Research Article

Traffic Incident Clearance Time Prediction and Influencing Factor Analysis Using Extreme Gradient Boosting Model

Jinjun Tang,¹ Lanlan Zheng,¹ Chunyang Han,¹ Fang Liu ,² and Jianming Cai ¹

¹Smart Transport Key Laboratory of Hunan Province, School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China

²School of Transportation Engineering, Changsha University of Science and Technology, Changsha 410205, China

Correspondence should be addressed to Fang Liu; rcliufang@163.com and Jianming Cai; jmcai@csu.edu.cn

Received 27 January 2020; Accepted 18 May 2020; Published 9 June 2020

Academic Editor: Feng Chen

Copyright © 2020 Jinjun Tang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Accurate prediction and reliable significant factor analysis of incident clearance time are two main objects of traffic incident management (TIM) system, as it could help to relieve traffic congestion caused by traffic incidents. This study applies the extreme gradient boosting machine algorithm (XGBoost) to predict incident clearance time on freeway and analyze the significant factors of clearance time. The XGBoost integrates the superiority of statistical and machine learning methods, which can flexibly deal with the nonlinear data in high-dimensional space and quantify the relative importance of the explanatory variables. The data collected from the Washington Incident Tracking System in 2011 are used in this research. To investigate the potential philosophy hidden in data, *K*-means is chosen to cluster the data into two clusters. The XGBoost is built for each cluster. Bayesian optimization is used to optimize the parameters of XGBoost, and the MAPE is considered as the predictive indicator to evaluate the prediction performance. A comparative study confirms that the XGBoost outperforms other models. In addition, response time, AADT (annual average daily traffic), incident type, and lane closure type are identified as the significant explanatory variables for clearance time.

1. Introduction

According to Lindley [1], traffic incidents result in about 60% of nonrecurrent traffic congestions. These congestions may cause lots of adverse effects such as reducing the roadway capacity, increasing the likelihood of secondary incidents [2], and unfavorable social and economic phenomenon [3]. When a traffic incident occurred, timely and reliable incident duration prediction plays an important role in the traffic authorities to design strategy for traffic guidance. In terms of Highway Capacity Manual, there are four phases in traffic incident duration [4]: detection time (the time from incident occurrence to detection), response time (the time from incident detection to verification), clearance time (the time from incident verification to clearance), and recovery time (the time from incident clearance to the normal traffic condition). Severe incidents that are not cleared in time may lead to a twice even three times incident duration [5]. Compared to other phases, clearance time is

the most important and time-consuming phase in the time incident process. Thus, the aims of this paper are to effectively predict the clearance time and investigate the significant influencing factors of clearance time.

Over the past few decades, a large number of works have been undertaken to predict the incident duration time. These approaches can be mainly categorized into statistical approaches and machine learning approaches. Statistical methods have their own model assumptions and predefined underlying relationships between dependent and independent variables [6] which provide the explainable ability to statistical methods. The widely used statistical methods are summarized as follows: probabilistic distribution analyses method [7, 8], regression method [9–13], discrete choice method [14], structure equation method [15], hazard-based duration method [16], Cox proportional hazards regression method [17–19], and accelerated failure time method [20–23]. Unlike statistical methods, machine learning methods are based on a more flexible mapping process that

requires no or less prior hypothesis. And flexible mapping allows machine learning methods to handle the nonlinear data in the high-dimensional space, but it cannot explore the potential relationship between dependent variables and independent variables. These widely used machine learning methods are categorized as K -nearest neighborhood method [24–27], support vector machine method [26–28], Bayesian networks method [29–34], artificial neural networks method [2, 35–37], genetic algorithm [37, 38], tree-based method [25, 39–41], and hybrid method [42].

In summary, conventional incident clearance time prediction studies rely on either statistical models with prior assumptions or machine learning models with poor interpretability [43]. To solve the abovementioned issues, we apply the extreme gradient boosting machine (XGBoost) method to predict the clearance time and then investigate the significant influencing factors of traffic incident clearance time. Because the XGBoost inherits both the advantages of statistical models and machine learning models, which can handle the nonlinear high-dimensional data when computing the relative importance among variables.

In this study, the prediction performance of XGBoost is examined by using the data from the Washington Incident Tracking System in 2011. In order to better explore the potential philosophy hidden in the original data, we cluster the original data in terms of their inherent properties. And then XGBoost model is built for each cluster. The framework of the proposed method is detailed in Section 3.5.

The remaining of this research is organized as follows. The data source is described in Section 2. Section 3 presents the K -means algorithm, the XGBoost algorithm, the Bayesian optimization algorithm, evaluation indicator, and the framework of the proposed method. The model results and discussion are outlined in Section 4. The last section is the conclusion.

2. Data Description

Traffic incident data were collected from the Washington Incident Tracking System (WITS), which occurred on the section from Boeing Access Road (Milepost 157) to the Seattle Central Business District (Milepost 165). This segment is not only a high incident-occurrence area but also takes on heavy traffic demand [44]. Therefore, it was chosen as the research object. And the annual average daily traffic (AADT) comes from the Highway Safety Information System (HSIS) database. The historical weather data were obtained from the National Oceanic and Atmospheric Administration (NOAA)'s weather stations in the region. The components of the data are detailed in Table 1. There are 14 discrete explanatory variables and 2 continuous explanatory variables in this dataset. In terms of their properties, they are divided into six categories: incident, temporal, geographical, environment, traffic, and operational. The detailed value sets of variables are presented as the third column in Table 1. In order to equalize the variability of independent variables, both response time and AADT variables are normalized [41, 43–46].

Totally, 2565 incident records were retrieved from the WITS database for the time period from 1 January to 31 December 2011. The mean and standard values of clearance time are, respectively, 13.10 minutes and 14.63 minutes. A big standard value (14.63 min) means that most of the clearance time values are quite different from their average values. That is, the original data should be processed to make the data organized well.

3. Methodology

3.1. K -Means Algorithm. K -means algorithm, developed by MacQueen [47], is one of the widely used methods in the field of dataset clustering. Samples in the dataset with similar characteristics can be clustered into the same class by using K -means [48]. The data we used in this research are expressed as $\{x_i = [x_{i1}, x_{i2}, \dots, x_{im}], y_i\}$, $i = 1, 2, 3, \dots, n$ and n represents the number of incidents, m is the number of explanatory variables, and the y denotes the actual clearance time. The detailed steps of the K -means algorithm are presented as follows:

Step 1: assuming the number of clusters (K clusters) and choosing the cluster centers from the dataset randomly.

Step 2: determining the clusters of other samples by the distance function as

$$x_i \in C_a, \quad \text{if } |x_i - O_a| < |x_i - O_b|. \quad (1)$$

Here, the O_a and O_b are the centers of the cluster a and cluster b , and C_a denotes the cluster a .

Step 3: after all samples have been clustered, the new center of each cluster should be calculated by using the following equation:

$$O_j = \frac{\sum_{i \in C_a} x_i}{N_C}, \quad j = 1, 2, 3, \dots, K, \quad (2)$$

where N_C is the number of the samples in cluster j .

Step 4: repeating step 2 and step 3 until the center of the cluster is within the permission.

Accordingly, we can find that the value of K and the cluster center are important to the clustering performance, as the clustering of K -means is extremely dependent on the selection of initial cluster center and the number of K . To obtain a reasonable K , we use the silhouette coefficient as the evaluation index, which is proposed by Rousseeuw [49] and defined as follows:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{if } a(i) < b(i), \\ 0, & \text{if } a(i) = b(i), \\ \frac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i). \end{cases} \quad (3)$$

TABLE 1: Description of explanatory variables for clearance time.

Category	Variable	Value set
	Response time	R+
Incident	Incident type	0 = others
		1 = disabled
		2 = debris
		3 = abandoned vehicle
		4 = collision
		0 = others
Incident	Lane closure type	1 = single lane
		2 = multiple lane
		3 = all travel lane
		4 = total lane
		0 = no; 1 = yes
		0 = no; 1 = yes
Temporal	Injury involved	0 = no; 1 = yes
	Fire involved	0 = no; 1 = yes
	Work zone involved	0 = no; 1 = yes
	Heavy truck involved	0 = no; 1 = yes
	Time of day	0 = daytime; 1 = night (22:00–6:00)
	Day of week	0 = weekdays; 1 = weekends
Temporal	Month of year	0 = other seasons
		1 = summer (Jun, Jul, and Aug)
Geographic	HOV	2 = winter (Dec, Jan, and Feb)
		0 = no; 1 = yes
Environment	Weather	0 = others
		1 = rainy
		2 = snowy
Traffic	Peak hours (6:00–9:00, 15:00–18:00)	0 = no; 1 = yes
	AADT	R+
Operational	Traffic control	0 = no; 1 = yes
	Washington State Patrol (WSP) involved	0 = no; 1 = yes

Here, the $a(i)$ is the average distance between sample i and other samples within the same cluster, and the $b(i)$ is the lowest average distance of sample i to all the remaining samples.

3.2. Extreme Gradient Boosting Machine Algorithm. Chen and Guestrin [50] proposed the extreme gradient boosting machine (XGBoost) algorithm. It is regarded as the advanced application of gradient boosting machine (GBDT) and adopts decision trees as the base learners for achieving classification and regression. Boosting is the integrated approach that can adjust the predicted error of the current model by adding new models to the model [41]. The predicted result of the boosting model is the sum scores of all models. Accordingly, the prediction of XGBoost is the sum scores of K boosted trees and is shown in the following equation:

$$\hat{y}_i = \varnothing(x_i) = \sum_{k=1}^K f_k(x_i), \quad f_k \in F, \quad (4)$$

where x_i is the i^{th} sample, $f_k(x_i)$ is the score of x_i at the i^{th} boosted tree, and F is the space composed of boosted trees. To decrease the fitting error of XGBoost, there is an improvement in regulation compared to GBDT, and it is presented as follows:

$$\text{obj}(\Theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad (5)$$

where y_i and $\hat{y}_i$ are the actual and predicted values of the i^{th} sample, the former item is the loss function, which needs to be a differentiable convex function, and the latter item is the penalty corresponding to the model complexity for avoiding overfitting. The second item of equation (5) can be detailed as follows:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2, \quad (6)$$

where both γ and λ are constants, T denotes the sum number of leaves, and w_j is the score of j^{th} leaf. When equation (6) equals zero, the $\text{obj}(\Theta)$ will convert to the conventional formula of GBDT.

According to equations (5) and (6), the training error and the model complexity are the two main sections of XGBoost. When the previous trees have been trained, the current tree can be trained by using additive training method. It means that when the t^{th} boosted tree is trained, the parameters of the previous trees (from the first tree to the $(t-1)^{\text{th}}$ tree) are fixed and their corresponding variables are constant. Taking the t^{th} boosted tree as an example, the loss can be expressed as follows:

$$\text{obj}(\Theta)^t = \sum_{i=1}^n l(y_i, \hat{y}_i^t) + \sum_{t=1}^T \Omega(f_t). \quad (7)$$

There are two formulas in these two items of (7):

$$\hat{y}_i^t = \hat{y}_i^{(t-1)} + f_t(x_i), \quad (8)$$

$$\sum_{t=1}^T \Omega(f_t) = \left(\sum_{k=1}^{T-1} \Omega(f_k) \right) + \Omega(f_t). \quad (9)$$

The first items of equations (8) and (9) are the sum score and sum regulation of former $(t-1)^{\text{th}}$ trees and the second items of them are the score and regulation of the t^{th} boosted tree, $\hat{y}_i^t$ is the predicted value of the t^{th} iteration, and $\sum_{t=1}^T \Omega(f_t)$ is the regulation of t^{th} iteration.

Equations (8) and (9) are substituted into equation (7), and then equation (7) is expanded by using the following Taylor formula:

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2. \quad (10)$$

The $\hat{y}_i^{(t-1)}$ is considered as x and the $f_t(x_i)$ is regarded as Δx . Then, equation (7) is transformed as follows:

$$\begin{aligned} \text{obj}(\Theta)^t &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{constant} \\ &\approx \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i (f_t(x_i))^2 \right] \\ &\quad + \Omega(f_t) + \text{constant} \\ &= \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i (f_t(x_i))^2 \right] + \Omega(f_t) + \text{constant}. \end{aligned} \quad (11)$$

As Chen and Guestrin [50] suggested, $f_t(x)$ can also be written as

$$f_t(x) = \omega_{q(x)}, \quad \omega \in R^K, q: R^d \longrightarrow \{1, 2, \dots, d\}, \quad (12)$$

where $q(x)$ is the leaf node of x , the $\omega_{q(x)}$ indicates the weight of $q(x)$ or that can be considered as the predicted value of the t^{th} iteration, and d is the number of leaf nodes. Then, equation (11) can be expressed as follows:

$$\begin{aligned} \text{obj}(\Theta)^t &= \sum_{i=1}^n \left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i (\omega_{q(x_i)})^2 \right] \gamma T \\ &\quad + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 + \text{constant} \\ &= \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] \\ &\quad + \gamma T + \text{constant}, \end{aligned} \quad (13)$$

where g_i and h_i are the first order and second order of gradient statistics. When the $q(x)$ is fixed, the optimal leaf weight and the metric function can be used to measure the quality of the tree structure $q(x)$ can be calculated:

$$\begin{aligned} w_j^* &= \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \\ \text{obj}(q) &= -\frac{1}{2} \sum_{j=1}^T \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \end{aligned} \quad (14)$$

3.3. Bayesian Optimization Algorithm. Bayesian optimization algorithm (BOA), one of the most famous extendible applications of the Bayesian network, is based on the construction of the probabilistic model. This model defines the distribution of objective function from the input data to output data. In this Bayesian optimization process, the global statistical characteristics are obtained from the optimal solutions and modeled by using the Bayesian network [51]. That is why the BOA shows its advantage in machine learning models because these machine learning models need more accurate parameters to flexibly handle nonlinear high-dimensional data [52]. In this study, the BOA is applied to optimize the parameters in the XGBoost with the aim to accurately predict the traffic incident clearance time.

The accomplishment of Bayesian optimization includes two core parts: prior function (PF) and acquisition function (AC), which is also called the utility function [51]. Gaussian process (GP) is generally considered as the PF. And the AC is used to balance the model exploration and exploitation. The framework of Bayesian optimization is presented in Figure 1 and the main steps are described as follows: (1) The data is split into training data and validation data by using the k -fold cross-validation method. Initialization parameters of the target model are defined as $\{\theta_1, \theta_2, \dots, \theta_n\}$. (2) The accuracy of the target model with initial parameters is evaluated by using validation data, and then the accuracy is recorded. The goal of the optimization is to minimize validation accuracy. (3) Gaussian process (GP) is employed to fit the recorded accuracy. (4) The parameters of the target model are updated in terms of the result of GP. Then, the maximum value of AC is used to select the next point, as it achieves the optimization by determining the next point to evaluate. Probability of improvement, expected improvement, and information gain are the three widely used AC [51]. In this study, expected improvement is chosen as the AC. Then, the best validation accuracy is mathematically written as follows:

$$\alpha(\theta, \text{GP}) = \int_{-\infty}^{\infty} \max(L - L^*, \text{GP}) P_{\text{GP}}(L | \theta) dL, \quad (15)$$

where L is the validation accuracy and $P_{\text{GP}}(L | \theta)$ is the probability of L with θ that is executed by using GP.

3.4. Evaluation Indicator. In general, the mean absolute percent error (MAPE) is a commonly used predictive indicator to evaluate the prediction performance of the regressive model. As mentioned above, the data are described as $\{x_i = [x_{i1}, x_{i2}, \dots, x_{im}], y_i\}$, $i = 1, 2, 3, \dots, n$, that can be considered as a matrix with the size of $n * m$. Specifically, n is

FIGURE 1: Parameter-tuning process of Bayesian optimization.

the number of incidents and y_i represents the actual value of the i^{th} incident. Considering p_i is the predicted value of the i^{th} incident. Then, the MAPE can be expressed as follows:

$$\text{MAPE} = \frac{\sum_{i=1}^n |(y_i - p_i)/y_i|}{n} \times 100\%. \quad (16)$$

In terms of this formula, the MAPE is a relative predictive indicator that can measure the prediction performance of the models based on actual values and predicted values.

3.5. Framework of the Proposed Method. As introduced in Section 2, we need a suitable way to handle the original dataset to organize the dataset well for exploring the potential philosophy hidden in data easier. To this end, in this research, we select the K -means algorithm as the method to cluster the original dataset into several categories in which the data are high similarity. Then, the XGBoost model is built for each category to perform prediction. The main steps of the proposed method are introduced as follows:

Step 1: clustering the original data into several categories by using the K -means algorithm. The number of clusters is determined by the optimal silhouette coefficient (the detailed information is introduced in Section 3.1).

Step 2: splitting the clustered data into training data and testing data for each category. Using the training data to constructs the XGBoost model.

Step 3: the BOA is used to optimize parameters for each constructed XGBoost model.

Step 4: inputting the testing data into the trained XGBoost, and then the predicted clearance time will be output and recorded.

Step 5: calculating the predictive indicator (MAPE) and the relative importance of explanatory factors

Noting that with the number of traffic incidents increasing, the dataset will be updated continuously, and thus the XGBoost should be retrained.

4. Prediction Result and Discussion

There are two objects of this study: (a) examining the performance of the XGBoost model in predicting clearance time and (b) investigating the significant factors of clearance time. We firstly process the original data, including data clustering, and clustering evaluation. Next, the data are split into training data and testing data with a ratio of 7:3. The XGBoost is trained by using training data, and the testing data are used for model evaluation. Then, comparison research examines the prediction performance of XGBoost. MAPE is chosen as a predictive measure. Finally, the relative importance of all the explanatory variables is calculated, and the significant explanatory variables of incident clearance time are analyzed. Overall, the proposed model is accomplished by coding and executing at Python.

4.1. Data Preprocessing. Before modeling, the original dataset has been processed by means of the K -means algorithm. As described in Section 3.1, the number of clusters (K) is the key parameter of the K -means algorithm. To find the best K , the values of K increasing from 2 to 10 are selected to calculate the corresponding silhouette coefficient, and the results are shown in Table 2. Assuming the iteration stops when the silhouette coefficients for continuous 5 iterations are not improved. The iteration stops when $K=7$, as the silhouette coefficients of continuous 5 iterations are decreasing. In terms of equation (3), a higher silhouette coefficient indicates a better clustering performance. According to Table 2, when $K=2$, the silhouette coefficient reaches the biggest value (0.613), which means K is set as 2 in this study. In this case, the original data are clustered into two clusters in this study. To present each cluster clearly, we draw the scatter plots of the target variable and one of the explanatory variables (which is chosen randomly), shown in Figure 2. The x -axis is clearance time and the y -axis denotes the response time. Figure 2(a) shows the scatter plot of these two variables in the original data, while Figure 2(b) shows the scatter plot of the clustered data. As shown in Figure 2(b), the cluster 1 marked with purple represents relative shorter clearance time, and cluster 2 marked with yellow indicates longer clearance time.

In order to knowledge the characteristic of two clusters clearly, several essential indexes are calculated and presented in Table 3. In total, there are 2246 incidents in cluster 1 and 319 incidents in cluster 2. Regarding cluster 1, the mean, standard, median, and range values of clearance time are 9 minutes, 5.44 minutes, 7.00 minutes, and 22 minutes. In respect to cluster 2, these values, respectively, are 39.25 minutes, 15.25 minutes, 35 minutes, and 75 minutes. Compared median value to mean value within each cluster, we can find that median values are, respectively, bigger than mean values for both two clusters. The result indicates that the distributions of clearance time in two clusters are skewed, instead of normal distribution. Then, we calculate the skew values of two clearance time distributions, and they are 0.92 in cluster 1 and 1.59 in cluster 2. Both of them present right-skewed, which are consistent with previous studies [26, 39, 41]. Distribution figures of clearance time in two clusters are shown in Figures 3(a) and 3(b).

Both Figures 3(a) and 3(b) present long-tail distributions with the range values of 22 and 75. It is difficult to handle the data with such a wide value range [53]. In this case, in order to make the distribution of clearance time closer to the normal distribution, we use data transformation to deal with clearance time data in two clusters. Regarding cluster 1, the skew value of clearance time is 0.92 which is between 0.5 and 1, indicating the median skewed. Therefore, according to the empirical method, we apply the square transformation to handle clearance time in cluster 1. In respect to cluster 2, the skewed value is 1.59 which is larger than 1, leading to a highly skewed. The log transformation is used to convert clearance time in cluster 2. Distributions of transformed clearance time are presented in Figures 3(c) and 3(d). In Figure 3, the blue line is the fitting curve of clustered data and the black line denotes the normal distribution curve

TABLE 2: Corresponding silhouette coefficient of each K .

K	2	3	4	5	6	7
Silhouette coefficient	0.613	0.447	0.422	0.418	0.396	0.352

which is fitted by their calculated mean and standard values. As shown in Figures 3(c) and 3(d), the distributions of transformed data are closer to normal distribution.

4.2. Parameter Optimization. In general, there are three approaches to optimize parameters, including the systematic grid search approach, the random search approach, and the Bayesian optimization approach. The grid search approach works well as it systematically searches the entire search space, but time-consuming. In contrast, the random search approach runs fast while it may miss the best value as it searches randomly in the search space. Bayesian optimization is the process of continuously sampling, calculating, and updating the model. In overall, we apply the Bayesian optimization method to find the optimal parameters in XGBoost. These parameters include max depth of the tree (max_depth), the number of trees ($n_estimators$), the learning rate of the tree (learning_rate), percent of randomly sampling for trees (subsample), sum of minimum leaf node sample weights (min_child_weight), and percentage of randomly sampled features (colsample_bytree). The increasing of $n_estimators$ may improve the accuracy of XGBoost but increase the computing time too. The max_depth is used to avoid overfitting. In contrast, the larger min_child_weight will result in underfitting. Both subsample and min_child_weight, respectively, denote the row and column sampling. The meaning of the learning rate is identified to avoid overfitting and increase the robustness of the model [54]. Therefore, all these parameters should be optimal for achieving the best model performance.

The Bayesian optimization is packaged in a module of python, called Hyperopt [55]. The objective function (f_{min}), search space (space), optimal algorithm (algo), and the maximum numbers of evaluations (max_evals) are four main objects of the Hyperopt, which is used to accomplish BOA. In this research, the XGBoost is the f_{min} , tree of Parzen estimator defaults as the algo, and the max_evals is generally set as 4. Regarding search space, we set $n_estimators \in [50, 500]$, learning_rate $\in [0.05, 0.1]$, max_depth $\in [2, 10]$, subsample $\in [0.1, 0.9]$, colsample_bytree $\in [0.1, 0.9]$, and min_child_weight $\in [2, 12]$. In addition, we use 5-fold cross-validation during parameter tuning, and the result is shown in Table 4.

Regarding cluster 1, the $n_estimators$, learning_rate, max_depth, subsample, colsample_bytree, and min_child_weight are, respectively, set as 140, 0.09, 6, 0.5, 0.7, and 3. In respect to cluster 2, the best prediction performance of XGBoost is obtained when the $n_estimators=100$, the learning_rate=0.05, the max_depth=5, the subsample=0.5, the colsample_bytree=0.3, and the min_child_weight=5. The XGBoost model reaches its best prediction performance when using these optimal parameters. And the MAPE values of optimized XGBoost for two clusters are 0.348 and 0.221, respectively.

FIGURE 2: Scatter plots of data. (a) Scatter plot of the original dataset. (b) Scatter plot of the clustered dataset.

TABLE 3: Statistical characteristics of clearance time.

Cluster	1	2
Count	2246.00	319.00
Mean	9.00	39.25
Standard	5.44	15.25
Min	3.00	21.00
25%	5.00	29.00
Median	7.00	35.00
75%	12.00	45.00
Max	25.00	96.00
Skew	0.92	1.59
Range	22.00	75.00

4.3. Comparison Analysis. To examine the prediction performance of XGBoost in clearance time prediction, we select several commonly used models including support vector regression (SVR) model, random forest (RF) model, and Adaboost model for comparison. To ensure fair comparison, the testing data and the parameter-tuning method (BOA) of all models are the same. For the SVR model, we select the radial basis function (RBF) as the kernel function. The gamma and penalty C are two key parameters of RBF and are set as 0.1, 64, and 0.15, 32 for two clusters. For the RF model, the number of trees ($n_{\text{estimators}}$), the maximum depth of the tree (max_depth), the minimum number of samples of internal node splitting (min_samples_split), and the minimum number of leaf nodes (min_samples_leaf) are the four key parameters, and they are set as 195, 8, 11, and 23 in the cluster 1 and 100, 13, 18, and 12 in the cluster 2. In regard to the Adaboost model, the same with RF model, $n_{\text{estimators}}$, max_depth , and min_samples_split should be identified. In addition, the learning_rate and the maximum features in splitting (max_features) also need to be optimized. These parameters of Adaboost in two clusters are set as 470, 6, 25, 0.05, 7 and 425, 9, 30, 0.11. The MAPE for four candidates is shown in Table 5, and the smallest values for two clusters are marked in bold.

As shown in Table 5, for cluster 1, the MAPE values of XGBoost, SVR, RF, and Adaboost are 0.348, 0.363, 0.357, and 0.383. The XGBoost represents the smallest MAPE, showing its superiority in clearance time prediction for cluster 1. As for cluster 2, the MAPE values of XGBoost, SVR, RF, and Adaboost are 0.221, 0.253, 0.228, and 0.231. Compared to other models, the XGBoost represents the smallest MAPE (0.221). It means the XGBoost model outperforms SVR, RF, and Adaboost in both two clusters. This result confirms the superiority of XGBoost in clearance time prediction.

4.4. Importance Evaluation for Explanatory Factors. Different explanatory variables have different effects on the target factor [56, 57]. To investigate the significant factors of clearance time, the relative importance of each explanatory factor is calculated by using the XGBoost with optimal parameters for two clusters. An explanatory factor with higher relative importance means that it generates a stronger effect on clearance time [41]. In this study, we assume that factors with relative importance greater than 8.0% are defined as significant explanatory factors, the relative importance of the general factor is from 2.5% to 8.0%, and the remaining explanatory factors are considered as insignificant factors. In this case, the explanatory factors with its importance are shown in Table 6.

As for cluster 1, AADT (17.70%), incident type (17.30%), response time (15.10%), and lane closure type (8.00%) are categorized into the significant explanatory factors of clearance time as their relative importance is bigger than 8.00%. The general factors of clearance time include six explanatory factors, such as WSP involved (7.60%), month of year (6.10%), traffic control (5.00%), weather (4.70%), day of week (4.60%), and peak hours (3.10%). And the remaining HOV (2.50%), time of day (2.10%), heavy truck involved (1.70%), injury involved (1.70%), and work zone involved

FIGURE 3: Distributions of clearance time. (a) Original clearance time distribution of cluster 1. (b) Original clearance time distribution of cluster 2. (c) Log-transformed clearance time distribution of cluster 1. (d) Square-transformed clearance time distribution of cluster 2.

TABLE 4: The optimal parameters in XGBoost.

Cluster	1	2
$n_estimators$	140	100
$learning_rate$	0.09	0.05
max_depth	6	5
$subsample$	0.5	0.5
$colsample_bytree$	0.7	0.3
min_child_weight	3	5

(0.30%) are regarded as the insignificant explanatory variables in cluster 1. Regarding cluster 2, four explanatory factors are included in significant explanatory factors to clearance time, including AADT (14.00%), incident type (12.8%), response time (22.30%), and lane closure type (8.40%). And fire involved (8.40%), weather (6.10%), month of year (6.10%), traffic control (6.10%), injury involved (5.00%), and HOV (2.80%) are the general explanatory factors. Peak hours (2.20%), heavy truck involved (2.20%),

TABLE 5: Prediction results for different models.

Cluster	XGBoost	SVR	RF	Adaboost
1	0.348	0.363	0.357	0.383
2	0.221	0.253	0.228	0.231

WSP involved (1.70%), day of week (1.10%), time of day (0.60%), and work zone involved (0.20%) are categorized into insignificant explanatory factors to incident clearance time.

That is, for both two clusters, AADT, incident type, response time, and lane closure type are considered as the significant explanatory factors of clearance time. But the same factor may generate varying impacts on clearance time in the different datasets [58]. In detail, the AADT is the greatest contribution to shorter clearance time in cluster 1 and generates the second impacts on longer clearance time in cluster 2 with the relative importance of 17.70% and 14.00%, respectively. Generally speaking, AADT represents

TABLE 6: Relative importance of explanatory factors on clearance time.

Cluster		1	2
Rank	Variable	Relative importance (%)	Relative importance (%)
Significant explanatory factors	1 AADT	17.70	Response time 22.30
	2 Incident type	17.30	AADT 14.00
	3 Response time	15.10	Incident type 12.80
	4 Lane closure type	8.00	Lane closure type 8.40
General explanatory factors	5 WSP involved	7.60	Fire involved 8.40
	6 Month of year	6.10	Weather 6.10
	7 Traffic control	5.00	Month of year 6.10
	8 Weather	4.70	Traffic control 6.10
	9 Day of week	4.60	Injury involved 5.00
	10 Peak hours	3.10	HOV 2.80
Insignificant explanatory variables	11 HOV	2.50	Peak hours 2.20
	12 Fire involved	2.50	Heavy truck involved 2.20
	13 Time of day	2.10	WSP involved 1.70
	14 Heavy truck involved	1.70	Day of week 1.10
	15 Injury involved	1.70	Time of day 0.60
	16 Work zone involved	0.30	Work zone involved 0.20

the characteristic of current traffic [59, 60]. That is, the traffic congestion with a high AADT may make the incident difficult to clear, leading to longer clearance time. As for incident type, it respectively contributes 17.30% and 12.80% to short and long clearance time and ranks the second in cluster 1 and the third in cluster 2. As shown in Table 1, the incident type factor consists of disabled vehicles, debris, abandoned vehicles, collision, and others. These incidents may block normal traffic [61, 62]. In this case, the transportation authorities may make a series of strategies to deal with the problems caused by these incidents [63, 64]. Interestingly, the longer clearance time seems less sensitive to incident type than shorter clearance time. Maybe a long clearing time means a high severity of the crash. With the relative importance of 15.10% and 22.3%, the response time factor is the third contributor for shorter clearance time in cluster 1 and yields the biggest impacts on longer clearance time in cluster 2. The result shows that longer clearance time is more sensitive to response time compared to shorter clearance time, which is consistent with the previous studies [18, 19]. For every minute, the response time increases, and the clearing time will increase by one percent [18, 19]. The lane closure type factor is the fourth contributed factor for both two clusters. It indicates the severity of incidents by restricting vehicles from entering the incident site [41].

5. Conclusions

In this study, XGBoost is applied to predict incident clearance time that occurred on the freeway and investigates the significant factors of clearance time by using the data collected from the Washington Incident Tracking System in 2011. We firstly introduce the original data and the proposed method briefly. The original data are clustered by using the *K*-means algorithm for better exploring the underlying relationship. Then, we built the XGBoost model for each cluster. Each clustered data is divided into 70% training data and 30% testing data. Training data are applied for modeling XGBoost and optimizing parameters on the basis of 5-fold

cross-validation BOA. Testing data are used to measure the prediction performance of XGBoost. And the MAPE is considered as the predictive indicators in this paper. To examine the model performance of XGBoost, support vector regression (SVR), random forest (RF), and Adaboost are chosen to predict the clearance time. The comparing study manifests that the XGBoost outperforms the other three models with the lowest MAPE in both two clusters. To obtain the significant factors of clearance time, we calculate the relative importance of each explanatory factor and then define the quantitative indexes about significant explanatory factors, general explanatory factors, and insignificant explanatory factors. The result is that response time, AADT, incident type, and lane closure type are the significant explanatory factors of clearance time.

It is worth noting that the traffic incident is the time-sequential process [65]. And almost the incident information is acquired from that process [66]. Modeling based on the acquired incident information is the limitation of the proposed method in this study. Because, during the initial stage of the incident, the prediction may not be accurate due to the acquired information is incomplete. For future research, multistage updates of information should be a promising future research direction. In addition, strategies about dealing with the unobserved heterogeneity of dependent variables, especially in traffic incidents filed, may be a hot topic, due to some omitted variables (e.g., driving behavior) that may generate potential impacts on the target variable.

Data Availability

The traffic incident data used to support the findings of this study are available from the corresponding author and first author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (71701215), Innovation-Driven Project of Central South University (no. 2020CX041), Foundation of Central South University (no. 502045002), Science and Innovation Foundation of the Transportation Department in Hunan Province (no. 201725), and Postdoctoral Science Foundation of China (nos. 2018M630914 and 2019T120716).

References

- [1] J. A. Lindley, "Urban freeway congestion: quantification of the problem and effectiveness of potential solutions," *IET Intelligent Transport Systems*, vol. 57, no. 1, pp. 27–32, 1987.
- [2] E. I. Vlahogianni, M. G. Karlaftis, and F. P. Orfanou, "Modeling the effects of weather and traffic on the risk of secondary incidents," *Journal of Intelligent Transportation Systems*, vol. 16, no. 3, pp. 109–117, 2012.
- [3] M. W. Adler, J. V. Ommen, and P. Rietveld, "Road congestion and incident duration," *Economics of Transportation*, vol. 2, no. 4, pp. 109–118, 2013.
- [4] H. C. Manual, *Highway Capacity Manual*, National Research Council, Washington, DC, USA, 2000.
- [5] S. Madanat and A. Feroze, "Prediction models for incident clearance time for borman expressway," Final Report FHWA/IN/JHRP-96/10, Purdue University, West Lafayette, IN, USA, 1997.
- [6] L.-Y. Chang and H.-W. Wang, "Analysis of traffic injury severity: an application of non-parametric classification tree techniques," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 1019–1027, 2006.
- [7] T. F. Golob, W. W. Recker, and J. D. Leonard, "An analysis of the severity and incident duration of truck-involved freeway accidents," *Accident Analysis & Prevention*, vol. 19, no. 5, pp. 375–395, 1987.
- [8] G. Giuliano, "Incident characteristics, frequency, and duration on a high volume urban freeway," *Transportation Research Part A: General*, vol. 23, no. 5, pp. 387–396, 1989.
- [9] A. J. Khattak, J. L. Schofer, and M.-H. Wang, "A simple time sequential procedure for predicting freeway incident duration," *IVHS Journal*, vol. 2, no. 2, pp. 113–138, 1995.
- [10] A. Khattak, X. Wang, and H. Zhang, "Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays," *IET Intelligent Transport Systems*, vol. 6, no. 2, pp. 204–214, 2012.
- [11] A. J. Khattak, J. Liu, B. Wali, X. Li, and M. Ng, "Modeling traffic incident duration using quantile regression," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2554, no. 1, pp. 139–148, 2016.
- [12] A. Garib, A. E. Radwan, and H. Al-Deek, "Estimating magnitude and duration of incident delays," *Journal of Transportation Engineering*, vol. 123, no. 6, pp. 459–466, 1997.
- [13] S. Peeta, J. L. Ramos, and S. Gedela, "Providing real-time traffic advisory and route guidance to manage borman incidents on-line using the hoosier helper program. Joint transportation research program," FHWA/IN/JTRP-2000/15, Indiana Department of Transportation and Purdue University, West Lafayette, IN, USA, 2000.
- [14] P. W. Lin, N. Zou, and G. L. Chang, "Integration of a discrete choice model and a rule-based system for estimation of incident duration: a case study in Maryland," in *Proceedings of the CD-ROM of Proceedings of the 83rd TRB Annual Meeting*, Washington, DC, USA, 2004.
- [15] J. Y. Lee, J. H. Chung, and B. Son, "Incident clearance time analysis for Korean freeways using structural equation model," *Proceedings of the Eastern Asia Society for Transportation Studies*, vol. 7, pp. 1850–1863, 2010.
- [16] N. E. Breslow, "Analysis of survival data under the proportional hazards model," *International Statistical Review/Revue Internationale de Statistique*, vol. 43, no. 1, pp. 45–57, 1975.
- [17] D. S. Bennett, "Parametric models, duration dependence, and time-varying data revisited," *American Journal of Political Science*, vol. 43, no. 1, pp. 256–270, 1999.
- [18] J.-T. Lee and J. Fazio, "Influential factors in freeway crash response and clearance times by emergency management services in peak periods," *Traffic Injury Prevention*, vol. 6, no. 4, pp. 331–339, 2005.
- [19] L. Hou, Y. Lao, Y. Wang et al., "Time-varying effects of influential factors on incident clearance time using a non-proportional hazard-based model," *Transportation Research Part A: Policy and Practice*, vol. 63, pp. 2–12, 2014.
- [20] D. Nam and F. Mannering, "An exploratory hazard-based analysis of highway incident duration," *Transportation Research Part A: Policy and Practice*, vol. 34, no. 1, pp. 85–102, 2000.
- [21] A. Stathopoulos and M. G. Karlaftis, "Modeling duration of urban traffic congestion," *Journal of Transportation Engineering*, vol. 128, no. 6, pp. 587–590, 2002.
- [22] A. T. Hojati, L. Ferreira, S. Washington, and P. Charles, "Hazard based models for freeway traffic incident duration," *Accident Analysis & Prevention*, vol. 52, pp. 171–181, 2013.
- [23] R. Li and P. Shang, "Incident duration modeling using flexible parametric hazard-based models," *Computational Intelligence and Neuroscience*, vol. 2014, Article ID 723427, 10 pages, 2014.
- [24] H. J. Kim and H.-K. Choi, "A comparative analysis of incident service time on urban freeways," *IATSS Research*, vol. 25, no. 1, pp. 62–72, 2001.
- [25] K. W. Smith and B. L. Smith, "Forecasting the clearance time of free-way accidents," Technical Report STL-2001-01, Center for Transportation Studies, University of Virginia, Charlottesville, VA, USA, 2014.
- [26] G. Valenti, M. Lelli, and D. Cucina, "A comparative study of models for the incident duration prediction," *European Transport Research Review*, vol. 2, no. 2, pp. 103–111, 2010.
- [27] Y. Wen, S. Y. Chen, Q. Y. Xiong, R. B. Han, and S. Y. Chen, "Traffic incident duration prediction based on K-nearest neighbor," *Applied Mechanics and Materials*, vol. 253–255, pp. 1675–1681, 2012.
- [28] W. W. Wu, S. Y. Chen, and C. J. Zheng, "Traffic incident duration prediction based on support vector regression," in *Proceedings of the ICCTP*, pp. 2412–2421, Nanjing, China, August 2011.
- [29] S. Boyles, D. Fajardo, and S. T. Waller, "A Naive Bayesian classifier for incident duration prediction," in *Proceedings of the TRB 86th Annual Meeting Compendium of Papers CD-ROM*, Washington DC, USA, 2007.
- [30] K. Ozbay and N. Noyan, "Estimation of incident clearance times using Bayesian networks approach," *Accident Analysis & Prevention*, vol. 38, no. 3, pp. 542–555, 2006.
- [31] H. Park, A. Haghani, and X. Zhang, "Interpretation of Bayesian neural networks for predicting the duration of detected incidents," *Journal of Intelligent Transportation Systems*, vol. 20, no. 4, pp. 385–400, 2015.
- [32] C. Chen, G. Zhang, R. Tarefder, J. Ma, H. Wei, and H. Guan, "A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes," *Accident Analysis & Prevention*, vol. 80, pp. 76–88, 2015.

- [33] C. Chen, G. Zhang, Z. Tian, S. M. Bogus, and Y. Yang, "Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations," *Accident Analysis & Prevention*, vol. 85, pp. 186–198, 2015.
- [34] F. Zong, X. Chen, J. Tang, P. Yu, and T. Wu, "Analyzing traffic crash severity with combination of information entropy and bayesian network," *IEEE Access*, vol. 7, pp. 63288–63302, 2019.
- [35] E. I. Vlahogianni and M. G. Karlaftis, "Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties," *Computer-Aided Civil and Infrastructure Engineering*, vol. 28, no. 6, pp. 420–433, 2013.
- [36] C.-H. Wei and Y. Lee, "Sequential forecast of incident duration using artificial neural network models," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 944–954, 2007.
- [37] C. X. Ma, W. Hao, F. Q. Pan, and W. Xiang, "Road screening and distribution route multi-objective robust optimization for hazardous materials based on neural network and genetic algorithm," *PLoS One*, vol. 13, no. 6, Article ID e0198931, 2018.
- [38] Y. Lee and C.-H. Wei, "A computerized feature selection method using genetic algorithms to forecast freeway accident duration times," *Computer-Aided Civil and Infrastructure Engineering*, vol. 25, no. 2, pp. 132–148, 2010.
- [39] C. Zhan, A. Gan, and M. Hadi, "Prediction of lane clearance time of freeway incidents using the M5P tree algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1549–1557, 2011.
- [40] Q. He, Y. Kamarianakis, K. Jintanakul, and L. Wynter, "Incident duration prediction with hybrid tree-based quantile regression," *Complex Networks and Dynamic Systems*, vol. 2, pp. 287–305, 2013.
- [41] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang, "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2303–2310, 2017.
- [42] W. Kim and G.-L. Chang, "Development of a hybrid prediction model for freeway incident duration: a case study in Maryland," *International Journal of Intelligent Transportation Systems Research*, vol. 10, no. 1, pp. 22–33, 2012.
- [43] J. J. Tang, L. L. Zheng, C. Y. Han et al., "Statistical and machine-learning methods for clearance time prediction of road incidents: a methodology review," *Analytic Methods in Accident Research*, vol. 27, Article ID 100123, 2020.
- [44] Y. J. Zou, J. J. Tang, L. T. Wu, K. Henrickson, and Y. H. Wang, "Quantile analysis of freeway incident clearance time," *Proceedings of the Institution of Civil Engineers-Transport*, vol. 170, no. 5, pp. 296–304, 2017.
- [45] Y. J. Zou, X. Z. Zhong, J. J. Tang et al., "A copula-based approach for accommodating the underreporting effect in wildlife-vehicle crash analysis," *Sustainability*, vol. 11, no. 2, pp. 1–13, 2019.
- [46] Y. Zou, X. Ye, K. Henrickson, J. Tang, and Y. Wang, "Jointly analyzing freeway traffic incident clearance and response time using a copula-based approach," *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 171–182, 2018.
- [47] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–296, 1967.
- [48] Y. Wang, K. Assogba, Y. Liu, X. Ma, M. Xu, and Y. Wang, "Two-echelon location-routing optimization with time windows based on customer clustering," *Expert Systems with Applications*, vol. 104, no. 104, pp. 244–260, 2018.
- [49] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [50] T. Q. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, CA, USA, 2016.
- [51] Q. Shang, D. Tan, S. Gao, and L. L. Feng, "A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis," *Journal of Advanced Transportation*, vol. 2019, Article ID 4202735, 11 pages, 2019.
- [52] E. Brochu, V. M. Cora, and N. D. Freitas, "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," Technical Report UBC TR-2009-23, Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, 2009.
- [53] S. Wang, R. Li, and M. Guo, "Application of nonparametric regression in predicting traffic incident duration," *Transport*, vol. 33, no. 1, pp. 22–31, 2018.
- [54] J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, "Crash injury severity analysis using a two-layer Stacking framework," *Accident Analysis & Prevention*, vol. 122, pp. 226–238, 2019.
- [55] J. Bergstra, B. Komer, D. Yamins, C. Eliasmith, and D. Cox, "Hyperopt: a Python library for model selection and hyperparameter optimization," *Computational Science & Discovery*, vol. 8, no. 1, Article ID 014008, 2015.
- [56] X. X. Ma, S. R. Chen, and F. Chen, "Correlated random-effects bivariate Poisson lognormal model to study single-vehicle and multivehicle crashes," *Journal of Transportation Engineering-ASCE*, vol. 142, no. 11, 2016.
- [57] Y. Yan, Y. Zhang, X. Yang, J. Hu, J. Tang, and Z. Guo, "Crash prediction based on random effect negative binomial model considering data heterogeneity," *Physica A: Statistical Mechanics and Its Applications*, vol. 547, Article ID 123858, 2020.
- [58] F. Chen, X. X. Ma, S. R. Chen, and L. Yang, "Crash frequency analysis using hurdle models with random effects considering short-term panel data," *International Journal of Environmental Research and Public Health*, vol. 13, no. 11, p. 1043, 2016.
- [59] Y. Wang, K. Assogba, J. Fan, M. Xu, Y. Liu, and H. Wang, "Multi-depot green vehicle routing problem with shared transportation resource: integration of time-dependent speed and piecewise penalty cost," *Journal of Cleaner Production*, vol. 2019, no. 232, pp. 12–29, 2019.
- [60] C. X. Ma, J. B. Zhou, X. C. Xu, F. Q. Pan, and J. Xu, "Fleet scheduling optimization of hazardous materials transportation: a literature review," *Journal of Advanced Transportation*, vol. 2020, Article ID 5070347, 16 pages, 2020.
- [61] C. X. Ma, W. Hao, W. Xiang, and W. Yan, "The impact of aggressive driving behavior on driver injury severity at highway-rail grade crossings accidents," *Journal of Advanced Transportation*, vol. 2018, Article ID 9841498, 10 pages, 2018.
- [62] C. X. Ma, D. Yang, J. B. Zhou, Z. X. Feng, and Q. Yuan, "Risk riding behaviors of urban E-bikes: a literature review," *International Journal of Environmental Research and Public Health*, vol. 16, no. 13, Article ID 2308, 2019.
- [63] Y. Yan, Y. Dai, X. Li, J. Tang, and Z. Guo, "Driving risk assessment using driving behavior data under continuous tunnel environment," *Traffic Injury Prevention*, vol. 20, no. 8, pp. 807–812, 2019.

- [64] C. Ding, X. Ma, Y. Wang, and Y. Wang, "Exploring the influential factors in incident clearance time: disentangling causation from self-selection bias," *Accident Analysis & Prevention*, vol. 85, pp. 58–65, 2015.
- [65] F. L. Mannering and C. R. Bhat, "Analytic methods in accident research: methodological frontier and future directions," *Analytic Methods in Accident Research*, vol. 1, pp. 1–22, 2014.
- [66] Y.-S. Chung, Y.-C. Chiou, and C.-H. Lin, "Simultaneous equation modeling of freeway accident duration and lanes blocked," *Analytic Methods in Accident Research*, vol. 7, pp. 16–28, 2015.