TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2018 Volume 33(1): 22–31
doi:10.3846/16484142.2015.1004104

# APPLICATION OF NONPARAMETRIC REGRESSION IN PREDICTING TRAFFIC INCIDENT DURATION

Shi Wang[1], Ruimin Li[1], Min Guo[2]

[1]*Dept of Civil Engineering, Tsinghua University, Beijing, China*
[2]*Beijing Traffic Management Bureau, Beijing, China*

Submitted 24 January 2014; resubmitted 29 April 2014; accepted 25 August 2014;
published online 28 January 2015

**Abstract.** Predicting the duration time of incidents is important for effective real-time Traffic Incident Management (TIM). In the current study, the *k*-Nearest Neighbor (*k*NN) algorithm is employed as a nonparametric regression approach to develop a traffic incident duration prediction model. Incident data from 2008 on the third ring expressway mainline in Beijing are collected from the local Incident Reporting and Dispatching System. The incident sites are randomly distributed along the mainline, which is 48.3 km long and has six two-way lanes with a single-lane daily volume of more than 10000 veh. The main incident type used is sideswipe and the average incident duration time is 32.69 min. The most recent one-fourth of the incident records are selected as testing set. Vivatrat method is employed to filter anomalous data for the training set. Incident duration time is set as the dependent variable in Kruskal–Wallis test, and six attributes are identified as the main factors that affect the length of duration time, which are 'day first shift', 'weekday', 'incident type', 'congestion', 'incident grade' and 'distance'. Based on the characteristics of duration time distribution, log transformation of original data is tested and proven to improve model performance. Different distance metrics and prediction algorithms are carefully investigated. Results demonstrate that the *k*NN model has better prediction accuracy using weighted distance metric based on decision tree and weighted prediction algorithm. The developed prediction model is further compared with other models based on the same dataset. Results show that the developed model can obtain reasonable prediction results, except for samples with extremely short or long duration. Such a prediction model can help TIM teams estimate the incident duration and implement real-time incident management strategies.

**Keywords:** traffic incident management; duration prediction; nonparametric regression approach; *k*-nearest neighbor; influence factors.

## Introduction

Traffic incidents represent an important component of non-recurrent traffic congestion (Kwon *et al.* 2006). According to estimates reported by the National Traffic Incident Management Coalition (NTIMC 2006), approximately one-quarter of all congestion in the US roadways is caused by traffic incidents, and blocking a freeway lane for a minute because of an incident results in four minutes of travel delay after incident clearance.

For more than two decades, many Traffic Incident Management (TIM) programs have been implemented in numerous areas and cities in the US under the cooperative work of transportation, public safety, and private sector professionals to clear traffic incidents safely and quickly (Owens *et al.* 2010). TIM has become a key solution to non-recurrent traffic congestion problems (Schrank *et al.* 2012). For an advanced incident management system, accurate and real-time prediction of incident duration is essential for traffic operators to provide timely information to travellers, particularly those approaching the incident scene, through various traveller information systems and traffic control measures.

In China, traffic incidents have also become a main cause of non-recurrent traffic congestion. However, rough estimation is often performed by traffic operators in most traffic control centres. Such estimations mainly depend on the working experience of the traffic operators, and the predictive accuracy may be poor for employees who lack professional skills. Therefore, developing a convincing prediction model for traffic incident duration remains essential.

This study aims to investigate an applicable model to predict traffic incident duration as soon as the traffic

Corresponding author: Ruimin Li
E-mail: *lrmin@tsinghua.edu.cn*



Taylor & Francis
Taylor & Francis Group

management centre obtains initial information from the reporter of the incident. Data on 3744 traffic incidents in the third ring expressway mainline in 2008 were collected from the Incident Reporting and Dispatching System (IRDS) in Beijing. The third ring expressway is a ring-shaped urban expressway in the city of Beijing, which connects a number of city nuclei areas and large residential districts. The present study focuses on the main road of the expressway, which is 48.3 km long with six two-way lanes. The design speed is 80 km/h and the single-lane daily volume is approximately 10000 veh. Around 60 ramps are placed in two directions, and a minimum space of 36 m is placed between on or off ramps. A nonparametric regression model, *k*-Nearest Neighbor (*k*NN) model, was proposed to predict traffic incident duration. The prediction performance and reliability of this algorithm were tested on the basis of the selected traffic incident dataset. Traffic incident duration is defined in this study as the time from when the incident information was obtained to when the incident scene was cleared.

## 1. Literature Review

Several methods have been implemented in the past few decades to develop models for predicting traffic incident duration. Most of these methods can be classified into the following categories:

1) *Linear regression*. To develop a series of truncated linear regression models, Khattak *et al.* (1995) examined 109 accidents that occurred in 1989 and 1990 in Chicago area freeways. Their research established a time sequential procedure for predicting traffic incident duration. Garib *et al.* (1997) developed a Multi-Linear Regression (MLR) model to estimate incident duration. This incident duration prediction model (with an adjusted *R*-square of 0.81) showed that incident duration can be predicted by the number of lanes affected, number of vehicles involved, truck involvement, time of day, police response time, and weather conditions. Khattak *et al.* (2012) applied Ordinary Least Squares (OLS) regression to estimate traffic incident duration. Dynamic models were then constructed based on data from the Hampton Roads area in 2006. A representative OLS model for incident duration prediction was found to have an $R^2$ of 0.255.
2) *Machine learning regression methods*. Boyles *et al.* (2007) developed a probabilistic model based on a naive Bayesian classifier; this model was proven superior to standard linear regression models in predicting traffic incident duration. Wei and Lee (2007) established an adaptive procedure to predict traffic incident duration, including two artificial neutral network-based models; the Mean Absolute Percentage Errors (MAPEs) of these prediction models are mostly under 40%. Yang *et al.* (2008) presented a new prediction model based on the Bayesian decision model to estimate traffic incident duration. Compared with most existing methods, the proposed model was suitable for incomplete data and exhibited better theoretical prediction accuracy. Zhan *et al.* (2011) used the M5P tree algorithm for lane clearance time prediction. The developed model achieved better prediction results than the traditional regression and decision tree models. Kim and Chang (2011) employed a hybrid model to develop a primary estimation system in which they combined rule-based tree model, multinomial logit model, and naïve Bayesian classifier. Wu *et al.* (2011) employed support vector regression to predict traffic incident duration and obtained high accuracy. He *et al.* (2013) proposed a hybrid tree-based quantile regression method that incorporated the merits of both quantile regression modelling and tree structured modelling.
3) *Survival analysis*. Chung (2010) applied survival analysis to develop a model for predicting traffic incident duration based on data from Korea in 2006 and 2007. The estimated parameters for the prediction model demonstrated temporal stability. Qi and Teng (2008) divided the incident management process into several stages and developed a hazard-based regression model for each stage with different variables to improve the accuracy of prediction. Kang and Fang (2011) applied survival analysis to develop the Weibull Accelerated Failure Time (AFT) duration prediction model with nearly three-year traffic incident data on an expressway in Zhejiang Province, China, from 2006 to 2008. The model prediction accuracy is acceptable with the MAPE measurement. Wang *et al.* (2013) applied the AFT models to predict traffic incident duration; the log-logistic distribution produced the best fit for the data from the freeway records maintained by Chinese policemen.
4) *Nonparametric regression*. Nonparametric regression is a form of regression analysis in which the result of regression is not in a predetermined form. The regression model is constructed on the basis of information derived from the data. Nonparametric regression has been applied in various fields, including prediction of traffic flow (Huang *et al.* 2011; Lam *et al.* 2006; Oswald *et al.* 2000) and of traffic incident duration (Smith, K., Smith, B. L. 2002; Boyles *et al.* 2007; Wei, Lee 2007).

*k*NN is a nonparametric method employed mainly for classification according to the closest training examples in the feature space (Tan *et al.* 2005). This methodology can also be used for regression if the target attribute is continuous instead of discrete. Lv *et al.* (2009) applied *k*NN in real-time highway traffic accident prediction and found that this method outperforms the conventional *C*-means clustering method. Oswald *et al.* (2000) discussed the use of nearest neighbor regression in real-time systems and provided general guidelines for optimizing computation speed and data structure. *k*NN is one of the simplest machine learning algorithms. As a type of instance-based learning, *k*NN is adaptive to various tasks in the field of transportation.

Smith, K. and Smith, B. L. (2002) employed the data on 7396 freeway accidents in Virginia to develop a *k*NN nonparametric regression model. A new distance metric,

instead of the traditional Euclidean distance, was used to provide a different weight factor for each independent variable. The straight average of the clearance time of each neighbor was calculated to generate the predicted value. The mean prediction error was the main criterion in determining the neighbor size. Approximately 49.6% of the test accidents were within 15 min of the actual time, and the average error was approximately 20 min.

Wen *et al.* (2012) established a new method to determine the weights of attributes. In the traffic incident duration prediction algorithm, the duration of events with a smaller distance weighs more. Error ≤ 15 min is used as the accuracy index in calculating the optimal value of *k*. When the error is not more than 15 min, lost-load incident has 75.66% accuracy, breakdown incident has 67.25% accuracy, and accident has 84.94% accuracy.

Valenti *et al.* (2010) developed an appropriate distance metric based on the number of matching independent variables between past and current incidents. The prediction value is obtained by weighing the contribution of *k*NNs. The error between the predicted and actual incident durations averages over 17 min.

For the traffic incident duration prediction model development based on the *k*NN algorithm, only a few previous studies have provided an in-depth discussion on data preprocessing. The present study implemented data preprocessing techniques, including data filtering and data transformation, to exclude outliers and improve model performance. Meanwhile, previous studies mainly employed a user-specified single definition of distance metric and prediction algorithm. The current study focuses on further statistical analysis to find the optimal combination of distance metric and prediction algorithm for the specific dataset.

## 2. Data Description

The 2008 data employed in this study were obtained from the IRDS in Beijing. The original data in the IRDS contain all incident events that occurred on all types of roads in the metropolitan area of Beijing. Thus, considering the geographic characteristics and traffic conditions of each road is important. However, specific information on each road is currently unavailable from the system. Several previous studies (Qi, Teng 2008; He *et al.* 2013) have shown that different roads have different effects on traffic incident duration time. Therefore, the incident data on the third ring expressway mainline were selected for further study on model development to exclude the influence of road characteristics. The total length of the study road is 48.3 km and the road has six two-way lanes. The mean value of lane width is 3.75 m and the shoulder width is 2.5 m.

Fig. 1 shows the distribution pattern of traffic incident duration time. The statistical analysis demonstrates that incident duration distribution is right skewed (*skewness* = 4.325). The mean value and standard deviation are 32.69 and 34.42 min, respectively. Approximately 90.0% of the incidents lasted less than 60 min, whereas 4.3% of the incidents lasted longer than 90 min.

In the original data obtained from the IRDS, each record has the same set of attributes, including the geographical information of the incident scene, incident characteristics, and temporal characteristics. To apply the distance metric mentioned in the following sections, all independent variables should be nominal attributes. The continuous variable (the *distance* variable) was divided into several groups to make all attributes nominal. Detailed information on the attributes is listed in Table 1.

Table 1. Variable Kruskal–Wallis test results

| Variable name | | Value | *p*-value |
|---|---|---|---|
| Temporal | Peak hour | 1 – peak hour (7:00–9:00 am and 5:00–7:00 pm);<br>0 – nonpeak hour | 0.194 |
| | Day first shift | 1 – 10:00 p.m – 6:00 a.m;<br>0 – 6:00 a.m – 10:00 p.m | <0.001 |
| | Weekday | 1 – weekday;<br>0 – weekend | 0.002 |
| | Season | 1 – spring;<br>2 – summer;<br>3 – autumn;<br>4 – winter | 0.822 |
| Incident | Incident type | 1 – sideswipe;<br>2 – rear-end crash<br>3 – include pedestrian or bike<br>4 – hit object;<br>5 – rollover;<br>6 – others | <0.001 |
| | Police | 0 – solved by drivers involved in incident;<br>1 – solved by polices | 0.100 |
| | Congestion | 0 – under non-congestion traffic condition;<br>1 – under congestion traffic condition | <0.001 |
| | Incident grade | 1 – include damaged vehicles, no casualties;<br>2 – include injured people, no death;<br>3 – include death | 0.006 |
| | Number of vehicles involved | 1 – one or two;<br>0 – more than two | 0.137 |
| | Taxi | 1 – incident involve taxi;<br>0 – no taxi | 0.208 |
| | Bus | 1 – incident involve bus;<br>0 – no bus | 0.152 |
| | Truck | 1 – incident involve small truck;<br>2 – incident involve large truck;<br>0 – no truck | 0.088 |
| Geographic | Distance | The distance from the incident scene to the central Beijing:<br>1 – less than 5 km;<br>2 – 5–6 km;<br>3 – 6–7 km;<br>4 – 7–8 km;<br>5 – more than 8 km | <0.001 |

## 3. Model Development

According to the basic concept of the *k*NN algorithm, the first step of modelling should establish the training sample dataset. Calibration of the distance metric is then required to measure the degree to which the given test sample is similar to each training sample. When *k*NN is used for regression, the property value of the test sample would be the average (or weighted average) of its *k*NN.

Several performance measures can be applied to evaluate prediction model development. The main measures of effectiveness include the Mean Absolute Error (MAE), MAPE, and Root Mean Squared Error (RMSE). Smith, K. and Smith, B. L. (2002) indicated that as a supplementary means, the percentage of predictions within a certain tolerance of their actual incident duration can be useful in predicting the duration of traffic incidents. These evaluation indices are calculated with the optimal value of neighbor size as:

$$MAE = \frac{\sum_{i=1}^{N} \left| t_{pi} - t_{0i} \right|}{N}; \quad (1)$$

$$MAPE = \frac{\sum_{i=1}^{N} \left| \frac{t_{pi} - t_{0i}}{t_{0i}} \right|}{N}; \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( t_{pi} - t_{0i} \right)^2}, \quad (3)$$

where: $t_{pi}$ is the predicted value of traffic incident duration; $t_{0i}$ is the actual value of traffic incident duration for the *i*th incident record; *N* is the total number of testing samples.

### 3.1. Division of Training and Testing Set

The total population of incident records on the third ring expressway mainline in 2008 was collected based on types such as fender bender, severe traffic crash, or vehicle on fire. Several studies (Wu *et al.* 2011; Demiroluk, Ozbay 2011) randomly selected the testing samples to examine the effectiveness of the prediction model. However, for a real-time TIM system, the goal of modelling is to predict the duration of the present case based on knowledge of history incident records in Beijing. Thus, the testing samples were selected in a chronological manner as in other studies (Lin *et al.* 2004; Kim *et al.* 2008).

In this study, the most recent one-fourth of incident records (936 records) were selected as testing samples for model evaluation and the earlier three-fourths (2808 records) as training samples to establish the historical database. The main incident type used is sideswipe for both the testing and training sets.

### 3.2. Filtering Anomalous Data

Before the next phases of model development, the extreme values in the training set should be excluded to reduce the effect of outliers. The Vivatrat method can be a useful tool to address anomalous values, which has been applied in processing injurious road crash rates (Dell'Acqua *et al.* 2013). In this study, the procedure is outlined as follows:

- dividing the samples into six groups by crash type;
- calculating the mean and deviation of duration time for each group of samples, and ordering the groups that increase the mean value of duration time; and
- calculating the representative dispersion $S_r$ for each group as follows:

$$S_r = \frac{1}{2} \min\left( S_{i+1} + S_i, S_{i-1} + S_i, S_{i+1} + S_{i-1} \right), \quad (4)$$

where: $S_{i+1}$, $S_i$ and $S_{i-1}$ represent the standard deviation of the $(i+1)$th, *i*th, and $(i-1)$th groups, respectively.

- For each group, the samples with duration time outside the range of $\mu_i \pm AS_r$ are removed, where $\mu_i$ is the mean of the duration time for the *i*th group. The value of *A* is set as 2.5.

### 3.3. Data Transformation

The typical right-skewed distribution pattern of incident duration time is consistent with the results of several relevant studies (Wang *et al.* 2013; Zhan *et al.* 2011; Valenti *et al.* 2010). However, modelling techniques often have difficulties dealing with data with a wide value range. For the data used in this study, most of the samples are located in the time interval of 20 min to 40 min, as shown in Fig. 1. The duration time of the selected nearest neighbors may be concentrated in a certain time interval, which might compromise the prediction accuracy. Thus, a certain type of data transformation may help make the data more applicable to the developed model. Previous studies (Garib *et al.* 1997; Valenti *et al.* 2010) that utilized multiple linear regression to predict traffic incident duration applies log transformation to induce symmetry and meet the normal distribution assumption. In the present study, log transformation of the

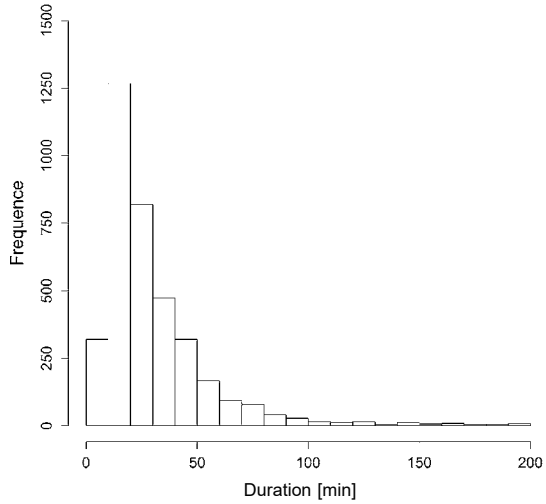


Fig. 1. Distribution of traffic incident duration

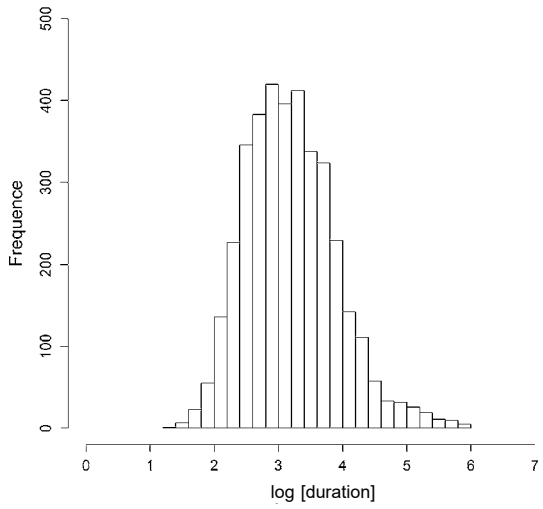


Fig. 2. Distribution of traffic incident duration after log transformation

dependent variable (incident duration time) was tested. Fig. 2 illustrates the relatively uniform distribution of the duration time after log transformation, which might mitigate the potential problems resulting from a right-skewed duration distribution with wide data range. The transformed data are compared with original data in Section 4.3 to validate the effect of the transformed data on prediction results and to investigate their influence on prediction performance.

### 3.4. Distance Metric

The core of the *k*NN algorithm lies in the approach to determine the *k*NNs. One of the most prevalent means to measure distance is Euclidean distance. However, this measure is mainly applied to data with continuous variables. According to the dataset employed in this study, all attributes are nominal for each incident record obtained from the IRDS, as listed in Table 1. Thus, several distance metrics suitable for nominal attributes are further examined. For nominal attributes, the simplest distance metric is the overlap metric (Li, C., Li, H. 2010). Nevertheless, this measure of distance remains rough because it does not consider the additional information provided by the nominal attribute values, which may be helpful in generalization (Wilson, Martinez 1997). Therefore, in the present study, attribute weighted overlap metric was adopted as follows:

$$d\left(x_1,x_2\right)=\sum_{i=1}^{n}w_i\delta\left(a_i\left(x_1\right),a_i\left(x_2\right)\right), \tag{5}$$

where: $n$ is the number of attributes; $a_i\left(x_1\right)$ and $a_i\left(x_2\right)$ are the values of the $i$th attribute of incident record $x_1$ and $x_2$, respectively; $\delta\left(a_i\left(x_1\right),a_i\left(x_2\right)\right)$ is 0 if $a_i\left(x_1\right)=a_i\left(x_2\right)$ and 1 otherwise; $w_i$ is the weight of the $i$th attribute.

Smith, K. and Smith, B. L. (2002) defined weight factor $w_i$ for each attribute based on the absolute difference in the means of different groups of samples for the particular attribute. With the attribute *season* as an example, the total population of incidents is divided into four groups (spring, summer, autumn, and winter). If incident sample $x_1$ occurred in *spring* and $x_2$ occurred in *summer*, then the absolute difference between the mean of incident duration time for all incidents in *spring*, $T_1$, and that for all incidents in *summer*, $T_2$, would be calculated as the weight factor for attribute *season* between $x_1$ and $x_2$:

$$w_i=\left|T_2-T_1\right|. \tag{6}$$

Wen *et al.* (2012) proposed a similar approach in which $\max\left(T_2/T_1,T_1/T_2\right)$ is calculated as weight factor $w_i$:

$$w_i=\max\left(T_2/T_1,T_1/T_2\right). \tag{7}$$

The two methods discussed are easy to understand but they consider only the mean value of the incident duration for different groups in the distance metric. A more elaborate attribute weighting method considers the degree to which a particular predictive attribute depends on the values of other attributes.

A decision tree can be employed for classification and regression, with the resulting model presented in the form of a tree structure. The decision tree as an independent prediction technique has been used in attribute selection (Ratanamahatana, Gunopulos 2003) and attribute weighting (Hall 2007) for naive Bayes algorithm. Based on the method employed by Hall (2007), in this study, the weight of each predictive attribute is inversely related to its degree of dependency on other attributes. After an ensemble of unpruned C4.5 decision trees is constructed, the minimum depth in which an attribute was tested at the built decision trees was used to weigh that attribute:

$$w_i=\frac{\sum_{j=1}^{N}\frac{1}{\sqrt{d_j}}}{N}, \tag{8}$$

where: $N$ is the number of bagging iterations; $d_j$ is the minimum depth in which the $i$th attribute was tested at the decision tree in the $j$th iteration.

$N$ is tested from 10 to 50. For each given $N$, the fraction of input data to sample with replacement from the input data for growing each new tree are tested from 10% to 100%. The attribute weight is determined by the ensemble of trees, which obtains the minimum mean square error for the training set.

For $w_i$ in this study, three methods were compared based on the same dataset to select the best one to apply.

### 3.5. Prediction Algorithm

In a typical *k*NN algorithm, two samples are usually assumed to have a high degree of similarity if the distance between them is small. Let $t_p$ be the predicted incident duration time of the given test sample $p$ and $t_1,t_2,...,t_k$ be the duration time of the $k$ neighbors. The distance between test sample $p$ and the $k$ neighbors is written as $d_1,d_2,...,d_k$. The predicted value is calculated by the straight average algorithm [Equation (9)] or weighted

average algorithm [Equation (10)]:

$$t_p = \frac{\sum_{i=1}^{k} t_i}{k}; \quad (9)$$

$$t_p = \sum_{i=1}^{k} W_i t_i, \quad (10)$$

where: $W_i$ is the weighting factor for the $i$th neighbor. Typically, the influence of each neighbor is determined by its distance, $W_i = 1/d_i^p$, a method known as inversed distance weighting. However, in actual modelling, the distance between the testing sample and the neighbor is sometimes equal to zero because all independent variables are nominal attributes. To obtain reasonable prediction results, a non-zero smoothing parameter δ was assigned in this study as follows:

$$W_i = \frac{\frac{1}{(d_i+\delta)^p}}{\sum_{i=1}^{k} \frac{1}{(d_i+\delta)^p}}, \quad (11)$$

where: δ and $p$ were set as 0.5 and 1.0, respectively, based on the rule of thumb. The predictive results between straight average and weighted average were compared in this study.

### 3.6. Optimal Value of *k*

The selection of neighborhood size $k$ can significantly influence the prediction result. A small value of $k$ may result in a large variance, and the predictive result can be compromised by abnormal values in the dataset. Conversely, with a large value of $k$, the developed model tends to include cases that are actually far from the test sample and therefore makes the neighbors hardly representative of the test sample. Thus, an appropriate value of $k$ is required to achieve balance. In this study, the optimal value of $k$ is determined through empirical testing of the model. More specifically, for each $k$ from 1 to 100, the prediction results of the testing set are obtained from the $k$ nearest neighbors in the training set. The overall MAPE for the testing set is then calculated. We assume that the optimal value of $k$ reaches the minimum overall MAPE.

## 4. Results

In the current study, the training set is used to determine the main influencing factors of the incident duration time and to calculate the attribute weights for the distance metric. Moreover, models with different distance metrics and prediction algorithms are measured separately against the testing set to find the optimal model specification.

### 4.1. Data Filtering

According to the methodology introduced in Section 3.2, the residual range of duration time for each incident type is calculated and presented in Table 2. A total of 3.28% of the original training set samples (92 records) are excluded, and the remaining data are included in the historical database to predict the duration time of the testing set.

### 4.2. Attribute Selection

As a large number of independent variables can be identified from the available incident data, determining the relationship between the value of each attribute and incident duration is an essential task. Statistical analysis indicated that the two basic assumptions of normality and homogeneity of variances requested by the ANOVA test cannot be met for all variables. Therefore, the nonparametric Kruskal–Wallis test was performed to identify the statistically relevant variables for predicting incident duration. With the level of significance set to 0.05, six independent variables were found to be statistically significant. The variables and corresponding Kruskal–Wallis test results are shown in Table 1. Consistent with the expected results, the $p$-values exhibited the following significant attributes: 'day first shift', 'weekday', 'incident type', 'congestion', 'incident grade' and 'distance'. The results in the following sections include all of these six variables.

### 4.3. Log Transformation Versus Original Data

To study the influence of data transformation on the prediction result, we first examined the effect of log transformation on the dependent variable (incident duration) before developing other parts of the model. Equation (6) was employed for the distance metric, and Equation (9) was used to generate the prediction value.

Table 2. Range of incident duration time for each indicent type

| Incident type | Mean | Standard deviation | $S_r$ | Residuals range for duration time | Remaining sample size | Removed sample size |
|---|---|---|---|---|---|---|
| Rollover | 87.22 | 101.29 | 53.46 | [–46.45, 220.88] | 10 | 1 |
| Hit object | 71.87 | 77.33 | 53.46 | [–61.79, 205.53] | 48 | 5 |
| Others | 71.50 | 29.60 | 44.63 | [–40.08, 183.08] | 5 | 1 |
| Include pedestrian or bike | 54.71 | 59.67 | 37.23 | [–38.37, 147.80] | 19 | 1 |
| Sideswipe | 32.24 | 33.99 | 29.84 | [–42.35, 106.83] | 2502 | 81 |
| Rear-end crash | 28.77 | 25.68 | 25.68 | [–46.45, 92.97] | 132 | 3 |

Table 3 demonstrates the divergence of prediction accuracy between the results before and after log transformation of incident duration. The transformed data are superior to the original data in all of the four measurements applied in this study, with a decrease of 8.58% in MAPE and a decrease of 1.59 min in MAE, which can be considered as a significant improvement. As an effective and helpful procedure in data preprocessing, log transformation was applied in all of the subsequent sections of this paper.

Table 3. Prediction results before and after log transformation of duration time

| | Optimal $k$ | MAE | MAPE | RMSE | $Err < 15$ min |
|---|---|---|---|---|---|
| Before | 5 | 16.66 | 62.38% | 28.05 | 56.64% |
| After | 53 | 15.17 | 53.80% | 27.78 | 73.18% |

### 4.4. Distance Metric and Prediction Algorithm

Three approaches to the distance metric and two to the prediction algorithm are discussed in this study. Considering the possibility of potential correlation of the distance metric and prediction algorithm, we combined different approaches to examine the prediction performance. For each combination, the optimal value of $k$ is determined and prediction accuracy is calculated. For the attribute weights based on bagging decision trees in Equation (10), the minimum square error for the training set is reached with 25 bagging iterations and a sampling fraction of 50%.

Table 4 shows the prediction results of these combinations. The optimal values of $k$ with combinations 1, 2, and 5 are significantly larger than the remaining combinations, thereby requiring additional computing time for practical application. The six combinations produced similar results in performance measurements. Generally, the mean average errors are approximately 15 min, and the mean average percentage errors are around 50%. Since none of the combinations are superior to others in all of these measurements, MAE and MAPE are given more consideration. The comparison of MAPE and MAE shows a slight difference. Combination 6 [Equations (8) and (10)], which is slightly better than the others in MAPE and MAE, has been retained for result analysis.

### 4.5. Result Analysis

The overall MAPE is 50.12% for the final model. Typically, a prediction model with its MAPE between 20% and 50% is capable of reasonable forecasting. If MAPE is larger than 50%, the forecasting is inaccurate (Lewis 1982). A careful examination of the results of different duration ranges reveals that the prediction model can provide reasonable prediction accuracy for incidents with duration between 15 and 60 min. However, for incidents with duration longer than 60 min or shorter than 15 min, the model hardly provides convincing results (Table 5). Thus, similar to the model in the previous study by Valenti *et al.* (2010), the present model cannot address extreme values in an acceptable manner.

The percentage of records within a given tolerance of prediction error was used as an evaluation index, as shown in the first three columns of Table 7. Approximately 3.4% of the total number of incidents has a prediction error larger than 1 h. These outliers cannot be effectively treated by the developed model. One possible explanation is the absence of several potential independent variables. For instance, the duration of a specific traffic incident can vary depending on the attitude of the drivers involved. The difference in the experience and knowledge of incident response personnel can also contribute to the divergence of traffic incident duration. Some important details in predicting the incident duration of outliers may be absent because of the relatively brief information provided by the system at the initial stage.

The $k$NN algorithm with model specification employed by Smith, K. and Smith, B. L. (2002) and stepwise MLR were also tested based on the same dataset used in this study. Table 6 demonstrates the model performance

Table 4. Prediction results with different distance metrics and prediction algorithms

| No | Distance | Prediction | Optimal $k$ | MAE | MAPE | RMSE | $Err < 15$ min |
|---|---|---|---|---|---|---|---|
| 1 | Equation (6) | Equation (9) | 53 | 15.17 | 50.26% | 27.78 | 73.18% |
| 2 | Equation (6) | Equation (10) | 87 | 15.18 | 50.39% | 27.60 | 72.40% |
| 3 | Equation (7) | Equation (9) | 35 | 15.22 | 50.25% | 27.86 | 73.08% |
| 4 | Equation (7) | Equation (10) | 35 | 15.18 | 50.28% | 27.77 | 72.86% |
| 5 | Equation (8) | Equation (9) | 87 | 15.20 | 50.16% | 28.01 | 73.82% |
| 6 | Equation (8) | Equation (10) | 35 | 15.15 | 50.12% | 27.92 | 72.86% |

Table 5. Prediction results for incidents lasting longer than 60 min or shorter than 15 min

| Incident duration | Count | Number of prediction with MAPE > 0.5 | Percent of prediction with MAPE > 0.5 |
|---|---|---|---|
| <15 min | 226 | 201 | 88.94% |
| >60 min | 79 | 73 | 92.41% |

for records with different incident durations. The other two models show similar overall prediction performance as the *k*NN model. The developed *k*NN model exhibits its superiority in predicting incidents that last less than 30 min, which is reflected in the improvement in MAE, RMSE, and MAPE for the first two groups.

Table 7 indicates the difference in the measurement of 'percentage of records within a given tolerance of prediction error'. Compared with the MLR model and the previous model (Smith, K., Smith, B. L. 2002), the *k*NN model performs more effectively in measuring prediction error less than 5, 10, and 15 min.

In other studies, Zhan *et al.* (2011) applied the M5P algorithm to predict incident duration. Seventy-eight percent of the total incidents were predicted with less than 30 min of prediction error. Chung (2010) employed the AFT metric model; 61% of the total incidents had a prediction error less than 15 min and 85% had a prediction error less than 30 min. In the present study, 72.9% and 89.4% of the total incidents have prediction errors less than 15 and 30 min, respectively (Table 7).

Generally, the model developed in this study can serve as an effective tool for traffic management teams to obtain a timely estimation of incident duration once the incident is reported by drivers or the police to the traffic control centre. System operators can implement further measures, such as traffic guidance and traffic control, to minimize the negative effects of incidents.

## Conclusions

The *k*NN algorithm was applied in this study to develop a prediction model of traffic incident duration based on 3744 incident data on the third ring expressway mainline. The data were recorded by the IRDS in Beijing, China.

Anomalous data are filtered from the training set by Vivatrat method. Six attributes are determined as the main influencing factors of traffic incident duration time by Kruskal–Wallis test, which are 'day first shift', 'weekday', 'incident type', 'congestion', 'incident grade' and 'distance'. Log transformation of original data is tested, which showed significant improvement.

Results demonstrate that the combination of weighted distance metric based on the decision tree and weighted prediction algorithm exhibits the best prediction accuracy. Overall, the developed model can obtain reasonable prediction results except for incidents that last shorter than 15 min or longer than 60 min. The model in this study is superior to the MLR model and *k*NN model developed by Smith, K. and Smith, B. L. (2002) in predicting incidents that last less than 30 min.

The study road is characterized by small ramp spacing and high traffic volume. The main incident type is sideswipe and the average duration is 32.69 min. The final model, which is determined based on specific data in this study, is optimized through data filtering, data transformation, and selection of optimal distance metric and prediction algorithm. The procedure of model development could provide a reference for further research on the application of *k*NN in incident duration time prediction with a different dataset.

Additional information, such as feedback from the involved drivers and the traffic police at the scene, may improve the prediction performance of the developed model. Moreover, the combination of the prediction results of different types of nonparametric techniques may lead to better model performance. Future studies could integrate different prediction methods to improve prediction accuracy.

## Acknowledgements

The research was supported by the Beijing Committee of Science and Technology, China (Grant No Z121100000312101).

Table 6. Prediction results for incidents with different duration ranges

| Incident duration | Count | *k*NN | | | Model by Smith | | | MLR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | MAPE | MAE | RMSE | MAPE | MAE | RMSE | MAPE |
| 0–15 | 226 | 10.32 | 10.74 | 100.96% | 12.20 | 12.79 | 118.23% | 12.15 | 13.15 | 117.82% |
| 15–30 | 373 | 4.31 | 5.36 | 21.05% | 5.19 | 6.69 | 26.71% | 5.57 | 6.73 | 27.26% |
| 30–60 | 258 | 17.35 | 19.45 | 41.28% | 16.25 | 17.28 | 35.39% | 16.31 | 18.21 | 38.22% |
| 60–120 | 64 | 54.10 | 56.32 | 67.57% | 51.08 | 53.39 | 63.69% | 49.45 | 52.55 | 61.88% |
| >120 | 15 | 153.14 | 161.67 | 84.64% | 150.29 | 159.18 | 82.87% | 146.53 | 134.49 | 80.80% |
| Total | 936 | 15.15 | 27.92 | 50.12% | 15.39 | 27.21 | 54.63 | 15.38 | 27.00 | 55.37% |

Table 7. Percent of records within a given tolerance of prediction error

| Prediction error | *k*NN | | Model by Smith | | MLR | |
|---|---|---|---|---|---|---|
| | Count | Percent | Count | Percent | Count | Percent |
| ≤5 min | 264 | 28.2% | 236 | 25.2% | 241 | 25.7% |
| ≤10 min | 523 | 55.9% | 473 | 50.5% | 485 | 51.8% |
| ≤15 min | 682 | 72.9% | 658 | 70.3% | 668 | 71.4% |
| ≤30 min | 837 | 89.4% | 845 | 90.3% | 844 | 90.2% |
| ≤60 min | 901 | 96.6% | 904 | 96.6% | 905 | 96.7% |

## References


Boyles, S.; Fajardo, D.; Waller, S. T. 2007. Naive Bayesian Classifier for incident duration prediction, in *TRB 86th Annual Meeting Compendium of Papers CD-ROM*, 21–25 January 2007, Washington, DC, 1–11.

Chung, Y. 2010. Development of an accident duration prediction model on the Korean freeway systems, *Accident Analysis & Prevention* 42(1): 282–289. http://dx.doi.org/10.1016/j.aap.2009.08.005

Dell'Acqua, G.; Russo, F.; Biancardo, S. A. 2013. Risk-type density diagrams by crash type on two-lane rural roads, *Journal of Risk Research* 16(10): 1297–1314. http://dx.doi.org/10.1080/13669877.2013.788547

Demiroluk, S.; Ozbay, K. 2011. Structure learning for the estimation of non-parametric incident duration prediction models, in *TRB 90th Annual Meeting Compendium of Papers DVD*, 23–27 January 2011, Washington, DC, 1–19.

Garib, A.; Radwan, A.; Al-Deek, H. 1997. Estimating magnitude and duration of incident delays, *Journal of Transportation Engineering* 123(6): 459–466. http://dx.doi.org/10.1061/(ASCE)0733-947X(1997)123:6(459)

Hall, M. 2007. A decision tree-based attribute weighting filter for naive Bayes, *Knowledge-Based Systems* 20(2): 120–126. http://dx.doi.org/10.1016/j.knosys.2006.11.008

He, Q.; Kamarianakis, Y.; Jintanakul, K.; Wynter, L. 2013. Incident duration prediction with hybrid tree-based quantile regression, in Ukkusuri, S. V.; Ozbay, K. (Eds.). *Advances in Dynamic Network Modeling in Complex Transportation Systems*, 287–305. http://dx.doi.org/10.1007/978-1-4614-6243-9_12

Huang, Z.; Ouyang, H.; Tian, Y. 2011. Short-term traffic flow combined forecasting based on nonparametric regression, in *Proceedings of the 2011 International Conference on Information Technology, Computer Engineering and Management Sciences (ICM)*, 24–25 September 2011, Nanjing, Jiangsu, China, 1: 316–319. http://dx.doi.org/10.1109/ICM.2011.89

Kang, G.; Fang, S.-E. 2011. Applying survival analysis approach to traffic incident duration prediction, in *ICTIS 2011: Multimodal Approach to Sustained Transportation System Development: Information, Technology, Implementation*, 30 June – 2 July 2011,Wuhan, China, 1: 1523–1531. http://dx.doi.org/10.1061/41177(415)193

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): 204–214. http://dx.doi.org/10.1049/iet-its.2011.0013

Khattak, A. J.; Schofer, J. L.; Wang, M.-H. 1995. A simple time sequential procedure for predicting freeway incident duration, *IVHS Journal* 2(2): 113–138. http://dx.doi.org/10.1080/10248079508903820

Kim, W.; Chang, G.-L. 2012. Development of a hybrid prediction model for freeway incident duration: a case study in Maryland, *International Journal of Intelligent Transportation Systems Research* 10(1): 22–33. http://dx.doi.org/10.1007/s13177-011-0039-8

Kim, W.; Chang, G.-L.; Rochon, S. M. 2008. Analysis of freeway incident duration for ATIS applications, in *Proceedings of the 15th World Congress on Intelligent Transport Systems and ITS America Annual Meeting 2008*, 16–20 November, New York, NY, USA, 2: 950–958.

Kwon, J.; Mauch, M.; Varaiya, P. 2006. Components of congestion: delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand, *Transportation Research Record* 1959: 84–91. http://dx.doi.org/10.3141/1959-10

Lam, W. H. K.; Tang, Y. F.; Tam, M.-L. 2006. Comparison of two non-parametric models for daily traffic forecasting in Hong Kong, *Journal of Forecasting* 25(3): 173–192. http://dx.doi.org/10.1002/for.984

Lewis, C. D. 1982. *Industrial and Business Forecasting Methods: a Practical Guide to Exponential Smoothing and Curve Fitting*. Butterworth Scientific. 143 p.

Li, C.; Li, H. 2010. A survey of distance metrics for nominal attributes, *Journal of Software* 5(11): 1262–1269. http://dx.doi.org/10.4304/jsw.5.11.1262-1269

Lin, P.-W.; Zou, N.; Chang, G.-L. 2004. Integration of a discrete choice model and a rule-based system for estimation of incident duration: a case study in Maryland, in *Transportation Research Board 83rd Annual Meeting Compendium of Papers CD-ROM*, 11–15 January 2004, Washington, DC.

Lv, Y.; Tang, S.; Zhao, H. 2009. Real-time highway traffic accident prediction based on the k-nearest neighbor method, *Proceedings of the ICMTMA'09: International Conference on Measuring Technology and Mechatronics Automation*, 11–12 April 2009, Zhangjiajie, Hunan, China, 3: 547-550. http://dx.doi.org/10.1109/ICMTMA.2009.657

NTIMC. 2006. *Benefits of Traffic Incident Management*. National Traffic Incident Management Coalition (NTIMC). 8 p. Available from Internet: http://ntimc.transportation.org/Documents/Benefits11-07-06.pdf

Oswald, R. K.; Scherer, W. T.; Smith, B. L. 2000. *Traffic Flow Forecasting Using Approximate Nearest Neighbor Nonparametric Regression*. Report UVA-CE-ITS_01-4. Center for Transportation Studies, University of Virginia. 115 p. Available from Internet: http://ntl.bts.gov/lib/23000/23500/23528/paper-Scherer-TrafficForecasting-Non-parametric.pdf

Owens, N.; Armstrong, A.; Sullivan, P.; Mitchell, C.; Newton, D.; Brewster, R.; Trego, T. 2010. *Traffic Incident Management Handbook*. US Department of Transportation, Federal Highway Administration, Office of Transportation Operations. 116 p. Available from Internet: http://www.ops.fhwa.dot.gov/eto_tim_pse/publications/timhandbook/index.htm

Qi, Y.; Teng, H. 2008. An Information-based time sequential approach to online incident duration prediction, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 12(1): 1–12. http://dx.doi.org/10.1080/15472450701849626

Ratanamahatana, C. A.; Gunopulos, D. 2003. Feature selection for the naive Bayesian classifier using decision trees, *Applied Artificial Intelligence* 17(5–6): 475–487. http://dx.doi.org/10.1080/713827175

Schrank, D.; Eisele, B.; Lomax, T. 2012. *TTI's 2012: Urban Mobility Report*. Texas A&M Transportation Institute, The Texas A&M University System. 70 p. Available from Internet: http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/mobility-report-2012.pdf

Smith, K.; Smith, B. L. 2002. *Forecasting the Clearance Time of Freeway Accidents*. Smart Travel Lab Report No STL-2001-01. Center for Transportation Studies, University of Virginia. 91 p. Available from Internet: http://ntl.bts.gov/lib/23000/23500/23524/paper-Smtih-IncidentDurationForecasting.pdf

Tan, P.-N.; Steinbach, M.; Kumar, V. 2005. *Introduction to Data Mining*. Addison-Wesley. 769 p.

Valenti, G.; Lelli, M.; Cucina, D. 2010. A comparative study of models for the incident duration prediction, *European Transport Research Review* 2(2): 103–111. http://dx.doi.org/10.1007/s12544-010-0031-4

Wang, J. H.; Cong, H. Z.; Qiao, S. 2013. Estimating freeway incident duration using accelerated failure time modeling, *Safety Science* 54: 43–50. http://dx.doi.org/10.1016/j.ssci.2012.11.009

Wei, C.-H.; Lee, Y. 2007. Sequential forecast of incident duration using artificial neural network models, *Accident Analysis & Prevention* 39(5): 944–954. http://dx.doi.org/10.1016/j.aap.2006.12.017

Wen, Y.; Chen, S. Y.; Xiong, Q. Y.; Han, R. B.; Chen, S. Y. 2012. Traffic Incident duration prediction based on *k*-nearest neighbor, *Applied Mechanics and Materials* 253–255: 1675–1681. http://dx.doi.org/10.4028/www.scientific.net/AMM.253-255.1675

Wilson, D. R.; Martinez, T. R. 1997. Improved heterogeneous distance functions, *Journal of Artificial Intelligence Research* 6: 1–34. http://dx.doi.org/10.1613/jair.346

Wu, W.-W.; Chen, S.-Y.; Zheng, C.-J. 2011. Traffic incident duration prediction based on support vector regression, in *ICCTP 2011: Towards Sustainable Transportation Systems*, 14–17 August 2011, Nanjing, China, 2412–2421. http://dx.doi.org/10.1061/41186(421)241

Yang, B.; Zhang, X.; Sun, L. 2008. Traffic Incident duration prediction based on the Bayesian decision tree method, in *Proceedings of the First International Symposium on Transportation and Development Innovative Best Practices*, 24–26 April 2008, Beijing, China, 338–343. http://dx.doi.org/10.1061/40961(319)56

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): 1549–1557. http://dx.doi.org/10.1109/TITS.2011.2161634