ARCHIVES OF TRANSPORT *ISSN (print):* *0866-9546*
Volume 43, Issue 3, 2017 ***e-ISSN (online):*** *2300-8830*
DOI: *10.5604/01.3001.0010.4228*

NON-PARAMETRIC MACHINE LEARNING METHODS FOR EVALUATING THE EFFECTS OF TRAFFIC ACCIDENT DURATION ON FREEWAYS

Ying Lee[1], Chien-Hung Wei[2], Kai-Chon Chao[3]
[1] National Kaohsiung Marine University, Kaohsiung, Taiwan
[2] National Cheng Kung University, Tainan, Taiwan
[3] THI Consultants Incorporation, Taipei, Taiwan
[1]e-mail: yinglee1017@gmail.com
[2]e-mail: louiswei@mail.ncku.edu.tw
[3]e-mail: stone_bhm1990@hotmail.com

Abstract: *Traffic accidents usually cause congestion and increase travel-times. The cost of extra travel time and fuel consumption due to congestion is huge. Traffic operators and drivers expect an accurately forecasted accident duration to reduce uncertainty and to enable the implementation of appropriate strategies. This study demonstrates two non-parametric machine learning methods, namely the k-nearest neighbour method and artificial neural network method, to construct accident duration prediction models. The factors influencing the occurrence of accidents are numerous and complex. To capture this phenomenon and improve the performance of accident duration prediction, the models incorporated various data including accident characteristics, traffic data, illumination, weather conditions, and road geometry characteristics. All raw data are collected from two public agencies and were integrated and cross-checked. Before model development, a correlation analysis was performed to reduce the scale of interrelated features or variables. Based on the performance comparison results, an artificial neural network model can provide good and reasonable prediction for accident duration with mean absolute percentage error values less than 30%, which are better than the prediction results of a k-nearest neighbour model. Based on comparison results for circumstances, the Model which incorporated significant variables and employed the ANN method can provide a more accurate prediction of accident duration when the circumstances involved the day time or drunk driving than those that involved night time and did not involve drunk driving. Empirical evaluation results reveal that significant variables possess a major influence on accident duration prediction.*

Key words: *accident duration, correlation, artificial neural networks, k-nearest neighbour method.*

1. Introduction

Traffic accidents usually cause considerable speed reduction and congestion on freeways due to lane closures or obstacles. For fifty large U.S. urban areas, the cost of extra travel time and fuel consumption due to congestion annually amounts to approximately $37.5 billion (Winston & Langer, 2006). To mitigate the impacts due to congestion, traffic management centres usually develop accident management programs. The aims of these programs include exploration of the important factors of accidents, detection of accidents, and provision of accident information forecasts. The impacts of each accident, i.e., duration and resulting congestion queue, may be affected by different features. Relevant features include continuous and/or categorical data, such as accident type, accident characteristics, the number of injuries or fatalities, illumination, type of vehicle involved, road geometry characteristics, and weather conditions. If these data can be processed and analysed effectively, traffic patterns under the influence of accidents could be adequately characterized for various applications in transportation.

Therefore, the objectives of this study are to collect, cross-check, and integrate accident features and traffic data for an accident duration forecasting model on a freeway. The model is based on a related accident database maintained by several public agencies. The proposed forecasting models apply a correlation analysis to select significant variables and employ k-nearest neighbour (kNN) and artificial neural network (ANN) approaches to develop a relationship between the selected variables and accident duration. To demonstrate the performance of the proposed procedure, model performance evaluation is conducted to compare the prediction performance of models with and without correlation analysis and compare the prediction performances for the significant circumstances.

The remainder of this paper is divided into the following sections. Relevant literature is reviewed and assessed in Section 2. Section 3 presents data sources and data analysis. Section 4 provides a brief introduction to the methodologies and a model evaluation indicator. Section 5 illustrates the evaluation results of four prediction models. Finally, Section 6 presents concluding remarks and suggestions for future research.

2. Literature Review

2.1. Accident duration forecast

Many types of incidents occur on highways. Whether it is a serious traffic accident or a falling object, the event can be referred to as an incident that occurs on the road. To reduce the uncertainty of travellers during an incident, several researchers have investigated the relationship between incident duration and traffic/incident data to estimate/forecast accident duration.

Kim and Chang (2011) developed a hybrid prediction model for freeway incident duration. It consists of a rule-based tree model (RBTM), a multinomial logit model (MNL), and a naïve Bayesian classifier (NBC). The decision tree model involves a five-step procedure. It classifies the incident duration data from a database according to incident type, and constructs a rule-based tree under the incident conditions. The results show that incident durations of 120 to 180 minutes and 180 to 240 minutes have satisfactory outcomes. The model performs well for incidents of less than 60 minutes or longer than 300 minutes.

Zhan et al. (2011) applied a regression method and the M5P tree algorithm to predict the lane clearance time of an incident for five scenarios. The model inputs included time of day, day of the week, lighting condition, the number of vehicles involved in the incident, vehicle type involved in the incident, and the number of lanes occupied in the incident. The results of the model showed that incidents that occurred during weekends or those that involved buses or trucks have longer lane clearance times. When the incidents occurred during the daytime period on weekdays, the lane clearance times were shorter. The mean absolute percent error (MAPE) of model performance during prediction was about 42%. When the incident duration was longer than 30 minutes, the prediction error increased and the MAPE value was higher than 78%.

Khattak et al. (2012) analysed traffic incidents and presented iMiT (incident management integration tool) to dynamically predict incident durations. Based on a statistical regression method, the prediction model incorporated time of day, weather conditions, incident location, the number of vehicles involved in the incident, and incident type as the inputs. The MAPE of model performance in estimation and prediction was lower than 55% and displayed reasonable estimation and prediction results.

Li (2015) applied a survival analysis model to develop an incident duration prediction model during three incident duration stages. When the incident duration was between 15 and 60 minutes, the MAPE of model performance was lower than 47% and exhibited reasonable prediction behaviour. When the incident duration was short (less than 15 minutes) or long (greater than 60 minutes), the prediction error was large and the MAPE value was higher than 61%.

Chung et al. (2015) proposed an accelerated failure time model to forecast accident duration and evaluate model performance for the number of lanes blocked. Their results indicated that the accident duration with no blocked lanes was less than those with two or three blocked lanes. However, the accident duration with one blocked lane was less than those with no blocked lanes.

Most studies agree that the data or information collected from management processes can improve the accuracy of predicted incident duration for model development. For incident duration model development, Qi and Teng (2008) defined four categories of input variables according to a USA incident database. Variables used in their model included:

- Weather characteristics: sunny, rainy, and snowy
- Temporal characteristics: AM peak, PM peak, night, and weekday
- Incident characteristics: lanes, property, severity, debris, road repair, and pothole
- Involved vehicle characteristics: bus, van, and truck

During the past few years, a variety of methods have been applied to develop freeway accident duration estimating/forecasting models. The most representative approaches can be classified into the following categories: multivariate regression (Garib et al., 1997; Smith K. & Smith B., 2001; Valenti et al., 2010), fuzzy logic model (Choi, 1996; Dimitriou & Vlahogianni, 2015), artificial neural network (Wang et al., 2005), and survival (Chung et al., 2015; Nam & Mennering, 2000; Chung, 2010; Hojati et al., 2013). Representative studies on highway accident duration prediction over the decade are summarized in Table 1. After assessing the methods frequently employed in the literature, survival analysis (accelerated failure time model) is a popular approach for most researchers and demonstrates acceptable results in freeway accident duration estimation or prediction.

This research differs from most previous studies, which used a regression method or an accelerated failure time model as the key analytical technique for model development. Two non-parametric machine learning methods, namely kNN and ANN, are demonstrated in the freeway accident duration prediction models and performance assessment in this study. Both methods are suitable for modelling complex systems and often achieve a reliable performance. Many studies have demonstrated that kNN and ANN have the potential to accurately predict traffic conditions on highways (Chien et al., 2002; Vlahogianni & Karlaftis, 2013) or on other traffic issue (Spławińska, 2015; Pamula, 2012). Thus, kNN and ANN were chosen as the key analytical techniques in this study.

2.2. Feature selection with correlation analysis

Most researches incorporate high-dimensional data to describe and distinguish complex objects. However, large feature vectors may result in some disadvantages to the model, such as longer model training time and more noise in model development. To avoid these problems, the feature vectors must be properly reduced (Lee & Wei, 2009).

Correlation is a technique for determining whether a linear relationship exists between two variables. The closer the correlation coefficient is to ±1, the stronger the linear relationship between the two variables is. Therefore, conducting a correlation analysis is useful for distinguishing significant independent features from dependent features before model development.

Zhang (2000) presented a prediction algorithm using artificial neural networks. The model was determined by correlation analysis. The parameters of the model can be obtained through nonlinear optimization. Preliminary studies showed that this approach can yield reasonably accurate results.

Table 1. Recent studies of highway accident duration prediction

Researcher	Methodology	Characteristics for model input	Best model performance	Study area
Chung, 2010	Accelerated failure time model	Temporal, Involved vehicle, Accident	MAPE<47%	Korea
Zhan et al., 2011	Regression	Temporal, Involved vehicle, Accident	MAPE<42.7% RMSE<63.46	USA
Khattak et al., 2012	Regression	Temporal, Weather, Accident, Location	MAPE<218% RMSE<17.47	USA
Hojati et al., 2013	Accelerated failure time model	Temporal, Weather, Accident, Traffic		
Li, 2015	Accelerated failure time model	Accident, Season	MAPE<238% RMSE<39.06	China
Li et al., 2015	Competing risk mixture model	Temporal, Traffic, Vehicle, Location	MAPE<94.7% RMSE<26.61	Singapore
Dimitriou and Vlahogianni, 2015	Fuzzy	Weather, Accident, Traffic	MAPE<36%	-
Chung et al., 2015	Accelerated failure time model	Temporal, Weather, Accident, Traffic	-	Taiwan

Guo and Nixon (2009) applied a correlation method to select the important features as the inputs for a pattern recognition model. The experimental results showed that the model selected 37 features from 73 features by the correlation method and achieved 90% classification accuracy rate in pattern recognition.

Based on the discussed literatures, it is clear that conducting research on accident duration is as important as on travel time prediction during an incident. In order to reduce the impact of incidents on travel time prediction, this study identifies significant accident features and develops accident duration prediction models.

3. Data

3.1. Study site

This study selected the Taiwan National Freeway No. 5 (from the Nan-Gang system interchange to the Su-Ao interchange) as the site of the case study. This double-lane road is 54-kilometers long and has seven interchanges as indicated in Figure 1. The distance between two neighbouring vehicle detectors (VDs) is about 2 km. There are five tunnels in the case study site, including the Hsueh-Shan Tunnel, which is the fifth longest tunnel in the world.

3.2. Data sources

Currently, two public agencies, namely the National Police Agency (NPA) and the National Freeway Bureau (NFB), maintain separate raw data regarding traffic accident information on National Freeway No. 5 in Taiwan. The traffic data from NFB includes incident duration and location. The accident data from NPA is the primary source providing detailed information of accident features and environmental factors at an accident site, such as the number of fatalities/injuries, weather conditions, and pavement conditions. To incorporate all information in this study, all data from these two databases require integration and cross-checking. The relevant features of these two databases are listed below:

The National Freeway Bureau

- Incident duration: response time and clearance time;
- Direction: north or south;
- Location: the mileage on National Freeway No. 5;
- Information of involved vehicle: name and phone number of driver, number of vehicle;
- The status of towing: towing or not, including leaving the vehicle on its own and clearing

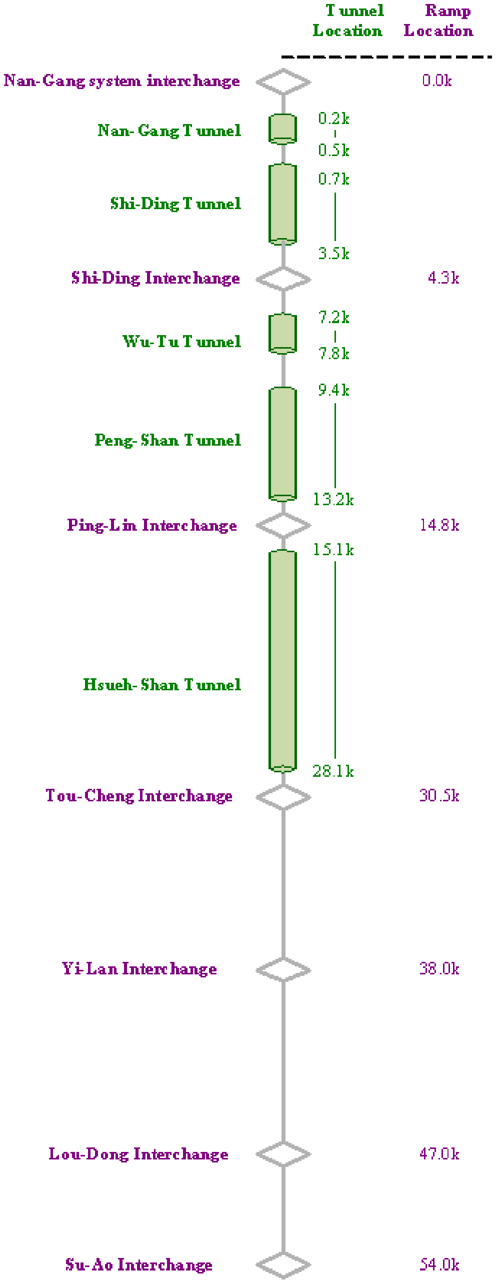

Fig. 1. Layout of Freeway No. 5

The National Police Agency

- Time: possible time of occurrence of confirmed accidents, response time;
- Location: the mileage on National Freeway No. 5;
- Direction: north or south;
- Injuries: number of injuries;
- Weather condition: sunny, rainy, or stormy;
- Type of road: tunnel or elevated road;
- Lighting condition: day time (exclude periods of dawn and dusk) or night time (includes tunnels and underpasses);
- Involved vehicles: type of vehicle involved (e.g., small truck, bus, tractor-semi trailer);
- Accident severity: A1, A2, or A3;
- Pavement condition: dry or wet;

In order to obtain the information of accident duration, this study integrates the above database with the Tow Truck Service Report database maintained by the NFB. Thus, the start time and clear-up time of an accident is practically available to evaluate the associated impact, i.e., accident duration.

A total of 239 accidents on National Freeway No. 5 were recorded during 2012 in the NPA database. However, most accident records are for the purpose of liability appraisal rather than for accident duration prediction. Therefore, integration with the NFB database is required due to the availability of clearance time of accidents. Consequently, a data set of 49 accidents on National Freeway No. 5 is obtained that combines the response time (Nation Police Agency) and clearance time (National Freeway Bureau).

The NFB installs VDs on highways to record traffic data such as speed, volume, and occupancy. Traffic patterns and variations during an accident can be adequately characterized by these data. Therefore, the accident duration can be obtained and verified by analysing these data. This study also incorporates the average speed and average volume as the model features. The traffic data from the VD were accumulated at an interval of five minutes.

3.3. Accident duration

The accident duration in this study represents the period between the time an accident is reported and the time when all handlers leave the accident site. The minimum, maximum, and average durations for 46 accidents were 14, 108, and 42 minutes. Table 2 shows the relative frequency of durations for 46 accidents. For about 56.5% of the accidents, the duration was less than 39 minutes. The percentage of accidents with durations between 40 minutes and 69 minutes was 28.3%, and 15.2% for durations greater than 70 minutes.

The 46 accidents in the sample set were divided into two parts: 60% of the samples were randomly selected as the training data while the remaining samples were categorized as the testing data. The accidents for model training and model testing were sampled randomly based on the relative frequency of duration.

Table 2. Relative frequency of accident duration

Accident duration (min)	# Samples	Frequency	Cumulative frequency
10~39	26	56.5%	56.5%
40~69	13	28.3%	84.8%
70~109	7	15.2%	100.0%

3.4. Independent variables

An accurate accident duration forecast will assist a driver to decrease uncertainty. The factors influencing an accident are numerous and complex. It is a challenge to accurately predict the impact of an accident due to the uncertainties involved. To capture the phenomenon of accidents, the independent variables incorporated in an accident duration forecasting model were selected from the NPA and NFB databases as shown in Table 3. Most accidents occurred during peak hours (52.2%+21.7%), rainy days (37%), night time (58.7%), at road sections with flexible pavements (95.7%), at road sections with direction facility (e.g., jersey barrier) (65.2%), and as a result of drunk driving (45.7%). Type A1 accidents did not occur during the data collection period. Most variables, such as average upstream speed, average upstream volume, time of day, and weather conditions, can be collected immediately from the database after an accident has been reported to the traffic management centre. This study incorporated all the collected variables to develop the accident duration prediction model and evaluate the model performance. Details are presented in Section 5.

Table 3. Independent variables

Features	Variables	Value	# Samples	%
Average speed at upstream	Average speed at upstream	Continuous variable: km/h		
Average volume at upstream	Average volume at upstream	Continuous variable: # vehicles every 1 min.		
Time of day	Non-peak hours during weekdays	Binary variable: 1: Yes, 0: No	9	19.5
	Peak hours during weekdays	Binary variable: 1: Yes, 0: No	24	52.2
	Non-peak hours during the weekend	Binary variable: 1: Yes, 0: No	3	6.5
	Peak hours during the weekend	Binary variable: 1: Yes, 0: No	10	21.7
Weather condition	Cloudy day	Binary variable: 1: Yes, 0: No	3	6.5
	Rainy day	Binary variable: 1: Yes, 0: No	17	37.0
	Stormy day	Binary variable: 1: Yes, 0: No	1	2.2
Illumination	Day time (excludes the dawn and dusk periods)	Binary variable: 1: Yes, 0: No	18	39.1
	Night time (includes tunnels or underpasses)	Binary variable: 1: Yes, 0: No	27	58.7
Road type	Tunnel	Binary variable: 1: Yes, 0: No	11	23.9
(Geographic characteristics)	Elevated road	Binary variable: 1: Yes, 0: No	4	8.7
# injuries	# injuries	Continuous variable: # injuries passengers		
Accident position	Main lane	Binary variable: 1: Yes, 0: No	5	10.9
	Ramp	Binary variable: 1: Yes, 0: No	4	8.7
	The lane to pass the toll station	Binary variable: 1: Yes, 0: No	1	2.2
Pavement type	Pavement type	Binary variable: 1: Flexible, 0: Rigid	44	95.7
Pavement condition	Pavement condition	Binary variable: 1: Wet, 0: Dry	17	37.0
Obstacle	Obstacle	Binary variable: 1: Yes, 0: No	2	4.3
Direction facility	Direction facility	Binary variable: 1: Yes, 0: No	30	65.2
Collision type	Crash into a roadside parapet	Binary variable: 1: Yes, 0: No	7	15.2
	Overtaking collision	Binary variable: 1: Yes, 0: No	6	13.0
	Crash into a safety island	Binary variable: 1: Yes, 0: No	2	4.3
	Turn over	Binary variable: 1: Yes, 0: No	1	2.2
	Crash into a tree	Binary variable: 1: Yes, 0: No	1	2.2
	Rush out of the road	Binary variable: 1: Yes, 0: No	1	2.2
Causation	Unsafe distance	Binary variable: 1: Yes, 0: No	21	45.7
	Drunk driving	Binary variable: 1: Yes, 0: No	1	2.2
	Changing lanes in an unsafe manner	Binary variable: 1: Yes, 0: No	13	28.3
	Breakdown	Binary variable: 1: Yes, 0: No	3	6.5
	Speeding	Binary variable: 1: Yes, 0: No	2	4.3
	Others	Binary variable: 1: Yes, 0: No	1	2.2
Accident severity	Accident severity (A2: People injured during an accident or died after an accident; A3: Property damage)	Binary variable: 1: A2, 0: A3	5	10.9
Type of involved vehicle	Small truck	Binary variable: 1: Yes, 0: No	12	26.1
	Bus	Binary variable: 1: Yes, 0: No	1	2.2
	Tractor-Semi Trailer	Binary variable: 1: Yes, 0: No	1	2.2
# involved vehicles	# involved vehicles	Continuous variable: # vehicles every 1 min.		

#: the number of

4. Methodology

Based on promising performance in the literature, this study employed two non-parametric machine learning methods, namely the k-nearest neighbour and artificial neural networks, to construct the accident duration prediction model. To prepare relevant data for model development, it was desirable to reduce the dimension of accident features using a correlation analysis for all accidents on National Freeway No. 5.

4.1. Model construction – k-Nearest Neighbour method

The kNN Method is a simple and nonparametric approach for both classification and estimation tasks. This effective method has been widely used in previous studies (Chan et al., 2009; Bustillos et al., 2011; Yu et al., 2011; Chen & Rakha, 2014) for travel time prediction.
The procedure of kNN for estimation is as follows:

(1) The samples are divided into two parts. 60% of samples are used for model training and the remaining 40% of samples are used for model testing. $X_{tr}=\{x_1, x_2, \cdots x_i\}$ and $Y_{tr}=\{y_1, y_2, \cdots y_i\}$ represent the training data sets; x_i denotes the data set of independent variables; and y_i indicates the dependent variables. Meanwhile, $X_{te}=\{x_1, x_2, \cdots x_j\}$ and $Y_{te}=\{y_1, y_2, \cdots y_j\}$ represent the testing data sets; x_j denotes the data set of independent variables; and y_j indicates the dependent variables. Further, *i* denotes the training data sample and *j* denotes the testing data sample. Each sample possesses *n* features.

(2) Calculate the distance between each training data x_i and each testing data x_j using the Euclidean distance shown in Equation 1.

$$d(x_i, x_j)=\sqrt{\sum_{n=1}^{N}(x_i^n - x_j^n)^2} \tag{1}$$

(3) Sort $d(x_i, x_j)$ for each testing data x_j in ascending order and select the first K closest training data set $\{y_1, y_2, \cdots y_k\}$ from $Y_{tr}=\{y_1, y_2, \cdots y_i\}$ for the testing data y_j.

(4) Use the kernel function in Equation 2 to average the first K closest training data set $\{y_1, y_2, \cdots y_k\}$ as the estimated y_j.

$$\hat{y}_j=\frac{\sum_{k=1}^{K} d(x_k, x_j) y_k}{\sum_{k=1}^{K} d(x_k, x_j)} \tag{2}$$

4.2. Model construction – Artificial Neural Networks

The ANN model is a structure describing the complex nonlinear relations between input and output variables. An artificial neuron is a computational model inspired by natural neurons. These consist of inputs, which are multiplied by weights to determine the activation of a neuron. Another function computes the output of the artificial neuron. ANNs combine artificial neurons in order to process information. The ANN model is based on the biological neural system and has been widely applied to prediction and classification problems. The most popular learning algorithm is the back-propagation network.
The ANN algorithm was run with the MATLAB software, which allows continuous and categorical data to be defined clearly. The initial weights from the input layer to the hidden layer were default random values and adjusted in the process until the result was stable and acceptable. In the model structure setting, one hidden layer was chosen and the optimal number of hidden units was determined with the performance index embedded in the software. Before model training, 60% of the samples were randomly selected as the training data and the remaining 40% were selected as the testing data. A typical ANN structure is shown in Figure 2, where the output "*y*" denotes the accident duration for the accident duration prediction model; the input "*x*" represents accident and traffic features; "*z*" stands for the node of the hidden layer; and "*w*" indicates the weight of the path. The network paradigm is a multilayer perceptron (MLP).
In this study, the number of neurons in the input layer was determined by the features discussed in Section 3.4. The output layer represented the accident duration for the accident duration prediction model.

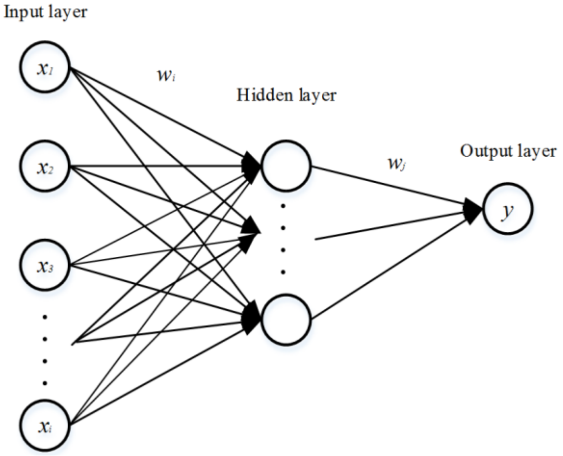

Fig. 2. The structure of an ANN model

4.3. Variable selection – Correlation analysis

To confirm the significant variables and resulting impacts for model construction, all the accident data collected in this study need to be analysed via a correlation analysis. The sample set of accident data collected is a comprehensive collection that covers most of the items recorded in the database. Thus, we first need to specify similar features in order to reduce the complexity of the data set. Relevant literature (Kim & Chang, 2011; Qi & Teng, 2008) indicates that weather conditions, illumination, temporal characteristics, involved vehicle characteristics, and cause of an accident are significant factors in model construction.

Many items for each accident are recorded in the two accident databases. After excluding irrelevant items, correlation analysis is conducted to identify features that have a significant impact on accident duration. The evaluation of accident features is based on the correlation coefficient, *r*. To describe the correlation coefficient, accident duration is defined as the dependent variable *y* and each individual feature is defined as an independent variable *x*.

When an independent variable is both quantitative and continuous, the Pearson correlation coefficient is used to compute the correlation coefficient.

Point-Biserial Correlation is a special case of the Pearson correlation in which the independent variable is a dichotomous variable. When the independent variable is a binary variable, the correlation coefficient is computed by Equation 3 as follows:

$$r_{xy} = r_{pb} = \frac{M_1 - M_0}{S_y} \times \sqrt{\frac{n_1 n_0}{n(n-1)}} = \\ = \frac{M_1 - M_0}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}}} \times \sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{3}$$

where:

- r denotes the correlation coefficient;
- M_1 represents the mean value of *y* when the value of the independent variable *x* is 1;
- M_0 indicates the mean value of *y* when the value of the independent variable *x* is 0;
- n_1 denotes the number of independent variables *x* whose values are 1;
- n_0 represents the number of independent variables *x* whose values are 0.

The results of the correlation analysis for the accident duration prediction model are presented in Table 4. The correlation analysis yielded a high significance at a confidence level of 99% for the following variables: Average volume at upstream, the number of injuries, crash into roadside parapets, and accident severity. An additional five variables, namely rainy day, day time, pavement condition, drunk driving, and the number of vehicles involved, significantly correlated with accident duration at a confidence level of 95%.

Table 4. Correlation coefficients of accident duration and features

Features	Variables	Coeff.	p-values
Average speed at upstream	Average speed at upstream	-0.107	0.478
Average volume at upstream	Average volume at upstream	-0.479 **	0.001
Time of day	Peak hours during weekdays	-0.142	0.347
	Non-peak hours during the weekend	0.156	0.301
	Peak hours during the weekend	-0.155	0.303
Weather condition	Cloudy day	0.179	0.235
	Rainy day	0.305 *	0.039
	Stormy day	0.063	0.677
Illumination	Day time (excludes dawn and dusk periods)	-0.291 *	0.049
	Night time (includes tunnels or underpasses)	0.278	0.062
Road type (Geographic characteristics)	Tunnel	-0.069	0.647
	Elevated road	0.374	0.100
# injuries	# injuries	0.409 **	0.005
Accident position	Main lane	-0.195	0.195
	Ramp	-0.214	0.153
	The lane to pass a toll station	0.152	0.313
Pavement type	Pavement type	0.002	0.987
Pavement condition	Pavement condition	0.314 *	0.034
Obstacle	Obstacle	0.052	0.732
Direction facility	Direction facility	0.130	0.387
Collision type	Crash into a roadside parapet	0.457 **	0.001
	Overtaking collision	0.066	0.665
	Crash into a safety island	0.058	0.702
	Turn over	0.040	0.791
	Crash into a tree	0.083	0.584
	Rush out of the road	0.152	0.313
Causation	Unsafe distance	0.345	0.109
	Drunk driving	0.329 *	0.026
	Changing lanes in an unsafe manner	0.124	0.411
	Breakdown	0.090	0.551
	Speeding	0.075	0.620
	Others	0.219	0.143
Accident severity	Accident severity (A2: People injured during an accident or died after an accident; A3: Property damage)	0.518 **	0.000
Type of involved vehicle	Small truck	0.172	0.253
	Bus	0.032	0.831
	Tractor-Semi Trailer	0.160	0.289
# involved vehicles	# involved vehicles	-0.282 *	0.038

#: the number of

*: indicates that the independent variable significantly correlated with accident duration at a confidence level of 95%.

**: indicates that the independent variable significantly correlated with accident duration at a confidence level of 99%.

4.4. Performance evaluation

Evaluation of model accuracy is required in order to assess the performance of the prediction models. The mean absolute percentage error (MAPE) is a summary measure widely used for evaluating the accuracy of prediction results (Zhan et al., 2011; Khattak et al., 2012; Li, 2015; Dimitriou & Vlahogianni, 2015; Chung, 2010; Li et al., 2015; Li et al., 2016). MAPE was applied in this study to fairly compare relative performance among various model settings.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{V_i^{actual} - V_i^{predicted}}{V_i^{actual}}\right| \times 100\% \qquad (4)$$

where:

- V_i^{actual} denotes the actual value of observation;
- $V_i^{predicted}$ represents the predicted value of observation;
- n indicates the sample size.

The lower the MAPE value is, the more accurate the prediction model will be. A MAPE value less than 10% indicates a highly accurate prediction; a MAPE value between 11% and 20% means a good prediction; and a MAPE value between 21% and 50% refers to a reasonable prediction. The threshold of MAPE was suggested by Lewis (1982).

5. Evaluation

Since the fundamental theory of the kNN and ANN training algorithms are stochastic-oriented, various combinations of initial weights and hidden units may lead to different states of convergence. The literature, however, does not offer a general guideline as to determining the best choice. Therefore, a suitable number of trials should be implemented to verify the performance of the proposed kNN and ANN models.

Given the sample set of 46 accidents prescreened with the corresponding accident duration times, ten experiments were conducted for examining the proposed methodology. In each experiment, 60% of the data was randomly selected as the training set from the sample set and the remaining 40% of data served as the testing set.

Four accident duration prediction models were developed in this study. Both Models 1 and 2 utilized the kNN method as the key algorithm. Model 1 incorporated all the variables in Table 3 while Model 2 incorporated the significant variables in Table 3. Both Models 3 and 4 used the ANN method as the key algorithm. Model 3 incorporated all the variables in Table 3 and Model 4 incorporated the significant variables in Table 3. The same training/testing set was applied to the four models in each experiment.

5.1. Results of the accident duration prediction model

Table 5 depicts the results of the ten experiments. The average MAPE values were below 48% for each type of model, yielding a level of reasonable prediction. For most experiments, Model 4 which applied the ANN method and incorporated the significant variables provided the best prediction results and the MAPE values close to 20%. Based on the results of model evaluation, Model 4 can provide good and reasonable predictions. The performances of the model that incorporated significant variables were better than those that incorporated all the variables. The models that applied the ANN method could predict the accident duration more accurately than those that applied the kNN method.

Table 5. The MAPE values of the accident duration prediction model

	Model 1	**Model 2**	**Model 3**	**Model 4**
Method for model construction	kNN	kNN	ANN	ANN
Independent variables in model	All variables	Significant variables	All variables	Significant variables
Experiments				
1	29.3%	28.8%	25.6%	20.9%
2	41.1%	40.4%	31.7%	21.7%
3	63.8%	57.6%	28.2%	20.1%
4	53.6%	45.3%	30.6%	21.5%
5	49.4%	44.9%	31.6%	21.9%
6	48.8%	46.1%	30.7%	23.7%
7	52.4%	49.8%	31.8%	26.7%
8	34.4%	40.6%	33.4%	29.9%
9	50.5%	57.0%	33.8%	33.8%
10	49.3%	45.9%	31.5%	34.2%
Average	47.3%	45.6%	30.9%	25.4%

Table 6 lists the performance difference among the four models for a further comparison of model performance. The first column of Table 6 compares the performance of the two models which applied the kNN method. The average performance of Model 2 was slightly better than the average performance of Model 1 and the average improvement was about 2.4%. The second column of Table 6 compares the performance of the two models which applied the ANN method. The average performance of Model 4 was better than the average performance of Model 3 and the improvement was about 18.4%. This result indicates that Model 2 and Model 4 may be considered a potential candidate approach to predict accident duration when a suitable set of accident variables is provided.

The third column of Table 6 compares the performance of the two models which incorporated all the variables. The average performance of Model

3 was better than the average performance of Model 1 and the average improvement was about 34.7%. The fourth column of Table 6 compares the performance of the two models which incorporated significant variables. Similarly, the average performance of Model 4 was better than the average performance of Model 2 and the average improvement was about 44.3%. Based on the aforementioned results, the ANN method is more efficient in developing the relationship between traffic/accident data and accident duration than the kNN method when the models incorporate the same variables.

Figure 3 shows the performance assessment with respect to the predicted accident duration vs. actual accident duration. Generally speaking, most data points are scattered along the 45° line with a reasonable distance (discrepancy), especially the plots of Model 4.

Table 6. The difference of model performance

Experiments	Performance improvement from Model 1 to Model 2	Performance improvement from Model 3 to Model 4	Performance improvement from Model 1 to Model 3	Performance improvement from Model 2 to Model 4
1	1.6%	18.3%	12.8%	27.6%
2	1.6%	31.5%	22.9%	46.4%
3	9.8%	28.5%	55.9%	65.0%
4	15.6%	29.8%	43.0%	52.5%
5	9.2%	30.7%	36.1%	51.3%
6	5.5%	22.9%	37.0%	48.6%
7	5.0%	15.9%	39.3%	46.3%
8	-18.1%	10.5%	2.8%	26.4%
9	-12.8%	0.2%	33.0%	40.7%
10	6.9%	-8.8%	36.2%	25.4%
Average	2.4%	18.0%	34.7%	44.3%

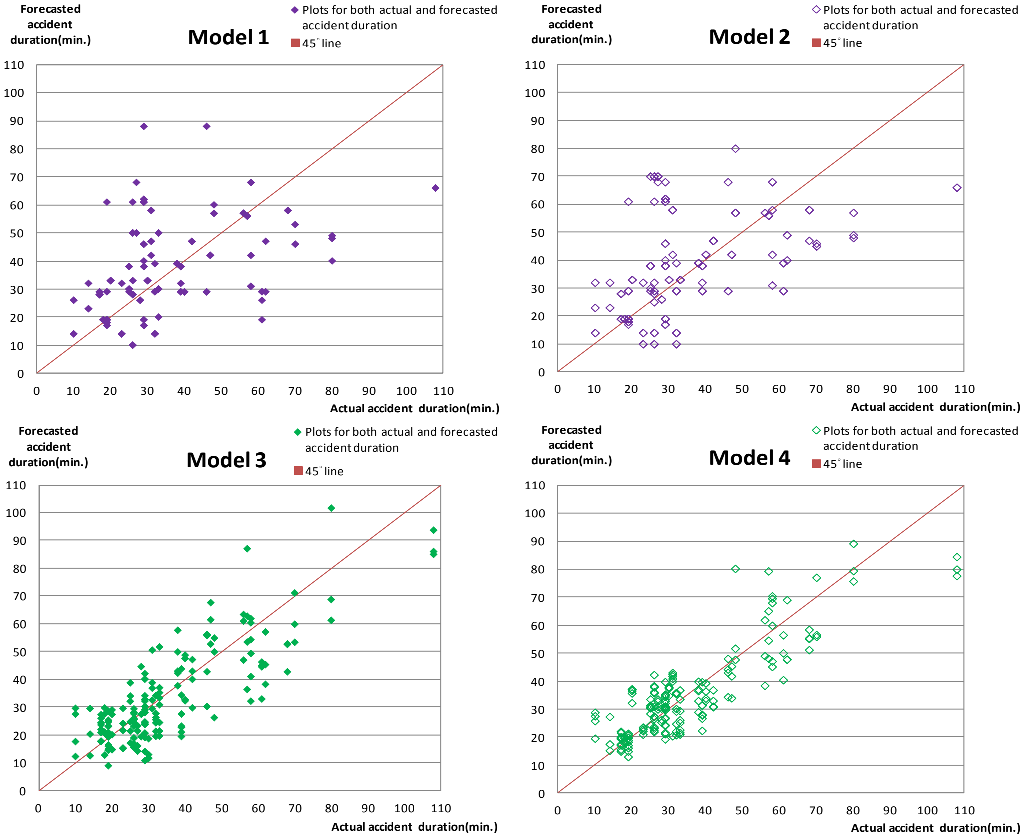

Fig. 3. Assessment results of the four models

As can be seen, most plots of Model 2 are closer to the 45° line than most plots of Model 1; meanwhile, most plots of Model 4 are closer to the 45° line than most plots of Model 3. This indicates that the prediction of models that incorporate significant variables can match the actual accident duration.

Most plots of Model 3 are closer to the 45° line than most plots of Model 1; meanwhile, most plots of Model 4 are closer to the 45° line than most plots of Model 2. This implies that the models that apply the ANN method may sufficiently capture the relationship between the inputs (accident features) and the output (accident duration).

5.2. Performance comparison for circumstances

Table 7 shows the MAPE values and p-values of a t-test for two circumstances of each feature for a comparison of the prediction performance. The statistic t-test was used to test the equality of MAPE values for two circumstances of each feature.

The p-values were greater than 0.05 and the MAPE values for the four models were not significantly different between (i) rainy days and other weather conditions, (ii) dry and wet pavement conditions, and (iii) a crash into a roadside parapet and other collision type. This result means that the four models can provide a similar prediction performance in all types of weather conditions, pavement conditions, and collision type.

Based on the p-values below 0.05, Model 4, which incorporated significant variables and employed the ANN method, can provide a more accurate prediction of accident duration when the circumstances involved the day time or drunk driving than those that involved night time and did not involve drunk driving.

For Model 1 and Model 2, the p-values between Type A2 and Type A3 accidents were less than 0.05 and the MAPE values of the Type A2 accident were significantly lower than those of the Type A3 accident. This result shows that the models that apply the kNN method may better capture the phenomenon of the Type A2 accident than that of Type A3.

Table 7. A comparison of model performance by circumstance

			Model 1	**Model 2**	**Model 3**	**Model 4**
		Method for model construction	**kNN**	**kNN**	**ANN**	**ANN**
Features	**Circumstances**	**Independent variables in model**	**All variables**	**Significant variables**	**All variables**	**Significant variables**
Weather	Rainy day	MAPE	39.3%	46.2%	27.3%	30.3%
	Others	MAPE	51.2%	45.4%	32.6%	23.1%
	t-test for the above two circumstances	p-value	0.118	0.920	0.242	0.109
Illumination	Day time	MAPE	54.9%	45.0%	29.2%	**17.8%**
	Night time	MAPE	42.1%	46.1%	32.1%	30.6%
	t-test for the above two circumstances	p-value	0.077	0.894	0.501	0.003
Pavement condition	Dry	MAPE	49.2%	46.0%	32.6%	23.3%
	Wet	MAPE	43.6%	45.0%	27.5%	30.0%
	t-test for the above two circumstances	p-value	0.461	0.903	0.249	0.160
Collision type	Crash into roadside parapet	MAPE	49.9%	52.3%	28.0%	24.6%
	Others	MAPE	46.8%	44.5%	31.4%	25.6%
	t-test for the above two circumstances	p-value	0.763	0.470	0.571	0.864
Drunk driving	Yes	MAPE	51.3%	44.2%	38.2%	**5.8%**
	No	MAPE	47.2%	45.7%	30.8%	25.8%
	t-test for the above two circumstances	p-value	0.884	0.960	0.652	0.226
Accident severity	A2	MAPE	**21.0%**	**19.8%**	31.7%	22.4%
	A3	MAPE	50.2%	48.5%	30.8%	25.8%
	t-test for the above two circumstances	p-value	0.013	0.022	0.895	0.632

6. Conclusions

An accident data set with 46 cases recorded in two databases during 2012 was built for the accident duration prediction models. The k-nearest neighbour (kNN) and artificial neural network (ANN) approaches were then employed to develop the prediction model when the relevant information regarding accident features/variables were provided. Before model development, a correlation analysis was applied to reduce the scale of interrelated features/variables. Based on the correlation analysis results, Average volume at upstream, the number of injuries, crash into roadside parapets, accident severity, rainy day, day time, pavement condition, drunk driving, and the number of vehicles involved significantly correlated with accident duration. The primary features identified on the case site were consistent with those reported in the literature. The evaluation results of prediction models indicate that the proposed ANN approach is promising as numerical experiments yielded good and reasonable performance in various model compositions based on mean absolute percent error values. The prediction performance can be improved, when the prediction model selected the significant variables and reduced variables dimension. Accurately forecasted accident duration will assist a driver to decrease uncertainty.

For future studies, more accident data should be collected and processed to facilitate the learning capability of the proposed models. Since default settings were mostly used for the time being, a number of parameters needed to be carefully set to further enhance the training mechanism.

References

[1] BUSTILLOS, B. I., CHIU, Y. C., 2011. Real-time freeway-experienced travel time prediction using N-curve and K nearest neighbor methods, *Transportation Research Record*, 2243, pp. 127–137.

[2] CHAN, K. S., LAM, W. H. K., TAM, M. L., 2009. Real-time estimation of arterial travel times with spatial travel time covariance relationships, *Transportation Research Record*, 2121, pp. 102–109.

[3] CHEN, H., RAKHA, H. A., 2014. Real-time travel time prediction using particle filtering with a non-explicit state-transition model, *Transportation Research Part C*, 43 (1), pp. 112–126.

[4] CHIEN, I. J., DING, Y., WEI, C. H., 2002. Dynamic Bus Arrival Time Prediction with Artificial Neural Networks, *Journal of Transportation Engineering*, 128(5), pp. 429–438.CHOI, H. K., 1996. Predicting Freeway Traffic Incident Duration an Expert System Context Using Fuzzy Logic, PhD Dissertation, University of Southern California, Los Angeles, USA.

[5] CHUNG, Y. S., CHIOU, Y. C., LIN, C. H., 2015. Simultaneous equation modeling of freeway accident duration and lanes blocked, *Analytic Methods in Accident Research*, 7, pp. 16–28.

[6] CHUNG, Y., 2010. Development of an accident duration prediction model on the Korean Freeway Systems', *Accident Analysis Preview*, 42(1), pp. 282–289.

[7] DIMITRIOU, L., VLAHOGIANNI, E. I., 2015. Fuzzy modeling of freeway accident duration with rainfall and traffic flow interactions, *Analytic Methods in Accident Research*, 5-6, pp. 59–71.

[8] GARIB, A., RADWAN, A.E., AL-DEEK, H., 1997. Estimating magnitude and duration of incident delays, *Journal of Transportation Engineering*, 123(6), pp. 459–466.

[9] GUO, B., NIXON, M. S., 2009. Gait feature subset selection by mutual information, *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans*, 39 (1), pp. 36–46.

[10] HOJATI, A. T., FERREIRA, L., WASHINGTON, S., CHARLES, P., 2013. Hazard based models for freeway traffic incident duration, *Accident Analysis and Prevention*, 52, pp. 171–181.

[11] KHATTAK, A., WANG, X., ZHANG H., 2012. Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays, *IET Intelligent Transport Systems*, 6(2), pp. 204–314.

[12] KIM, W., CHANG, G. L., 2011. Development of a Hybrid Prediction Model for Freeway Incident Duration: A Case Study in Maryland. *International Journal of Intelligent Transportation Systems Research*, 10(1), pp. 22–33.

[13] LEE, Y., WEI, C. H., 2009. Freeway travel time forecast using Artificial Neural Networks with Cluster Method, *12th International Conference on Information Fusion*, pp. 1331–1338.

[14] LEWIS, C. D., 1982. *Industrial and Business Forecasting Method*. London: Butterworth Scientific.

[15] LI, R., 2015. Traffic incident duration analysis and prediction models based on survival analysis approach, *IET Intelligent Transport Systems*, 9(4), pp. 351–358.

[16] LI, R., PEREIRA, F. C., BEN-AKIVA, M. E., 2015. Competing risks mixture model for traffic incident duration prediction, *Accident analysis and prevention*, 75, pp, 192–201.

[17] LI, T., YANG, Y., WANG, Y., CHEN, C., YAO, J., 2016. Traffic fatalities prediction based on support vector machine, *Archives of Transport*, 39(3), pp. 21–30.

[18] NAM, D., MANNERING F., 2000. An exploratory hazard-based analysis of highway incident duration, *Transportation Research Part A*, 34(2), pp. 85–102.

[19] Pamula, T., 2012. Classification and Prediction of Traffic Flow Based on Real Data Using Neural Networks, *Archives of Transport*, 24(4), pp. 519–529.

[20] QI, Y., TENG, H., 2008. An Information-Based Time Sequential Approach to Online Incident Duration Prediction, *Journal of Intelligent Transportation Systems*, 12(1), pp. 1–12.

[21] SMITH, K., SMITH, B. 2001. Forecasting the Clearance Time of Freeway Accidents, Research Report STL-2001-01. Center for Transportation Studies, University of Virginia, Charlottesville, VA.

[22] Spławińska, M., 2015. Development of models for determining the traffic volume for the analysis of roads efficiency, *Archives of Transport*, 33(1), pp. 81–91.

[23] VALENTI, G., LELLI, M., CUCINA, D., 2010. A comparative study of models for the incident duration prediction, *European Transport Research Review*, 2(2), pp. 103–111.

[24] VLAHOGIANNI, E. I., KARLAFTIS, M. G., 2013. Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties, *Computer-Aided Civil and Infrastructure Engineering*, 28(6), pp. 420–433.

[25] WANG, W., CHEN, H., BELL, M. C., 2005. Vehicle breakdown duration modeling', *Journal of Transportation and Statistics*, 8(1), pp. 75–84.

[26] WINSTON, C., LANGER, A., 2006. The effect of government highway spending on road users' congestion costs. *Journal of Urban Economics*, 60(3), pp. 463–483.

[27] YU, B., LAM, W. H. K., TAM, M. L., 2011. Bus arrival time prediction at bus stop with multiple routes', *Transportation Research Part C*, 19(6), pp. 1157–1170.

[28] ZHAN, C., GAN, A., HADI, M., 2011. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 12(4), pp. 1549–1557.

[29] ZHANG, H. M., 2000. Recursive prediction of traffic conditions with neural network models', *Journal of Transportation Engineering*, 126(6), pp. 472–481.