Clearance time prediction of traffic accidents: A case study in Shandong, China

Anyi Zhang[1],
Fanyu Meng[2],
Wenwu Gong[1],
Yiping Zeng[1],
Lili Yang[1],
Diping Yuan[3]

[1] Department of Statistics and Data Science, Southern university of Science and Technology, Shenzhen, China.

[2] Academy for Advanced Interdisciplinary Studies & Department of Statistics and Data Science, Southern university of Science and Technology, Shenzhen, China.

[3] Shenzhen Urban Public Safety and Technology Institute, Shenzhen, China.

Author correspondence:
Lili Yang,
Department of Statistics and Data Science,
Southern University of Science and Technology,
Shenzhen,
China.
Email: yangll@sustech.edu.cn

URL: http://trauma.massey.ac.nz/issues/2022-IS/AJDTS_26_IS_Zhang.pdf

Abstract

Accurate predictions of the clearance time of highway accidents can help make more effective decisions and reduce the economic losses caused by the accidents. This paper compares two representations of traffic accidents with mixed vehicle types and establishes two different classification models. The traffic accident data in Shandong Province, China from 2016 to 2019 are used as a case study. The interpretability of the parametric model indicates that the types of vehicles involved in the accident, the type of accident, and the weather can significantly affect the clearance time of the accident. The results of this study can not only provide evidence of whether the types of vehicles involved in the accident will affect the accident clearance time, but also provide advice for the authorities to quickly clear accident scenes and prevent further accidents.

Keywords: *Highway, Accident clearance time, Vehicle types, Passenger vehicles*

Introduction

As of the end of 2019, the total mileage of highways in China reached 149,600 km (Statistical Bulletin, 2020). With the highway network growing and the number of motor vehicles increasing rapidly, traffic accidents are occurring more frequently. While expressways promote economic development, they are more prone to major traffic accidents than urban roads due to their large traffic volumes and fast speeds, causing a large number of casualties and huge property losses every year (Lin et al, 2016; Park & Haghani, 2016). In the literature, a lot of research has been carried out to improve the efficiency of expressway traffic safety.

The accident duration prediction can be used to predict the clearance time of a certain accident. At the same time, real-time event duration prediction can help event managers determine the best emergency rescue and traffic control strategies (Ji et al., 2008). In addition, based on the prediction of accident impacts traffic managers can provide the drivers with guidance information so that the drivers dynamically correct their routes to reach the destination in the shortest time (Baykal-Gürsoy et al., 2009; Schrank et al, 2015). Thus effectively mitigating the traffic congestion and improving the level of accident management.

Traffic accident management is of great importance to transportation organizations. Delays caused by traffic accidents directly increase the possibility of secondary accidents, leading to more serious traffic congestions (Chung et al., 2015; Mannering et al., 2014; Meng et al., 2020). For every minute the primary accident remains on the highway, the average risk of the second collision will increase (Cassandra et al., 2012). Accurately prediction of the clearance time of the accidents can facilitate the decision making of the transportation management department and reduce the adverse effects caused by the traffic accidents.

In the past few decades, various statistical models have been applied to model and predict the clearance time of highway accidents. Regression models are typical methods used for estimating and predicting the incident duration for highway incidents (Garib et al., 1997; Valenti et al., 2010). Hazard-based duration models were also widely used to model and predict the accidents duration.

Separate hazard-based duration models were developed by Nam and Mannering to analyze detection/reporting time, response time, and clearance time of highway accidents duration (Nam & Mannering, 2000). AFT models and topic modeling was also applied to predict accident duration, but due to the limitations of the topic model, they did not study the impact of every single variable (Ruimin et al., 2015).

Artificial intelligence-based methods were also adopted to capture the relationship between accidents duration and its influential factors. The K-Nearest Neighbor and artificial neural network model are typical methods to model the clearance time of highway accidents (Wei & Lee, 2007; Wen et al., 2012). Lin et al. (2016) proposed an improved model based on M5P. They replaced linear regression of each leaf by HBDM algorithm and compared the traditional M5P model, HBDM algorithm, and the proposed M5P-HBDM. The results showed that M5P-HBDM could identify more important and meaningful variables. Recently, a complex network algorithm, which combines the modularity-optimizing community detection algorithm and the association rules learning algorithm, was proposed to identify the factors that affect highway accidents clearance time (Lin et al., 2014).

Previous studies (Li et al., 2017; Ding et al., 2015) have identified various factors that influence the incident clearance time, including incident characteristics(e.g., number of vehicles involved in an incident, truck/taxi/bus involvement); weather conditions(e.g., rain, fog, and/or snow); temporal factors (e.g., time of day, day of the week, and/or season); traffic characteristics (e.g., traffic volume) and some other factors. In particular, Crashes that occur in rainy or foggy weather are more likely to have long accident clearance times. And accidents with hazardous material dumping may take longer to clear (Nam & Mannering, 2000). When a large vehicle is involved in a crash, such as large trucks or buses, the accident clearance time may be longer (Chung, 2010). Traffic flow and upstream and downstream speeds around the accident location can also affect accident clearance times (Lee et al., 2010).

As expected, incidents involving chemical spills, hazardous materials and large vehicles have longer clearance time. In addition, incidents during congested periods typically take longer to clear (Hou, 2013). While the way these factors affect incident clearance times is consistent across multiple studies, some studies have found the opposite to be true. For example, while incident clearance during peak hours typically lasts longer, Hojati et al. (2013) observed shorter clearance times for incidents that occurred during the afternoon peak hour.

In the previous studies of clearance time prediction, researchers have considered the number of vehicles involved in the accident, whether there were large vehicles (Xia, 2016), and the number of heavy vehicles involved in the accident (Xu et al., 2013). Few studies have considered different types of vehicles involved in the accident to better represent their impact to the clearance time of the accidents. Specifically, for the same number of vehicles, if the types of the vehicles are different, the clearance time of the accident may be different.

This paper considers two different representations of vehicle types and measures their impact on predicting the clearance time of traffic accidents. A case study using the data obtained from the highways of Shandong province, China is presented. We introduced the data and its preprocessing firstly. When processing data, we delete the data with the accident clearance time exceeding 400 minutes. Because the reasons for the excessively long accident clearance time are single and there is the possibility of erroneous records. Then two models, the generalized linear model and the mixed-effects model, are compared for the task of predicting the clearance time of traffic accidents. The impact of different types of vehicles involved in the accident on the clearance time is also analyzed, based on which policy suggestions are provided to improve traffic accident management.

Data Collection and Analysis

Accident Database and Variable Definition

The accident data used in this article comes from four expressways including G2, G25, G35, and G1511 in Shandong Province between 2016 and 2019. These four expressways are important arterial roads in Shandong Province. In the four years from 2016 to 2019, there were a total of 4,255 accidents. The dataset contains information about the location, the weather, the time, and the vehicles involved in the accident. Based on this information, we carry out two preprocessing operations. First, we group the time of the accident into four time periods to generate four new variables, namely time of day1, time of day2, time of day3 and time of day4. Second, all the categorical variables are converted into dummy variables as shown in Table 1.

There are 29 variables in the extracted data set. In the latter two models, we use dummy variables to define the types of vehicles involved in the accident, so there are 34

variables before the screening variables. The dependent variable is binary, with a value indicating, "long" or "short", in which "long" refers to the clearance time of the accidents is longer than 120 minutes. The 2009 edition of the Manual on Uniform Traffic Control Devices defines accidents with clearance duration longer than 120 min as large-scale traffic accidents (Zhang et al., 2012). Large-scale traffic accidents can cause more serious congestion and economic losses.

There are 28 independent variables, of which 6 are continuous variables and 22 are dummy variables. To alleviate the influence of diverse value ranges, we divide the variables medium truck flow, large truck flow and embedding congestion, respectively by 1,000, 10,000, or 100,000 such that their value ranges are aligned to a range from 0 to 10. The dataset used by the mixed-effects model is slightly different. The categorical variable location is added (location contains information on 15 different areas on the four highways), and the weather variable is transformed from four dummy variables to one categorical variable. The variable are shown in Table 1.

The passenger car unit (PCU) is calculated at a later stage. According to the traffic volume survey vehicle classification and vehicle conversion coefficient, different weights are assigned to different vehicle types, as shown in Table 2. Then we calculate the weighted sum of vehicles involved in each accident, as shown in Table 3, which represents the number of vehicles involved in the corresponding accident.

Data Preprocessing

There are 332 records that contain missing information, and they are excluded from this study. As a result, there are 3923 remaining data. It is known from the accident description that most of the accidents with long duration are difficult to be clear in a short time, such as the spontaneous combustion of the truck, collision or rollover of the loaded truck,

Table 1.
Variable definition

Variable name	Type	Description
minivan flow	Continuous	minivan flow/10000
medium truck flow	Continuous	medium truck flow/1000
extra large truck flow	Continuous	flow of extra large trucks/10000
container truck flow	Continuous	container flow/1000
embedding congestion	Continuous	The ratio of the total traffic volume of the road network to the total capacity allowed by the road network
time of day1	Dummy	The accident happened during 6:00~10:00:1; other time :0
time of day2	Dummy	The accident happened during 10:00~16:00:1; other time :0
time of day3	Dummy	The accident happened during 16:00~22:00 :1; other time :0
time of day4	Dummy	The accident happened during 22:00~6:00:1; other time :0
night	Dummy	19:00-07:00 :1; 07:00 -19:00 :0
cloudy	Dummy	Cloudy:1; other weather :0
sunny	Dummy	Sunny :1; other weather :0
rainy	Dummy	Rainy /snowy:1; other weather :0
car	Dummy	The car is responsible for the accident :1; Other vehicles are responsible for the accident :0
passenger car	Dummy	The passenger car is responsible for the accident :1; Other vehicles are responsible for the accident :0
truck	Dummy	The truck is responsible for the accident :1 Other vehicles are responsible for the accident :0
rear end	Dummy	The type of accident is rear end collision :1; The accident type is other type :0
Crash barrier	Dummy	The type of accident is guardrail collision :1; The accident type is other type :0
spontaneous combustion	Dummy	The type of accident is spontaneous combustion :1; The accident type is other type :0
other types of accidents	Dummy	Types of accidents except rear end collision, guardrail collision and spontaneous combustion :1
pcu	Continuous	Passenger car unit.
spilled goods	Dummy	Goods were spilled in the accident :1
car-involved accident	Dummy	Whether there is a car involved in the accident.
bus-involved accident	Dummy	Whether there is a bus involved in the accident.
coach-involved accident	Dummy	Whether there is a coach involved in the accident.
small truck-involved accident	Dummy	Whether there is a small truck involved in the accident.
van-involved accident	Dummy	Whether there is a van involved in the accident.
large truck-involved accident	Dummy	Whether there is a large truck(Semi Trailer) involved in the accident.
short clearance time	Binary	dependent variable Accident duration 1 refers to the accident duration is shorter than 120 minutes, and 0 refers to the opposite.

and the reason for the error recording is not excluded. The clearance time of some highway accidents is even more than 720min (12h), so we suspect the possibility of wrong records or other objective reasons. The data set does not explain in detail the reasons for the excessively long clearance time of highway accident, and the data is not representative. Because of the particularity of these accidents with extremely long clearance time, the classification model will have poor performance on such cases. Therefore, the accident data with a clearance time of longer than 400 min is removed, and finally we obtained 3832 observations.

Table 2.
Weights of different vehicle types

Vehicle type	Car	Bus	Coach	Small truck	Van	Large truck
weight	1	1.5	1.5	1	1	3

Table 3.
Calculation of PCU

Vehicle types	Vehicle types	Vehicle types	pcu
Small truck	Small truck	Small truck	3
Car	Small truck		2
Car	Large truck		4

Table 4.
The maximum, quantile and mean of the accident clearance time of the processed data

Minimum	1st-quantile	Median	Mean	3rd-quantile	Maximum
1.0	28.0	51.0	70.21	88.0	400.0

Table 5.
Descriptive statistics of variables left after screening

Variable name	Type	Mean (percentage for dummies)	variance	min	max
minivan flow	Continuous	0.2968	0.0416	0.0002	1.6216
medium truck flow	Continuous	1.2900	1.1778	0.0000	9.8170
extra large truck flow	Continuous	0.5823	0.1821	0.0001	3.5261
container truck flow	Continuous	1.1543	2.3984	0.0000	9.5250
embedding congestion	Continuous	1.0440	0.1817	0.0010	3.5020
time of day1	Dummy	0.1683	-	0	1
time of day2	Dummy	0.3852	-	0	1
time of day3	Dummy	0.2578	-	0	1
time of day4	Dummy	0.1887	-	0	1
night	Dummy	0.3617	-	0	1
sunny	Dummy	0.6665	-	0	1
car	Dummy	0.5055	-	0	1
rear.end	Dummy	0.7059	-	0	1
Crash barrier	Dummy	0.1561	-	0	1
spontaneous combustion	Dummy	0.0287	-	0	1
pcu	Continuous	2.2192	1.5520	1.0000	13.0000
spilled goods	Dummy	0.0284	-	0	1
short clearance time	Binary	-	-	0	1

Figure 1.
Accident clearance time distribution of processed data

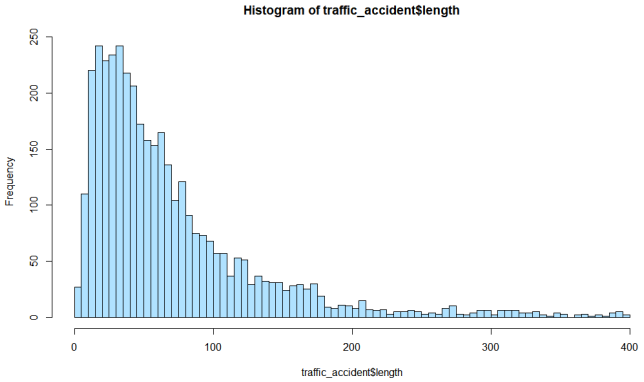

Figure 1 shows the distribution of accident clearance time after removing the accident data with clearance time longer than 400 min. It can be seen from Table 4 that the average value of the retained data is 70.21, and the median value is 51. Most of the data are within the range from 0 to 200 min. With a ratio of 4:1, we obtained a training set with 3,000 observations and a test set with 832 observations.

Variable Selection

It is found that the two variables, i.e., vehicle equivalent and large truck flow, have a strong correlation with other variables. Hence, these two variables are removed from the data set. Because there are too many variables, the Akaike information criterion (AIC) is used for variable screening (Akaike, 1974). AIC is a standard metric to measure the goodness of fit of statistical models. It is based on the concept of entropy, which can measure the complexity of the estimated model and the goodness of the model fitting the data.

The formula of AIC is

$$AIC = 2 * k - 2\ln(L) \quad (1)$$

where k represents the number of parameters in the fitted model. L represents the likelihood of the model.

First, a generalized linear model is established which includes all the

variables except the two highly correlated variables. Then use the 'step()' function in the R software for variables selection. The step function is based on the AIC, and by selecting a model which has the smallest AIC, the set of variables that contains the most useful information is kept. The variables obtained after the variable selection are shown in Table 5. There are 17 variables left (not including dependent variables), of which 6 are continuous variables and the other 11 are dummy variables. The mean, variance, and extreme values of the continuous variables are also shown in Table 5.

When we use dummy variables to represent the vehicle types involved in the accident instead of the PCU, the variables filtered using the AIC are shown in Table 6. In this case, there are 17 variables left (not including dependent variables), four of which are continuous variables, and the other 13 are dummy variables. The mean, variance, and extreme values of the continuous variables are also shown in Table 6. Although six dummy variables were introduced to represent different types of vehicles involved in accidents, only the two variables car-involved accident and small truck-involved remained after the screening.

Table 6.
Results of logistic regression

Variable name	Type	Mean (percentage for dummies)	variance	min	max
minivan flow	Continuous	0.2968	0.0416	0.0002	1.6216
medium truck flow	Continuous	1.2900	1.1778	0.0000	9.8170
container truck flow	Continuous	1.1543	2.3984	0.0000	9.5250
embedding congestion	Continuous	1.0440	0.1817	0.0010	3.5020
time of day1	Dummy	0.1683	-	0	1
time of day2	Dummy	0.3852	-	0	1
time of day3	Dummy	0.2578	-	0	1
time of day4	Dummy	0.1887	-	0	1
night	Dummy	0.3617	-	0	1
sunny	Dummy	0.6665	-	0	1
car	Dummy	0.5055	-	0	1
rear end	Dummy	0.7059	-	0	1
Crash barrier	Dummy	0.1561	-	0	1
spontaneous combustion	Dummy	0.0287	-	0	1
car-involved accident	Dummy	0.6649	-	0	1
small truck-involved accident	Dummy	0.4468	-	0	1
spilled goods	Dummy	0.0284	-	0	1
short clearance time	Binary	-	-	0	1

Methodology

Generalized Linear Model

Logistic regression is a generalized linear model, which is commonly applied to binary or multi-class classification problems. Moreover, logistic regression can show the influence of each independent variable on the dependent variable compared to other classification algorithms.

In the generalized linear model, the dependent variable Y follows the exponential family distribution. The relationship with the covariate $X_1,...,X_p$ is through the formula $\eta = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. $g(x)$ is the link function.

$$g(\mu) = \eta \quad (2)$$

$$\mu = E(Y) \quad (3)$$

For Bernoulli distribution, if the probability of Y = 1 is p_1 , then

$$E(Y) = P(Y = 1 \mid X_1,..., X_n) = p_1 \quad (4)$$

Through Equation 3.2,

$$\mu = p_1 \quad (5)$$

In logistic regression, the logistic function is $h(\eta) = \frac{1}{1+e^{-\eta}}$, and thus $\eta = \ln(\frac{h(\eta)}{1-h(\eta)})$.

It can be found that when η is in the range of negative infinity to positive infinity, $h(\eta)$ range from 0 to 1 and increases monotonically: when η >0, $h(\eta) > 0.5$ when η < 0, $h(\eta) < 0.5$.

From Equations 3.1 and 3.4, we have:

$$p_1 = \mu = g^{-1}(\eta) = h(\eta) = \frac{1}{1+e^{-\eta}}. \quad (6)$$

That is

$$g(p_1) = \ln(\frac{p_1}{1-p_1}) \quad (7)$$

Then we get the logistic regression model as follows:

$$\ln(\frac{p_1}{1-p_1}) = \ln(\frac{1}{e^{-\eta}})$$
$$\eta = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \quad (8)$$

$\frac{p_1}{1-p_1}$ is called the odds, when the value is greater than 1, $p_1 > 0.5$, we think that event {Y = 1} is more likely to happen.

Mixed-effects Model

The principle of the generalized linear model has been introduced in (3.1). By adding a random effect term u_i to the model, the conditional distribution

expectation of the dependent variable Y_{ij}, is defined as followed:

$$\mu_{ij} = E(Y_j \mid u_i, X_j) \quad (9)$$

The conditional mean value is combined with the conditional linear prediction value η_{ij} through the link function :

$$g(\mu_{ij}) = \eta_j = X_j^{'}\beta + Z_j^{'}u_i \quad (10)$$

Equation (3.9) is the general form of the generalized linear mixed model, and Y_{ij} : indicates the j^{th} observed response variable of the i^{th} category, $i = 1,...,m, j = 1,...,n_i$. It is independent under the condition of random effects u_i and follows the exponential distribution family, which can be binomial distribution, Poisson distribution, Gamma distribution, etc. X_{ij} indicates the explanatory variables; β indicates the fixed effect parameter vector; u_i indicates the random effect and it follows the multi-normal distribution with zero mean and a variance-covariance matrix of γ . u_i represents the heterogeneity between the classes caused by the hidden factors and the observed correlation within the same class and are independent of each other between different classes. Z_{ij} indicates the explanatory variable related to random effects. The design matrix has two parts, i.e., fixed effects X and random effects Z .

The generalized linear mixed model is also called the conditional model. When $Z = 1$, $\eta_{ij} = X_j^{'}\beta + u_i$, is the simplest mixed-effects model, namely the random-intercept model. u_i represents the influence of the i^{th} category on the observed value within the class (variation that cannot be explained by the covariate can be observed). σ_u^2 reflects the heterogeneity between different classes.

Due to the non-linear relationship between the dependent variable and the independent variables and the existence of random effects u_i in the model, it is difficult to estimate the parameters of the model. Assume that the likelihood function of the *i*th category is:

$$L_i(\beta, u_i) = f_y(y_i \mid u_i, X_i, \beta) = \prod_{j=1}^{n_i} f_y(y_j \mid u_i, X_j, \beta) \quad (11)$$

Suppose the density function $f_u(u_i, G)$ of random effects u_i is, with marginal likelihood function:

$$L_i(\beta, \gamma) = \int_{u_i} L_i(\beta, u_i) f_u(u_i, \gamma) du_i$$
$$= \int_{u_i} \left\{ \prod_{j=1}^{n_i} f_y(y_{ij} \mid u_i, X_{ij}, \beta) \right\} f_u(u_i, \gamma) du_i, \quad (12)$$

γ is the variance covariance matrix of u_i, and is the parameter estimate of G. The following likelihood function is constructed:

$$L(\beta, \gamma) = \prod_i L_i(\beta, \gamma) \quad (13)$$

It can be seen from the above equations that the calculation of the likelihood function is much more complicated than the linear mixed-effects model, and the problem of high-dimensional integration of random effects u_i needs to be solved. Many approximate inference methods for maximizing the likelihood function have been proposed, e.g., the main integral approximation methods are Laplace approximation (Breslow & Clayton, 1993), Adaptive Gaussian integration, first-order Taylor sequence expansion approximation (Li et al., 2007).

Results Analysis and Disscussions

Model Results with Vehicle Types Encoded as Dummies

First, we use dummy variables to encode the types of vehicles involved in the accident instead of PCU to build the model. Then we use AIC for feature selection. The results of the generalized linear model with vehicle types encoded as dummies are shown in the Table 7. (Only the variables with significance levels above 0.1 are shown in the results.) The AIC of the model is 1969.8, and thus most of the variables are significant. The variables representing the embedding congestion, accident type, responsible car type, and whether there were cars or small trucks involved in the accident are all significant with a level of 99% .

Then we establish a mixed effect model with vehicle types encoded as dummies by using these variables and location. The variable representing whether it is a sunny day is replaced by the variable weather (here weather1 represents the sunny day, weather2 represents cloudy, weather3 represents the rainy day, weather4 represents fog and haze, weather5 represents snowy day). Moreover, we choose time of day, night, and location as random intercept terms to establish three random intercept models.

Use the “anova()” function of the R software to test the AIC and significance of the three models. The results show that the AIC of the model with location as the random intercept term is the smallest, and it is more significant than the other two models. The results show that the medium trucks flow, minivan flow, container truck flow, and extra large truck flow are not significant in the model, and it is also found that the time of day is not very significant. Therefore we remove these five variables.

Next, we take location as the random intercept term and embedding congestion, car1involved accident, and spilled goods as the random slope terms to establish three generalized linear mixed-effects models. Use

Table 7.
Results of logistic regression

	Estimate	Std. Error	z value	Pr(>|z|)	Significance level
(Intercept)	0.7691	0.2859	2.690	0.0071	**
minivan flow	-0.8980	0.4078	-2.202	0.0276	*
medium truck flow	0.1538	0.0778	1.978	0.0479	*
embedding congestion	0.6512	0.1844	3.531	0.0004	***
night1	-0.4612	0.1614	-2.858	0.0043	**
sunny1	0.3871	0.1243	3.115	0.0018	**
car1	1.5619	0.2530	6.174	<0.0001	***
rear end1	0.7848	0.1617	4.854	<0.0001	***
crash barrier	0.7259	0.2087	3.478	<0.0001	***
spontaneous combustion	-0.5984	0.2771	-2.160	0.0308	*
spilled goods	-1.5372	0.2549	-6.030	<0.0001	***
car-involved accident	1.0538	0.1765	5.970	<0.0001	***
small truck-involved accident	0.5644	0.2118	2.665	0.0077	**

Null deviance: 2584.3 Residual deviance: 1935.8 AIC: 1969.8
*parameter significant at the 0.1 level;
**parameter significant at the 0.05 level;
***parameter significant at the 0.01level.

Table 8.
The random effects in the generalized mixed-effects model

Group name	Variance	Std. Error	Corr
Location (Intercept)	0.4820	0.6942	
embedding congestion	0.4257	0.6524	-0.9100

Table 9.
The fixed effects in the generalized mixed-effects model

	Estimate	Std. Error	z value	Pr(>|z|)	Significance level
(Intercept)	0.3242	0.2767	1.172	0.2413	
night1	-0.6648	0.1232	-5.396	<0.0001	***
weather2	-0.4750	0.1430	-3.323	0.0009	***
weather5	-0.9329	0.3235	-2.884	0.0039	**
car1	1.5937	0.2535	6.288	<0.0001	***
rear end1	0.7616	0.1635	4.658	< 0.0001	***
crash barrier	0.7175	0.2125	3.376	0.0007	***
spontaneous combustion	-0.6111	0.2811	-2.174	0.0297	*
car-involved accident	1.1404	0.1745	6.535	<0.0001	***
small truck-involved accident	0.5751	0.2140	2.687	0.0072	**
spilled goods	-1.5597	0.2569	-6.070	<0.0001	***

AIC: 1976.8 Log-likelihood: -972.4 Number of observations: 3000
*parameter significant at the 0.1 level;
**parameter significant at the 0.05 level;
***parameter significant at the 0.01level.

the “anova ()” function(analysis of variance or deviance tables for one or more fitted model objects.) to test the AIC and significance of the three models. The results show that the AIC of the model with embedding congestion as the random slope is the smallest and the most significant. The results are shown in Tables 8 and 9. We found that the variables including whether the accident occurred at night, the weather of the accident, the type of accident, the type of vehicle responsible for the accident, and whether there were cars or small trucks involved accident in the accident, were all significant at the 99% level.

Same as the generalized linear model with vehicle types encoded as dummies, the coefficients of car-involved accident and small truck-involved accident of this model are also positive. The AIC of this model is 1976.8, which is larger than the generalized linear model with vehicle types encoded as dummies, meaning that the mixed effect model with vehicle types encoded as dummies has induced less information loss by introducing the random parameters.

Model Results with PCU Equivalents

The logistic regression model with PCU equivalents was established with the variables selected by AIC. Check the coefficient and significance of each variable, the regression results are shown in Table 10. The AIC of the model is 1982.3, and most of the variables are significant. Embedding congestion, accident type, responsible car type, and PCU are all significant with a level of 99%. The generalized mixed-effects model with PCU equivalents was established by using the selected 14 variables and location. We use time of day, night, and location as random intercept terms to establish three random intercept models.

Similarly, we use the “anova()” function of the R software to test the AIC and significance of the three models. The results show that the AIC of the model with location as the random intercept term is the smallest, and it is more significant than the other two models. The results show that the medium trucks flow, minivan flow, container truck flow, and extra large truck flow are not significant in the model, and it is also found that the time of day is not very significant. Therefore these five variables are removed.

Table 10.
Results of logistic regression

	Estimate	Std. Error	z value	Pr(>|z|)	Significance level
(Intercept)	0.0236	0.2389	0.099	0.9213	
minivan flow	-0.8665	0.4148	-2.089	0.0367	*
medium truck flow	0.1939	0.0847	2.289	0.0221	*
extra large truck flow	-0.5738	0.2723	-2.107	0.0351	*
container truck flow	-0.0900	0.0394	-2.285	0.0223	*
embedding congestion	1.0927	0.2787	3.921	<0.0001	***
time of day3	0.4052	0.1888	2.146	0.0319	*
night1	-0.5074	0.1603	-3.165	0.0016	**
sunny1	0.3845	0.1233	3.119	0.0018	**
car1	1.7684	0.1654	10.691	<0.0001	***
rear.end1	1.4421	0.1770	8.149	< 0.0001	***
crash barrier	0.7678	0.2087	3.680	<0.0001	***
spontaneous combustion	-0.6714	0.2781	-2.414	0.0158	*
pcu	-0.2633	0.0525	-5.012	<0.0001	***
spilled goods	-1.6344	0.2533	-6.451	<0.0001	***

Null deviance: 2584.3 Residual deviance: 1948.3 AIC: 1982.3
*parameter significant at the 0.1 level;
**parameter significant at the 0.05 level;
***parameter significant at the 0.01level.

Table 11.
The random effects in the generalized mixed-effects model

Group name	Variance	Std. Error	Corr
Location (Intercept)	0.4932	0.7023	
embedding congestion	0.4456	0.6675	-0.9300

Table 12.
The fixed effects in the generalized mixed-effects model

	Estimate	Std. Error	z value	Pr(>|z|)	Significance level
(Intercept)	1.2830	0.2024	6.339	<0.0001	***
night1	-0.8007	0.1201	-6.666	<0.0001	***
weather2	-0.4939	0.1415	-3.491	0.0005	***
weather5	-0.9089	0.3177	-2.860	0.0042	**
car1	1.9383	0.1594	12.160	<0.0001	***
rear end1	1.4573	0.1779	8.194	< 0.0001	***
crash barrier	0.7504	0.2124	3.533	<0.0001	***
spontaneous combustion	-0.6597	0.2812	-2.346	0.0190	*
pcu	-0.2524	0.0519	-4.861	<0.0001	***
spilled goods	-1.7033	0.2554	-6.669	<0.0001	***

AIC: 2000.9 Log-likelihood: -985.5 Number of observations: 3000
*parameter significant at the 0.1 level;
**parameter significant at the 0.05 level;
***parameter significant at the 0.01level.

Next, we take location as the random intercept term and take embedding congestion, PCU, and spilled goods as the random slope terms to establish three generalized linear mixed-effects models. We then use the "anova ()" function to test the AIC and significance of the three models. The results show that the AIC of the model with embedding congestion as the random slope is the smallest and the most significant. The results are shown in Table 11 and Table 12. We found the variables representing whether the accident occurred at night, the weather of the accident, the accident type, the type of vehicle responsible for the accident, and the PCU were all significant with a level of 99%.

It can be seen from the coefficients that the variable coefficients of the fixed effects part of the generalized mixed-effects model with PCU equivalents are similar to the variable coefficients of the generalized linear model with PCU equivalents. Except for the significant degree of changes in some variables, there is almost no difference. The AIC of this model is 2001, which is larger than the generalized linear model with PCU equivalents. So the generalized linear model with PCU equivalents performs better in predicting the clearance time of highway accidents.

Comparison and Discussions

Compared with other weather, the accident duration is more likely to be "short" on sunny days. The results were the same as those found in previous studies (Nam & Mannering, 2000). The results show that the highway accidents clearance time attends to be "longer" at night. Studies have also found that accidents that occur at night are more likely to be severe (Ding et al., 2015)

An accident with a car as the responsible party is more likely to last shorter than 120 minutes compared with an accident with a truck or a bus as the responsible party. The accidents with rear-ending collisions and crash barrier collisions are more likely to have a shorter duration than other types of car accidents. When the accident type is spontaneous combustion, the clearance time of the accident is more likely to be longer than 120min. When PCU is larger, i.e., there are more vehicles involved in the accident or the vehicle involved in the accident is a large truck, the accident clearance time is more likely to be "long" which is consistent with certain previous studies (Li et al., 2017). If goods

are spilled during an accident, it is more likely that the accident lasts longer than 120 minutes.

In addition, the results show that the heavier the embedding congestion is, it is more likely that the clearance time of the accident is shorter than 120 minutes, which is inconsistent with our intuition. An explanation is that the heavier the embedding congestion is when an accident occurs, the person in charge will clear the traffic more efficiently to avoid more serious congestion. It can be seen from the model results that the variables representing car-involved accidents and small truck-involved accidents are significant. A positive coefficient means that when a car or a small truck is involved in a car accident, it is more likely that the accident clearance time is less than 120 minutes. On the other hand, if there is a large truck or bus in the accident, it may take longer to clear the accident.

Conclusions

This study analyzed the traffic accident data of 4 expressways including G2, G25, G35, G1511 in Shandong Province to predict whether the clearance time of a traffic accident is greater than 120 minutes. Comparing the results of two generalized linear models and two mixed-effects models, we found the factors that affect the duration to clear traffic accidents. These factors are night, weather, embedding congestion, car, rear end, crash barrier, PCU, spilled goods, car involved accident, and small truck involved accident.

The results of the four-parameter estimation models show that embedded congestion, weather, accident type, and accident vehicle type are the factors that affect the accident clearance time most significantly. If the weather is sunny, it is more likely that the accident clearance time is less than 120min. The accident with a car as the responsible party is more likely to last shorter than 120 minutes. The larger the number of vehicles involved in the accident is or the greater the PCU is, it is more likely that the accident clearance time will be greater than 120 minutes. When the type of vehicle involved in the accident is a small vehicle, such as a small car or a small truck, it is more likely that the clearance time of the accident is 'short'. From the results, when the accident involves multiple vehicles or the type of accident is spontaneous combustion, the person in charge shall deal with the accident scenes as soon as possible. Prevent such accidents from causing serious congestion or more serious accidents.

Due to the limited data we have, the lane closure type and the number of injured are not included. This information may affect the duration of highway accidents. In a future study, we can collect more data to apply to our model. For future work, we are planning to obtain more traffic accident data and investigate the application of artificial intelligence-based methods for accident clearance time prediction. Future studies can compare the fitting and prediction Performance of these models, and we can also introduce more model evaluation criteria, such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Besides, different variable screening methods will be explored, as when the variables used in the models are different, the factors related to the clearance time of the accidents will also be different.

Acknowledgements

This research is supported by the National Key Research and Development Program of China under Grant Nos. 2019YFC0810705 and 2018YFC0807000, and the National Natural Science Foundation of China under Grant No. 71771113.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716-723. https://doi.org/10.1109/TAC.1974.1100705

Baykal-Gürsoy, M., Xiao, W., & Ozbay, K. (2009). Modeling traffic flow interrupted by incidents. *European Journal of Operational Research, 195*(1), 127-138. https://doi.org/10.1016/j.ejor.2008.01.024

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9-25. https://doi.org/10.1080/01621459.1993.10594284

Chung, Y. (2010). Development of an accident duration prediction model on the Korean Freeway Systems.*Accident Analysis and Prevention, 42*(1), 282–289. https://doi.org/10.1016/j.aap.2009.08.005

Chung, Y., Chiou, Y., & Lin, C. (2015). Simultaneous equation modeling of freeway accident duration and lanes blocked. *Analytic Methods in Accident Research, 7*, 16-28. https://doi.org/10.1016/j.amar.2015.04.003

Dimitriou, L., & Vlahogianni, E. I. (2015). Fuzzy modeling of freeway accident duration with rainfall and traffic flow interactions. *Analytic Methods in Accident Research, 5-6,* 59–71. https://doi.org/10.1016/j.amar.2015.04.001

Hojati, A.T., Ferreira, L., Washington, S., & Charles, P. (2013). Hazard based models for freeway traffic incident duration. Accident; Analysis and Prevention, *52*, 171–181. https://doi.org/10.1016/j.aap.2012.12.037

Hou, L., Lao, Y., Wang, Y., Zhang, Z., Zhang, Y.i., & Li, Z. (2013). Modeling freeway incident response time: a mechanism-based approach. *Transportation Research Part C: Emerging Technologies, 28,* 87–100. https://doi.org/10.1016/j.trc.2012.12.005

Garib, A., Radwan, A.E., & Al-Deek, H., (1997). Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering. 123* (6), 459–466. https://schlr.cnki.net/Detail/doi/GARJ8099_1/SJCE13092200026588

Ji, Y., Zhang, X., & Sun, L. (2008). A Review of the Traffic Incident Duration Prediction Methods. *Highway Engineering, 33*(3), 72-79+141. http://dx.chinadoi.cn/10.3969/j.issn.1674-0610.2008.03.017

Lee, Y., & Wei, C. H. (2010). A computerized feature selection method using genetic algorithms to forecast freeway accident duration times. *Computer-Aided Civil and Infrastructure Engineering, 25*(2), 132–148. https://doi.org/10.1111/j.1467-8667.2009.00626.x

Li, R., Pereira, F.C., Ben, A., & Moshe, E. (2015). Competing risk mixture model and text analysis for sequential incident duration prediction. *Transportation Research Part C: Emerging Technologies, 54,* 74-85. https://doi.org/10.1016/j.trc.2015.03.009

Li, R., Guo, M. & Lu, H. (2017). Analysis of the Different Duration Stages of Accidents with Hazard-Based Model. *International Journal of Intelligent Transportation Systems Research, 15*(1), 7-16. https://doi.org/10.1007/s13177-015-0115-6

Li, L., Hao, Y., Zhang, P., Zou, Y., Zou, Z., Zhang, Y., & Zhou, S. (2007). Generalized linear mixed effect model and its application. *Modern Preventive Medicine, 11*, 2103-2104. http://dx.chinadoi.cn/10.3969/j.issn.1003-8507.2007.11.038

Lin, L., Wang, Q., & Sadek, A. (2016). A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations. *Accident Analysis & Prevention. 91,* 114-126. https://doi.org/10.1016/j.aap.2016.03.001

Lin, L., Wang, Q., & Sadek, A. (2014). Data mining and complex network algorithms for traffic accident analysis. *Transportation Research Record, 2460,* 128–136. https://doi.org/10.3141/2460-14

Mannering, F., & Bhat, C. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research, 1*, 1-22. https://doi.org/10.1016/j.amar.2013.09.001

Meng, F., Xu, P., Song, C., Gao, K., Zhou, Z., & Yang, L. (2020). Influential factors associated with consecutive crash severity: A two-level logistic modeling approach. *International Journal of Environmental Research and Public Health, 17*, 15. https://doi.org/10.3390/ijerph17155623

Nam, D., & Mannering, F. (2000). An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy Practice, 34* (2), 85–102. https://doi.org/10.1016/S0965-8564(98)00065-2

Park, H., & Haghani, A. (2016). Real-time prediction of secondary incident occurrences using vehicle probedata. *Transportation Research Part C: Emerging Technologies, 70,* 69-85. https://doi.org/10.1016/j.trc.2015.03.018

Schrank, D., Eisele, B., Lomax, T., & Bak, J. (2015). 2015 Urban mobility scorecard and appendices. *Texas A&M Transp. Institute.* https://rosap.ntl.bts.gov/view/dot/61407

Statistical bulletin on the development of the transportation industry in 2019. (2020). *Transportation Accounting, 6*, 86-91. http://dx.chinadoi.cn/10.13646/j.cnki.42-1395/u.2020.05.016

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. https://doi.org/10.1007/s12544-010-0031-4

Wei, C., & Lee, Y. (2007). Sequential forecast of incident duration using Artifificial Neural Network models. *Accident Analysis and Prevention 39*(5), 944–954. http://dx.doi.org/10.1016/j.aap.2006.12.017

Wen, Y., Chen, S., Xiong, Q., Han, R., & Chen, S. (2012). Traffic incident duration prediction based on K-nearest neighbor. *Applied Mechanics and Materials, 253–255,* 1675–1681. https://doi.org/10.4028/www.scientific.net/AMM.253-255.1675

Xia, Z. (2016). Probability prediction model of expressway traffic accident duration. *Highway and Automobile Transportation, 3*, 52-55. http://dx.chinadoi.cn/10.3969/j.issn.1671-2668.2016.03.014

Xu, Z., He, Y., & Sun, X. (2013). Review of Expressway Accident Clearance Time Prediction Methods. *Transportation Standardizatio*n, *21*, 130-134. http://dx.chinadoi.cn/10.16503/j.cnki.2095-9931.2013.21.038

Zhang, H., Zhang, Y., & Khattak, A. (2012). Analysis of Large-Scale Incidents on Urban Freeways. *Transportation Research Record, 2278*(1), 74-84. https://doi.org/10.3141/2278-09