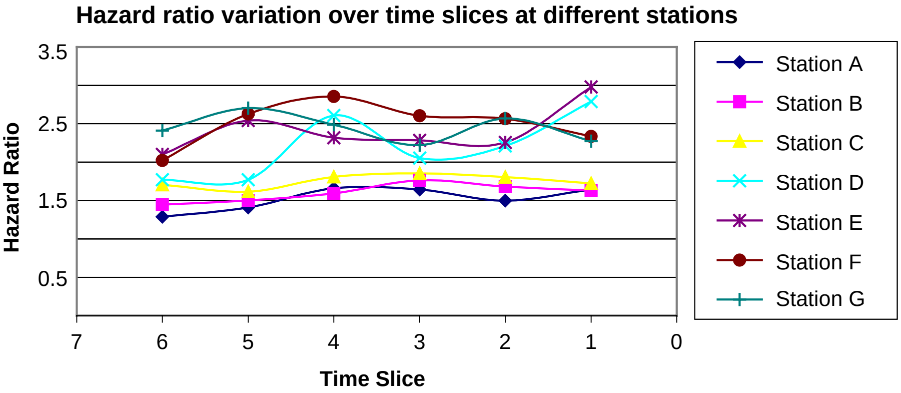

Fig. 4. Hazard Ratio of log coefficient of variation of speed over time slices observed at different stations.

Table 5
The number of patterns in the datasets created for training and evaluation of neural networks

Data set	Crash data points	Non-crash data points	Number of training data points		Number of evaluation data points	
			Crash	Non-crash	Crash	Non-crash
Complete	148	2857	100	100(1883)	48	974
Time-limited	116	2289	78	78(1526)	38	763

points and 100 non-crash cluster centers the dataset for creation of PNNs was ready. It should be noted, however, that non-crash data points in the evaluation dataset were not clustered and were used as is. The number of points in creation and evaluation datasets are shown in Table 5 (refer to the first row). It may be seen that the test (evaluation) dataset had a total of 48 crashes (148–100 = 48) and 974 non-crash data points (2,857–1,883 = 974).

Because crashes during late-night and early morning hours may be attributed mostly to human errors rather than ambient traffic conditions (e.g., a study by Stutts et al. (2003) in which the fatigue/drowsiness has been associated with late-night crash involvement), a reduced dataset (referred to as "time-limited") was prepared in which only the crashes (and corresponding non-crash data points) that occurred from 7 a.m. to 10 p.m. were included. The number of data points available for training and testing was obviously reduced in the time-limited dataset. The composition of this dataset is also shown in Table 5 (refer to second row). The figure in parenthesis in the column containing non-crash training data points is the number of patterns from which the cluster centers, equal to the number of crash data points, are obtained.

3. Classification models: results and discussion

The hazard ratio values for Logcvs were higher than corresponding values for average volume and occupancy, therefore the first experiment with PNN involved deciding on the combination of Logcvs to be used as inputs. Based on the hazard ratio values for the variables and the practical consideration described earlier, various combinations of Logcvs were used as PNN inputs and the resulting performance of the models on the evaluation dataset was carefully examined. The classification performance of the models was evaluated in terms of two parameters, namely, percentage of overall (crash and non-crash) patterns classified correctly on the test dataset and percentage of crash identification over the test dataset. The criterion for the optimal model was the maximum overall classification accuracy for at least 70% of crashes identified correctly. The overall classification accuracy criterion will ensure that even at good crash identification rate, too many false warnings aren't issued.

It was observed that the three-dimensional input pattern involving the Logcvs at stations D, E and F (which are the two stations upstream and the station of the crash itself, respectively) during time slice 3 (10–15 minutes prior to the time of the crash) meets the requirement of providing the optimal crash identification of 72.5% (with 62.1% of overall crash and non-crash identification accuracy) on the evaluation dataset. In order to further explore the parameters that might improve the classification, in addition to the three dimensional traffic speed pattern, 5-minute average occupancies from various time slice-station combinations were included as inputs. It was found that when occupancy at G3 (Station G, downstream of crash location at time slice 3) with LogcvD3, LogcvE3, LogcvF3 is used as part of a 4-dimensional input pattern, the maximum crash identification that could be achieved was 62.34%. It was the maximum among various 4-dimensional patterns explored with occupancy data from time slices 3 to 6. Note that it is less than that achieved through traffic speed patterns only and doesn't even satisfy the minimum requirement of 70%. It was observed that when occupancy at G3 was replaced with Occupancy at F2 or F1 (at station of the crash, during time period 5–10 and 0–5 minutes prior to crash, respectively) the classification accuracy improved marginally. This finding conforms to the literature (Lee et al., 2003) as the occupancy seems to have a relatively short-term effect on crash occurrence as compared to temporal variation in speed. It is worth mentioning that such models would have little practical application since there will not be enough time to "predict" a crash. Therefore, they were discarded from further considerations and the final models used only the three-dimensional input pattern with Logcvs at stations D, E and F to represent the real-time traffic characteristics.

Table 6 shows the results of the model using the aforementioned 3-dimensional input patterns and classifying them as crash or non-crash over a range of σ (the spread parameter) values. It may be recalled from an earlier section in the paper that the spread parameter has a profound impact on the estimated PDFs. The optimal performance based on the criteria adopted is highlighted in the table. It may be seen that at very small spread values (e.g., 0.005), the model has very high accuracy for crashes (above 95%) but the overall classification accuracy is poor (less than 20%). What this essentially means is that most of the data points from the test data set are being classified as crashes, which from a practical point of view would lead to excessive "false alarms." At near zero spread values the PNN act as a nearest neighbor classifier with class of "single nearest neighbor" exerting too much influence on the resulting class of test data point. It is highly likely to have non-crash data near to at least one of the crash data points (the reason being that

Table 6
Performance of PNN models classifying observations from the complete dataset as crash vs. non-crash

Spread Value	Result parameters for classical PNN (%)		Result parameters for modified PNN (%)	
	Overall classification accuracy (test crash and non-crash data)	Accuracy on test crash data	Overall classification accuracy (test crash and non-crash data)	Accuracy on test crash data
0.005	18.9	97.5	19.9	98.0
0.01	21.5	97.5	20.0	97.5
0.015	27.9	90.0	25.5	90.8
0.02	37.8	87.5	34.2	88.5
0.025	48.6	85.0	46.7	84.2
0.03	56.2	77.5	54.3	76.3
0.035	62.1	72.5	59.8	73.7
0.04	66.7	67.5	63.9	68.4
0.045	70.0	65.0	67.7	65.8
0.05	72.5	62.5	70.3	63.2

sometimes even the alarming conditions may not culminate into a crash due to driver's ability). Therefore, if for a non-crash case its "single nearest neighbor" lies in the crash category, then at near zero spread value it will be classified as crash even though multiple data points from the competing class (i.e., non-crash cases) might be present in the vicinity of this unknown input pattern. When the value of spread parameter was increased gradually (i.e., with an increment of 0.005), it was found that although the overall classification accuracy increases, the percentage of crashes correctly identified decreases. This means that at even higher spread values such a network will classify everything as non-crash and achieve high overall accuracy but would be of no use, as the primary aim of this research is to identify the crashes correctly. The reason for missing out on crashes is because so much blurring is caused by the high spread parameter value (σ) that it loses the details of density function of the crash data. Therefore, an appropriate spread value providing optimal classification based on the 70% crash identification criterion should be chosen. Note that the performance of PNN model at various spread parameter values conform to the properties discussed earlier in the paper while discussing the Parzen estimator.

Two more PNN models were created and evaluated, incorporating the roadway alignment (straight vs. curved) at the crash location and time of the day when crash occurred, respectively, in the classes to be identified. Inclusion of time of day into the classes to be identified degrades the performance of the network drastically. To explain this, one must first note that the 70% minimum crash identification criterion for this PNN model is achieved at spread parameter value of $\sigma = 0.015$, which is far less than that for any other models (see row 3 in Table 7). Inclusion of late night crashes would mean that a lot of daytime non-crash cases will be similar to (i.e., less distant neighbors of) night time crashes (in late night hours 5-minute average speeds will be high with less variance, a pattern which we expect day-time non-crashes to follow). Since a spread value as low as 0.015 will force PNN to become merely a nearest neighbor classifier, a lot of day time non-crashes will be classified as late night crashes (which are their nearest neighbors). This would mean poor overall classification accuracy at desired crash identification rate (greater than 70%). Also, as mentioned earlier, the late night crashes may be attributed more to human errors, rather than any crash prone interactions between vehicles resulting from congestion or turbulence in the traffic speed patterns, making them difficult to "predict." These factors result in significantly poor performance of this particular PNN model.

A careful analysis of the missed (i.e., unidentified) crashes led to the conclusion that quite a few of these crashes occurred during late night hours, so a model classifying the speed patterns into crash and non-crash

Table 7
The optimal classification performances by various PNN models

Dataset used for training and evaluation	Roadway alignment in the classes to be identified	Time of the day in the classes to be identified	Parameters for classical PNN			Parameters for modified PNN		
			Spread Value	Overall accuracy (test crash and non-crash data)	Accuracy on test crash data	Spread Value	Overall accuracy (test crash and non-crash data)	Accuracy on test crash data
Complete	×	×	0.035	62.1 %	72.5 %	0.035	59.8 %	73.7 %
Complete	√	×	0.045	74.6 %	71.7 %	0.045	73.2 %	71.6 %
Complete	×	√	0.015	17.8 %	72.3 %	0.035	18.8 %	72.0 %
Time-limited	×	×	0.050	80.0%	70.1 %	0.045	72.6 %	73.9 %

√ Incorporated; × Not Incorporated.

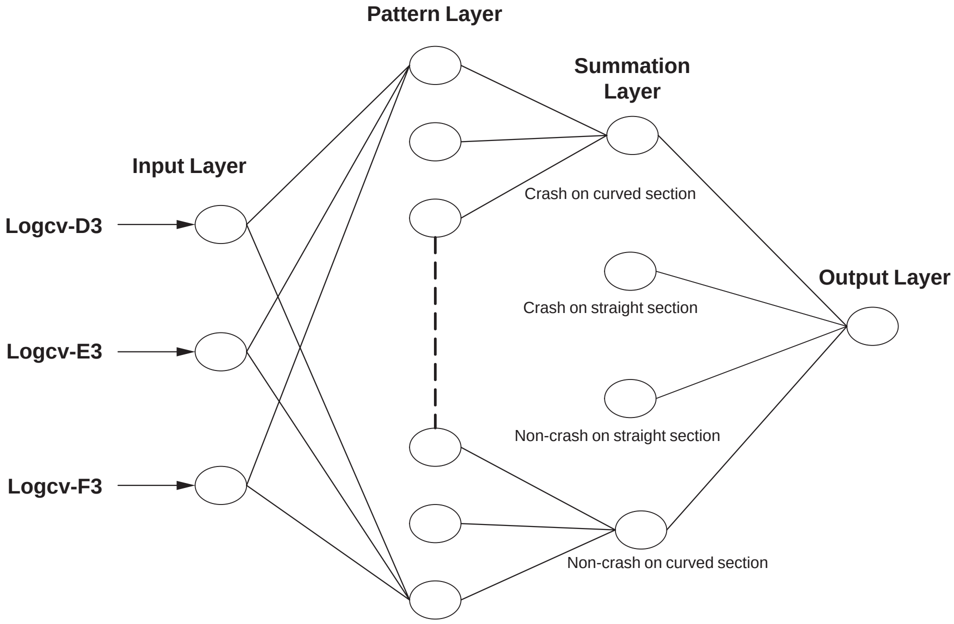

Fig. 5. PNN with best classification accuracy on evaluation dataset when complete crash and non-crash data are used for creation and evaluation.

was developed using the time-limited dataset. In all, four PNN classifiers and the optimal results obtained from them are depicted in Table 7. Note that the input patterns to all these PNN models are three-dimensional, consisting of Logcv-D3, Logcv-E3 and Logcv-F3 (i.e., logarithms of coefficient of variation in speed (Logcvs) at stations D, E and F during time slice 3).

To compare across various models we may observe that the PNN model improves its performance (i.e., reasonable crash identification rate at moderate false alarm rate) when the roadway alignment is incorporated into the classes to be identified (results shown in second row, Table 7). The topology of this network is shown in Fig. 5. The classification performance also improved when time-limited dataset was used for classification between crash and non-crash (i.e., without including the roadway alignment; results shown in the fourth row in Table 7).

It was not possible to develop a time-limited model that accounts for the roadway alignment using these data since relying on the time-limited dataset separating the crashes belonging to straight and curved sections would have resulted in insufficient evaluation sample size. Another point to be noted here is that there is no marked difference between the performances of classical and modified PNN. This implies that in this dataset whether the Euclidean or statistical distance is applied as a measure of nearness in the PNN models, no difference is observed. The reason might be that the three Logcvs used as inputs are equally important and explain the variance in the data in almost equal proportions.

3.1. Proposed real time application

The results from the PNN classifiers show that it is possible to identify more than 70% of the crashes at a reasonable "false alarm" rate using the traffic speed loop data collected from a series of three consecutive loop detector stations, 10–15 minutes prior to the crashes. The real-time application of the models developed here is conceptually simple. On a stretch of a freeway one may collect data from sets of three consecutive loop stations (e.g., a series of 10 loop detectors on a freeway section) which may be divided into sets of three detectors as (1, 2 and 3), (2, 3 and 4), (3, 4 and 5), and so on. The logarithm of coefficients of variation in speed for five minute interval can be continuously calculated and subjected to the PNN models. If patterns emerging from any set of detectors is classified as crash, the freeway segment in the vicinity of the station most downstream of the set of three (as it will correspond to station "F;" station of the crash), may be flagged as a potential crash location.

Once a location is identified for having high potential of crash occurrence it may be flagged with warnings issued through variable message signs (VMS). However, warning the drivers about an impending crash needs more investigation. The effects of such warnings on drivers need to be thoroughly studied. Also, the concept of variable speed limits could be used to intervene and reduce the variation in speeds. Higher speed limits on upstream while lower speed limits on the downstream of a potential crash location, identified by the crash prediction model, could be the basic strategy in applying variable speed limits.

The strategies suggested in this study require extensive research before they may be implemented in the field. However, as an immediate application of the model, some freeway locations that show the hazardous speed variability due to their configuration (e.g., presence of onramp) may be identified. The drivers merging on the freeway through such onramp locations may be warned abut the existing/impending conditions on the freeway. Interstate-4 segment in the vicinity of SR 408 (East-west Expressway) onramp on to I-4 (Eastbound) is one such location with a high number of rear-end crashes. Another possible application for the model may be to have the crash mitigation squad ready near to the locations with high potential of crashes.

4. Conclusions

The performance of multiple PNN models having different combinations of Logcvs (logarithms of coefficient of variation in speed) and average occupancy as input was examined. It was observed that the model achieving optimal classification performance included Logcvs observed 10–15 minutes prior to crash occurrence from three stations: the station of the crash and two stations immediately preceding it in the upstream direction. The parameters used as inputs represent “alarming” traffic conditions, with coefficient of variation defined as standard deviation of speed in five-minute intervals divided by the average speed over the same interval. Lower speeds associated with high variance (resulting in a high value for coefficient of variation) measured on loop detectors depict frequent formation of queues followed by their quick dissipation. In practice these are crash (in fact rear-end crash) prone driving conditions where the drivers need to be very alert while following the vehicles ahead since they would have to slow down and speed up again very often.

The performance of the PNN classifier improves when additional information regarding alignment of the roadway at crash location is provided to the model by increasing the number of classes. Inclusion of time of the day (day time or late night) doesn’t improve the performance of the models. When a time-limited dataset (excluding late-night crashes) was used for training and evaluation of neural networks, the best model, in terms of overall classification accuracy, was achieved. In fact it may always be difficult to predict late-night crashes, since so many of them appear to be caused by sporadic driver errors rather than any turbulence in the traffic flow. The authors also acknowledge the fact that these models are developed using data from a dense urban segment of the freeway where the traffic, crash, and geometric characteristics remain largely uniform (i.e., same *AADT*/peak hour, little or no variation in the geometry along the segment, and mostly rear-end crashes caused by frequent formation and dissipation of ephemeral queues). Hence, these models would perform much better while predicting crashes in the congested regime than compared to a free-flow regime. To predict crashes in the free flow regime, the study area would need to be expanded thereby allowing for diversity in traffic, geometric, and crash characteristics.

The study demonstrates the applicability of loop detector data for identifying crash prone conditions (especially of rear-end type). Once a potential crash location is identified in real-time, measures for reducing the speed variance may be taken in order to reduce the risk. The strategy for such measures, however, should be carefully investigated prior to field application.

Acknowledgement

The authors wish to thank the Florida Department of Transportation for funding this research. We want to also thank Dr. Haitham Al-Deek of UCF and his team for providing the data that was used in the analysis. All opinions and results are those of the authors and do not reflect the opinion of any other party.

References

Abdulhai, B., & Ritchie, S. (1999). Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network. *Transportation Research Part C: Emerging Technologies*, *7*(5), 261 280.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.) New York: John Wiley and Sons, Inc.

Garber, N., & Ehrhart, A. (2000). The effect of speed, flow, and geometric characteristics on crash frequency for two-lane highways. *Transportation Research Record, No. 1717, Transportation Research Board, National Research Council, Washington, D.C.* (pp. 76 83).

Lee, C., Saccomanno, F., & Hellinga, B. (2002). Analysis of crash precursors on instrumented freeways. *Transportation Research Record*, 1784.

Lee, C., Saccomanno, F., & Hellinga, B. (2003). Real-time crash prediction model for the application to crash prevention in freeway traffic. *Transportation Research Record*, 1840.

Masters, T. (1995). *Advanced algorithms for neural networks: A C++ sourcebook*. New York: John Wiley and Sons, Inc.

Oh, C., Oh, J., Ritchie, S., & Chang, M. (2001). Real time estimation of freeway accident likelihood. *Presented at*. *The 80th annual meeting of Transportation Research Board, Washington, D.C.*

Shinar, D. (1999). Speed and crashes: A controversial topic and an elusive relationship. *Traffic Eng.*, *41*, 52 55.

Specht, D. (1996). Probabilistic neural networks and general regression neural networks. In C. H. Chen (Ed.), *Fuzzy Logic and Neural Network Handbook* (pp. 3.1 3.37). Berlin: McGraw-Hill.

Stutts, J. C., Wilkins, J. W., Osberg, J. S., & Vaughn, B. V. (2003). Driver risk factors for sleep-related crashes. *Accident analysis and prevention*, *35*, 321 331.