Applied Mathematics, Modeling and Computer Simulation
C.-H. Chen et al. (Eds.)

doi:10.3233/ATDE230984

Research on Predicting Duration of Urban Submarine Tunnel Traffic Accidents Based on PCA-LGBM Model

Yuting WANG[a], Xiaoyu CAI[b,1] Bo PENG[b]
[a] *School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou (510275), Guangdong, China*
[b] *College of Smart City, Chongqing Jiaotong University, Chongqing (400074), China*

Abstract. Tunnels are an essential component of urban transportation. Compared to open roads, accidents inside tunnels tend to have longer durations, and the road closures resulting from accidents have a greater impact on traffic operations, particularly in the case of submarine tunnels. This study focuses on investigating the characteristics and trends of accident duration in submarine tunnels based on accident data from Qingdao Jiaozhou Bay in China from 2018 to 2020. Firstly, the study examines the influence of ten variables, including the number of vehicles involved, accident types, and weather conditions, on the duration of accidents. The data indicate that the manner and quantity of vehicles leaving the accident scene are critical factors affecting accident duration, while weather conditions have no significant impact. Furthermore, considering the correlations among the influencing factors and the high-dimensional sparsity of the data, a PCA-LGBM model for accident duration prediction is constructed. This model combines the dimensionality reduction capability of Principal Component Analysis (PCA) with the powerful prediction capability of LightGBM (LGBM). Finally, experimental results demonstrate that compared to other models such as MLR, BPNN, PCA-BPNN, and LGBM, the proposed model exhibits superior performance with a minute-level accuracy rate of 75%.

Keywords. Urban submarine tunnel; traffic accidents; accident duration; PCA; LGBM

1. Introduction

Underwater tunnels in urban areas have more complex traffic environment features compared to open roads. Consequently, they are prone to frequent traffic accidents. After an accident occurs, tunnel management authorities close the lanes to ensure safety, and the duration of lane closure depends on the duration of the accident. Therefore, accurate prediction of accident duration is crucial for decision-making in the management of underwater tunnel traffic operations. It also helps users choose alternative routes to avoid congestion.

Statistical methods and machine learning are two primary approaches for predicting the duration of accidents. Regression models were the earliest ones used for accident duration prediction [1-4]. When dealing with complex and highly nonlinear relationships

1 Corresponding author, Xiaoyu CAI, Chongqing Jiaotong University, E-mail: caixiaoyu@cqjtu.edu.cn

between dependent and independent variables, machine learning methods have shown superior performance [5-9]. Statistical methods have strict mathematical assumptions and functional structures, making them superior to machine learning methods in explaining the mathematical relationships between accident duration and influencing factors. However, machine learning methods exhibit higher accuracy and stability in model predictions, with tree-based models performing exceptionally well. Currently, most research focuses on a single method, and the research scenarios are mostly limited to open roads, with relatively fewer studies focusing on tunnel scenarios, especially underwater tunnel scenarios. Current studies on the factors influencing accident duration include variables related to road conditions, environment, accident characteristics, etc. [10]. However, no research has been found on the mode of vehicle departure after an accident, which is one of the reasons for the relatively low accuracy in time prediction.

To address the aforementioned issues, this paper proposes a combination model, PCA-LGBM, based on Principal Component Analysis (PCA) and the tree-based model Light Gradient Boosting Machine (LGBM). This model aims to improve the prediction accuracy and algorithm interpretability and achieve minute-level prediction of accident duration in underwater tunnel scenarios.

2. Materials

2.1. Data Resource

This study collected data on 2,047 traffic accidents that occurred in the Qingdao Jiaozhou Bay Undersea Tunnel in China between 2018 and 2020. The data includes 11 variables, such as accident duration, accident location, and accident type.

2.2. Data Analysis

2.2.1. Accident Duration

In this study, accident duration is used to represent the time interval from accident discovery to accident clearance.

According to statistical analysis, the overall distribution of accident duration in the tunnel is shown in figure 1, which exhibits a long-tail distribution. The maximum duration is 120 minutes, the minimum duration is 1 minute, and the average duration is 11 minutes.

2.2.2. Traffic Accident Characteristics

This study collected data on 10 accident characteristics, including accident type, vehicle departure mode, need for traffic police, vehicle driving direction, accident location, accident time period, weather conditions, number of involved vehicles, number of towed vehicles, and number of mediating vehicles, as shown in table 1. It is worth noting that the information regarding vehicle exit mode, number of towed vehicles, and number of mediating vehicles has not been previously addressed in previous studies.

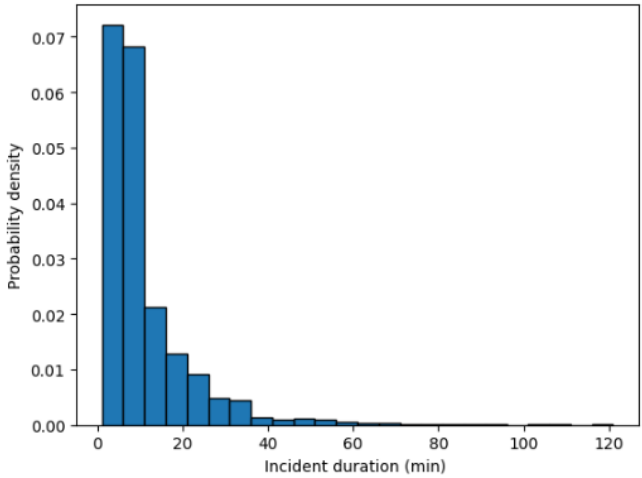


Figure 1. Histogram of accident duration.

Table 1. Characteristics of traffic Accident in and out of tunnels (2018–2020).

Variable	Definition	Quantity	Proportion (%)
Accident type	1 = Hit-fixed-object	14	0.7%
	2 = Rear-end	2033	99.3%
Traffic police	0 = No	1991	97.3%
	1 = Yes	56	2.7%
Direction	1 = North to south	981	48.0%
	2 = South to North	1066	52.0%
Mileage marker	1 = 1~2 km	32	1.6%
	2 = 2~3 km	216	10.6%
	3 = 3~4 km	269	13.1%
	4 = 4~5 km	514	25.1%
	5 = 5~6 km	520	25.4%
	6 = 6~7 km	300	14.7%
	7 = 7~8 km	159	7.7%
	8 = 8~9 km	37	1.9%
Day or night	0 = Bright	1993	97.4%
	1 = Dark	54	2.6%
Weather	1 = Wet	147	7.2%
	2 = Dry	1900	92.8%
Departure mode	1 = Tow away	395	19.3%
	2 = Mediated departure	1269	62.0%
	3 = Self-driving departure	383	18.7%
Number of vehicles	1	15	0.7%
	2	1380	67.4%
	3	444	21.7%

Variable	Definition	Quantity	Proportion (%)
	4	147	7.2%
	5	46	2.2%
	6	9	0.4%
	7	3	0.1%
	9	2	0.1%
	0	1614	78.8%
	1	326	15.9%
Number of towed vehicles	2	100	4.9%
	3	5	0.2%
	4	2	0.1%
	0	253	12.4%
	1	854	41.7%
	2	674	32.9%
Number of mediating vehicles	3	178	8.7%
	4	70	3.4%
	5	15	0.7%
	6	3	0.1%

In terms of accident types, rear-end collisions account for 99.3% of the total. The dominance of rear-end collisions in tunnels can be attributed to factors such as the prohibition of lane changing except at merging/diverging sections. Collisions with fixed objects, such as tunnel sidewalls, are usually associated with hazardous driving behaviors (e.g., speeding) and violations (e.g., drunk driving), often requiring the intervention of traffic police.

Regarding spatial characteristics, there is not much difference in the direction of travel. In contrast to previous research conclusions, traffic accidents typically occur at tunnel entrances and exits, while in underwater tunnels, accidents are concentrated in the middle section (50.5%). Underwater tunnels often have a “U” or “V” shaped longitudinal profile, and drivers tend to experience increased pressure in long tunnels, leading to a tendency to unconsciously accelerate to maintain visual continuity. Consequently, rear-end collisions are more likely to occur when vehicles travel too fast or fail to brake in time at the bottom of the slope. In terms of temporal characteristics, the majority of accidents occur during the daytime, with a smaller proportion occurring at night. Most accidents occur in dry weather conditions, with only 7.2% occurring during rainy or snowy weather.

The number of vehicles involved in accidents has been mentioned in previous research, but the mode of vehicle departure and the corresponding quantity have not been addressed. In this study, the mode of vehicle departure and its quantity are defined as the vehicle departure situation. The vehicle departure modes are classified into three categories: (1) accidents that do not cause any damage, and the vehicle owner drives away on their own, (2) accidents that result in minor damage but lead to disagreements among vehicle owners, requiring third-party intervention for resolution before departure, and (3) accidents with severe consequences where the accident vehicle cannot be started

and requires towing. According to data statistics, 19.3% of vehicles need to be towed, 62.0% require mediation before departure, and 18.7% of vehicles leave on their own.

3. Methodology

3.1. PCA-LGBM Combination Model

First, the original correlated variables are transformed into a set of linearly uncorrelated variables using PCA. The number of principal components is determined based on the cumulative contribution rate. Then, an LGBM model is built based on these principal components. The PCA algorithm workflow in this study is as follows.

(1) The standardized collection of the original indicator data results in a p-dimensional factor vector $x = (x_1, x_2, \dots, x_p)^T$, with n ($n = 2047$) samples denoted as $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, where $i = 1, 2, \dots, n$, and $n > p$. The data is transformed using standardization. Standardization matrix Z:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, j = 1,2, \dots, p \tag{1}$$

where $\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$, $s_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n-1}}$.

(2) Calculation of the covariance matrix R from the standardized matrix Z:

$$R = \frac{Z^T Z}{n-1} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2}}, i, j = 1,2, \dots, p \tag{2}$$

(3) The covariance matrix R is calculated, and its characteristic equation $|R - \lambda I_p| = 0$ is solved to obtain p eigenvalues. The principal components are then determined based on the cumulative contribution rate β, and the number of principal components m is determined.

$$\beta = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{3}$$

For each λ_j, $j = 1,2, \dots, m$, solving the equation system $Rb = \lambda_p b$, we obtain the unit eigenvectors b_j^o.

(4) Transform the standardized indicator variables into principal components.

$$U_{ij} = z_i^T b_j^o, j = 1,2, \dots, m \tag{4}$$

LGBM (Light Gradient Boosting Machine) is an ensemble algorithm for gradient boosting frameworks. One of the innovative ideas of this algorithm is the use of the

Gradient-based One-Side Sampling (GOSS) algorithm. GOSS selectively retains instances with larger gradients while randomly sampling instances with smaller gradients. The GOSS algorithm first sorts the instances based on the absolute values of their gradients and selects the top "*a*" instances. Then, it randomly samples "*b*" instances from the remaining data. When calculating the information gain, the algorithm multiplies the gradients of the sampled instances with small gradients by (1-*a*)/*b*. This approach allows the algorithm to focus more on the undertrained instances without significantly altering the distribution of the original dataset. Let O be the training datasets on a fixed node of the decision tree. The variance gain of splitting feature j at point d for this node is defined as

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{\left(\sum_{\{x_i \in O: x_{ij} \le d\}} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum_{\{x_i \in O: x_{ij} > d\}} g_i\right)^2}{n^j_{r|O}(d)}\right) \quad (5)$$

where $n_O = \sum I[x_i \in O]$, $n^j_{l|O}(d) = \sum I[x_i \in O: x_i \le d]$, $n^j_{r|O}(d) = \sum I[x_i \in O: x_i > d]$.

The formula for calculating the estimated variance gain $\tilde{V}_j(d)$ of the GOSS algorithm is as

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{\{x_i \in A: x_{ij} \le d\}} g_i + \frac{1-a}{b}\sum_{\{x_i \in B: x_{ij} \le d\}} g_i\right)^2}{n^j_l(d)} + \frac{\left(\sum_{\{x_i \in A: x_{ij} > d\}} g_i + \frac{1-a}{b}\sum_{\{x_i \in B: x_{ij} > d\}} g_i\right)^2}{n^j_r(d)}\right) \quad (6)$$

A represents the subset with larger gradients and *B* represents the subset with smaller gradients. And $\frac{1-a}{b}$ is used to normalize the sum of the gradients over *B*.

Furthermore, the exclusive feature bundling algorithm can combine many exclusive features into fewer dense features, effectively avoiding unnecessary computation for zero feature values.

The data samples processed by the PCA algorithm are divided into training and testing sets in a ratio of 0.8:0.2, where 80% of the data is used to train the proposed model and 20% is used to test the trained model.

3.2. Model Evaluation Index

This study primarily evaluates the prediction accuracy and performance of the model based on the error between predicted values and actual values. Commonly used indicators include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Error Ratio ($\Delta\varepsilon$).

4. Results and Discussions

4.1. Principal Component Analysis Results

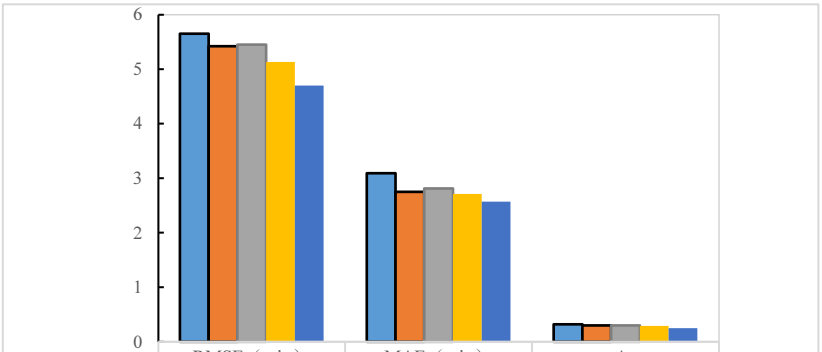
Based on the PCA, four principal components were obtained with a cumulative contribution rate (β) of 75%. Principal Component 1 mainly reflects the information of the number of vehicles, departure mode, number of towed vehicles, and number of

persuaded vehicles. These four features primarily describe the departure information of the accident participants. Principal Component 2 mainly reflects the driving direction and tunnel mileage marker of the accident location. These two features describe the spatial location information of the accident. Principal Component 3 reflects the accident type and whether police intervention is required, while Principal Component 4 reflects the accident occurrence time information.

4.2. Principal Component Analysis Results

To compare the performance of the proposed PCA-LGBM model with other algorithms, the results were compared with MLR, BPNN, PCA-BPNN, and LGBM models.

According to figure 2, it can be observed that the PCA-LGBM model has the best predictive performance, with predicted values closely matching the true values. And the PCA-LGBM model has the lowest errors in all aspects, indicating that the model has the optimal performance.



	RMSE（min）	MAE（min）	Δε
MRL	5.65	3.09	0.32
BPNN	5.42	2.75	0.3
LGBM	5.45	2.81	0.3
PCA-BPNN	5.13	2.71	0.29
PCA-LGBM	4.7	2.57	0.25

Figure 2. Evaluation index of each model.

5. Conclusions

This study investigates the duration of traffic accidents in urban underwater tunnels using real-world data. It proposes a PCA-LGBM model that combines the dimensionality reduction capability of PCA with the powerful predictive ability of LGBM. It is the first study to incorporate the departure mode and quantity of vehicles as factors influencing accident duration, thereby further improving the predictive accuracy of the model. Experimental results demonstrate that the performance of the proposed model is significantly better than four other models, namely MRL, LGBM, BPNN, and PCA-BPNN. Future research will consider collecting more comprehensive accident-related information, such as traffic volume and casualties, and expanding the sample size to ensure higher accuracy and generalizability of the model.

References

[1] Chung Y, Walubita L, Choi K. Modeling Accident Duration and Its Mitigation Strategies on South Korean Freeway Systems [J]. Transportation Research Record Journal of the Transportation Research Board, 2010, 2178:49-57.
[2] Hojati A T, Ferreira L, Washington S, et al. Hazard based models for freeway traffic incident duration [J]. Accident Analysis & Prevention, 2013, 52(12): 171-181.
[3] Zou Y, Henrickson K, Lord D, et al. Application of Finite Mixture Models for Analyzing Freeway Incident Clearance Time [J]. Transportmetrica A Transport Science, 2015, 12(2): 1-23.
[4] Zou Y, Tang J, Wu L, et al. Quantile analysis of factors influencing the time taken to clear road traffic incidents [J]. Proceedings of the Institution of Civil Engineers, 2017, 170(5): 296-304.
[5] Valenti G, Lelli M, Cucina D. A comparative study of models for the incident duration prediction[J]. European Transport Research Review, 2010, 2(2): 103-111.
[6] Li D, Wu J, Peng D. Online Traffic Accident Spatial-Temporal Post-Impact Prediction Model on Highways Based on Spiking Neural Networks [J]. Journal of Advanced Transportation, 2021: 1-20.
[7] Karuppaiah K S, Palanisamy N P G. Heterogeneous ensemble stacking with minority upliftment (HESMU) for churn prediction on imbalanced telecom data [J]. Materials Today Proceedings, 2021: 1-8.
[8] Ma X, Ding C, Luan S, et al. Prioritizing Influential Factors for Freeway Incident Clearance Time Prediction Using the Gradient Boosting Decision Trees Method [J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(9): 2303-2310.
[9] Grigorev A, Mihaita A S, Lee S, et al. Accident duration prediction using a bi-level machine learning framework with outlier removal and intra-extra joint optimisation [J]. arXiv, 2022. DOI:10.48550/arXiv.2205.05197.
[10] Luo Q, Liu C. Exploration of road closure time characteristics of tunnel traffic accidents: A case study in Pennsylvania, USA [J]. Tunnelling and Underground Space Technology, 2023, 132: 104894.