Statistical Distribution and Influencing Factor Analysis of the Impact Range of the Traffic Accident

Fujian Wang[1,3,a], Hui Tian[1,2,b], Shengqiang Jia[3,c], Zhenyu Mei[1,2,d]

{[a]ciewfj@zju.edu.cn, [b]2547730830@qq.com, [c]ciewfj@126.com, [d]meizhenyu@zju.edu.cn}

[1]College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, Zhejiang, China 310058;
[2]Center for Balance Architecture, Zhejiang University, Hangzhou, Zhejiang, China 310058;
[3]The Architectural Design & Research Institute of Zhejiang University Co. Ltd, Hangzhou, Zhejiang, China 310012

Abstract: In order to provide a reference for managers to control traffic flow after traffic accidents, multivariate ANOVA is used to study the significant influencing factors of the impact range of the traffic accident. The impact range of the traffic accident is the length of the road extent affected by the accident. Through the real traffic accident dataset of Humboldt County, California, the distribution characteristics of the impact range of the accident were analyzed, and then the factors influencing the impact range of the accident such as weather, peak time, day/night, season, weekend and intersection were analyzed by the intersubjective effect test, and then the factors with significant influence were analyzed by the post-hoc test to further determine the significances of differences among different categories within the factors. The results show that the distribution of the impact range of the accident has the phenomenon of right skewness, and the lognormal distribution fitting effect is the best. Intersection, weather, and peak hour have significant effects on the impact range of the accident. In addition, the effects of sunny day and rainy day on the impact range of the accident are significantly different, the effects of sunny day and haze on the impact range of the accident are significantly different, the effects of cloudy day and rainy day on the impact range of the accident are significantly different, and the effects of morning peak and off-peak on the impact range of the accident are significantly different.

Keywords: traffic accident; impact range of the traffic accident; lognormal distribution; multivariate ANOVA; intersubjective effect test; post-hoc test.

1 INTRODUCTION

Traffic accidents have been a major focus of research in the traffic safety community. The occurrence of traffic accidents not only brings loss of life and property to people, but also affects the upstream traffic flow, so that the quality of vehicle operation around the accident location is reduced and traffic congestion is formed, and untimely control may also cause secondary accidents[1]. Identifying the distribution and influencing factors of the impact range of the accident is of great significance for managers to take traffic interventions, evacuate traffic flows and avoid secondary accidents[2].

With regard to the impact range of the accident, previous studies have focused on modelling estimates of the scope of impact of accidents. Some scholars calculate the duration of the

ICPDI 2023, September 01-03, Chongqing, People's Republic of China

DOI 10.4108/eai.1-9-2023.2338839

accident in the freeway section from the delay analysis based on traffic engineering theory, and estimate the impact range of the accident[3-5].

Chung et al. (2015) developed a binary integer programming method to estimate the spatiotemporal impact of freeway accidents[6]. Unlike freeway sections, urban roads are more complex road networks, and Sun et al. (2018) estimates and evaluates the impact of accidents on urban roads by extracting the characteristics of the road network[7], and finds that accidents in locations with significant road network connectivity have a wider range of impacts. Sun et al. (2019) established a congestion judgment model based on speed difference to describe the temporal and spatial distribution characteristics of the impact of accidents from the consideration of the speed change of vehicles on urban roads caused by accidents[8]. Tang et al. (2022) considered the characteristics of vehicle queuing to estimate the impact range of traffic accidents in the urban road network[9], and simulated and analyzed the urban area road network built by collecting Changsha license plate data and VISSIM, and found that the duration of the accident and the number of lanes had a greater effect on the impact scope of the accident. Eboli et al. (2020) investigated the factors influencing accident severity, and found that the related factors were grouped in different categories referring to road, external environment, and driver[10]. AlKheder et al. (2022) studied the effects of fog, rain, dust, and fine weather conditions on traffic accident types and frequency, and the weather conditions showed a significant impact on the types of traffic accidents[11]. Retallack and Ostendorf (2020) analyzed the relationship between traffic volume and accident frequency at intersections, and found that accident frequency increases nonlinearly in the higher levels of congestion[12].

In the era of intelligent transportation, accurate and diversified traffic information data collection technology has been widely used, but at present, no scholar has directly analyzed the statistical distribution of the impact range of the accident through the measured data, and the factors affecting the impact range of the accident, such as weather, time characteristics and other factors, have not been comprehensively considered.

Therefore, we analyze the data on the impact of traffic accidents captured by traffic sensors in the road network in recent years, covering most of the year, including various types of weather and different characteristics of the road network. Through the real accident data set, the distribution characteristics of the accident impact range are explored, and the multivariate ANOVA analysis is carried out to find out the significant influencing factors of the accident impact range, so as to provide a reference for managers to take reasonable countermeasures for traffic flow after traffic accidents.

2 ACCIDENT DATA PREPROCESSING

2.1 Description of the dataset

The traffic accident data in this article is from the US traffic accident dataset "US-Accidents"[13]. This is a nationwide traffic accident dataset covering 49 states in the United States. The raw data was collected by U.S. and its state departments of transportation, law enforcement agencies, and traffic cameras and sensors within the road network from February 2016 to December 2020. The traffic accidents in this dataset record the spatial attributes that describe the accident, such as the accident location, and whether there is an intersection near

the accident location; time attributes such as the time of the accident; and meteorological properties such as weather, temperature, humidity, etc. In this paper, accident data from Humboldt County, California, in "US-Accidnts" for the whole year of 2020 was selected as research data.

2.2 Description of the impact range of the accident

Traffic accidents on the road can block some lanes or even block traffic, and the affected vehicles upstream of the accident location will slow down or stop to cause queues. Therefore, the occurrence of traffic accidents has a certain range of influence, and this range value has been given in the original dataset used in this article, expressed as Distance, in miles. As shown in Figure 1, the distance between A and B is defined as the impact range of the accident, which means the length of the road extent affected by the accident.

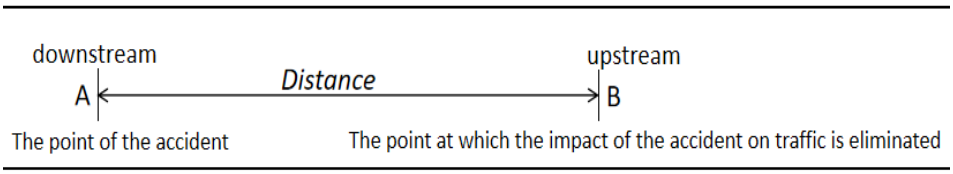

Figure 1. The length of the road extent affected by the accident

2.3 Outlier preprocessing

The accident impact range is measured by sensing devices such as induction loops, and data errors may occur during data transmission and data recording, so the accident impact range Distance value is analyzed to improve data quality. In addition, this article only studies those accidents that have an impact on traffic. The detection methods of outliers in statistics usually include Z-score method, Rheinda criterion, absolute median deviation (MAD) method and interquartile range (IQR) method, the first three methods are suitable for data obeying normal distribution, while IQR method does not depend on a specific distribution and is more universal, so the IQR method is selected to treat Distance.

IQR is a measure of variability by dividing a dataset into quartiles, it divides a hierarchically ordered data set into four equal parts, so there are Q1 (1st quartile), Q2 (2nd quartile), and Q3 (3rd quartile), and IQR is defined as Q3-Q1. The accident impact range Distance values greater than (Q3+1.5IQR) and less than (Q1-1.5IQR) are considered outliers. Then the outliers are excluded according to the upper and lower limits obtained by the IQR method.

2.4 Missing value preprocessing

In the process of data transmission and recording, it is inevitable that there will be missing values. If the missing data is not handled well, the reliability of the analysis results will be affected. The absence rate of weather factors in this study was 9.3%. Machine learning methods are superior to statistical methods in missing value filling and have high stability[14]. At the same time, since there are also temperature, humidity, rainfall and other factors that are highly related to the weather in the dataset, this paper uses machine learning methods to fill in the missing weather values.

Extreme Gradient Boosting (XGB) is an ensemble algorithm based on gradient boosting decision tree (GBDT), which is more accurate than GBDT and can prevent overfitting. The XGB method is as follows. Assuming that there are a total of *K* decision trees in an ensemble model, the estimate given by the model on sample *i* is:

$$y_i = \sum_{k=1}^{K} f_k\left(x_i\right) \tag{1}$$

where y_i represents the estimation result of the sample x_i, $f_k\left(x_i\right)$ represents the sum of the estimated scores of the leaf nodes of the *k*th tree when the sample x_i is input to the *k*th tree. This paper fills in the missing values of weather based on XGB, where weather is used as a label, and temperature, humidity and rainfall are formed a matrix *X*. The steps are as follows:

1) Divide the data into datasets with missing weather and datasets without missing weather.

2) The dataset without missing weather is divided into training and validation sets according to a 7:3 ratio.

3) For several important parameters of the model, grid search is used to traverse the candidate parameters for the training set to find out the optimal parameter combination.

4) Set the model parameters selected in step 3 to train the XGB model.

5) Using the validation set to validate the state and convergence of the model.

6) Enter the dataset with missing weather into the best XGB model, and fill in the missing values through model estimation.

Through the experimental calculation of the above steps, the accuracy of the XGB model constructed in this paper reaches 87.8%. So the missing values of the weather are filled in with high quality.

2.5 Factor selection

The focus of this paper is to consider how to analyze the influencing factors of the impact range of the accident through some easily available factors after the traffic accident, so as to provide managers with references before adopting traffic control measures. In the accident scenario, weather, time characteristics, and road network characteristics are the most readily available factors, and it is necessary to analyze which factors have the most significant influence under the combined actions of these factors. Based on this, this paper selects six influencing factors representing time, space, and external conditions, such as peak time, day/night, weekend, season, intersection, and weather in Table 1.

Table 1. Summary of the classification of selected factors

Number	Factor name	Description	Category within factor
F1	Weather_Condition	Show weather conditions	(1)sunny, (2)cloudy, (3)haze, (4)rainy
F2	Peak_time	Whether it's peak hour	(1)morning peak, (2)evening peak, (3)off-peak

F3	Day_Night	Shows day or night	(1)day, (2)night
F4	Weekends	Show weekend or other	(1)weekends, (2)other
F5	Season	Display the season in which the incident occurred	(1)spring, (2)summer, (3)autumn, (4)winter
F6	Junction	Shows if there is an junction in the area around the accident	(1)yes, (2)no

3 ANALYSIS OF INFLUENCING FACTORS OF THE IMPACT RANGE OF THE ACCIDENT

3.1 Fitting of the accident impact range distribution

In past research, scholars have mainly focused on the modeling of the impact range of accidents, and most of the data in the analysis of the models are indirectly obtained by simulation, rather than real data. The data in this paper are from the real accident data collected through traffic information collection equipment in recent years, covering most of the year, including accidents under various weather and different road network characteristics, which can better conform to the research of the impact range distribution of accidents.

First, with the help of Python's third-party package Fitter, the data samples are fitted, and Fitter can fit the distribution that these data obey, and its scientific calculation library covers all the distributions we know. Taking the accident impact range Distance as the sample input, all the distributions in the fitter package were tried, and finally the fitter output showed that the lognormal distribution (lognorm), inverse gauss distribution (invgauss), and gamma distribution (gamma) were more in line with the distribution of the data samples in this study.

The effects of the three fitted distributions are tested by the K-S test which is a test method that compares the distributions of two observations. The K-S tests were performed on samples at significant levels of 0.05 and 0.01, respectively, and the results are shown in Table 2.

From the fitting results, at a significance level of 0.05, all three distributions are H = 0, indicating that the null hypothesis is accepted, in which the P value of the lognormal distribution is the largest, indicating that the fitting effect of the lognormal distribution is the best. At a significance level of 0.01, only the lognormal distribution accepts the null hypothesis. In summary, the lognormal distribution can better describe the distribution of the impact range of the accident. The effect of lognormal distribution fitting is shown in Figure 2.

In fact, in most traffic accidents, accidents with little impact on traffic flow always account for the majority, so the smaller Distance value accounts for most of the totality, which is shown in the distribution map that the accident impact range is concentrated on the left, and a small number of accidents will have a greater impact on the traffic flow, which is reflected in the phenomenon of tailing on the right side of the distribution map. Therefore, the lognormal distribution can better describe the distribution of accident impact range.

Table 2. Results of K-S test

Distribution type	α = 0.05		α = 0.01	
	H	P	H	P
lognormal	0	0.2689	0	0.0814
inverse Gaussian	0	0.1543	1	0.0037
gamma	0	0.0621	1	0

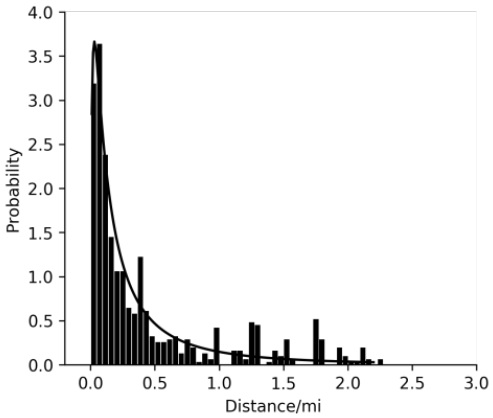

Figure 2. The effect of lognormal distribution fitting

3.2 Multivariate ANOVA of the impact range of the accident

As mentioned above, the optimal distribution of the accident influence range is the lognormal distribution, that is, its logarithms follow the normal distribution. Therefore, some parameter test methods in statistics can be used for factor analysis. Multivariate ANOVA is used to study whether two or more influencing factors have a significant impact on observations. The intersubjective effect test in the results of multivariate ANOVA indicates that under multiple levels of multiple factors, it describes the overall influence of one factor on the observed value at each level of other factors. The influence range of the accident is affected by many factors in the accident scene, and there are 6 factors in the study of this paper, including factors such as the time of the accident, the weather, and the characteristics of the road network at the time of the accident. This article uses SPSS 26 to analyze which factors have a significant impact on the observed values.

Multivariate ANOVA generally requires that the observations conform to a normal distribution and meet the homogeneity of variance, which refers to the consistent fluctuation of the observed values corresponding to each category within the factor. Since the accident impact range data in this paper satisfy the lognormal distribution, its logarithms satisfy the normal distribution, and the accident impact range Distance value is logarithmically transformed. Then, a relatively robust Levin homogeneous test of variance is carried out, the significance level is set to 0.05, and the null hypothesis is proposed: H0—variance homogeneity; H1—variance. From the results of Levin's homogeneity of variance test based

on mean and median, the P values of all six factors are greater than 0.05, so the factors obey the null hypothesis, satisfy the homogeneity of variance and meet the conditions of ANOVA.

The intersubjective effect test is as follows. The significance level is set to 0.05, and the null hypothesis is: H0—different levels of factor F_i have no significant effects on the impact range of the accident; H1—different levels of factor F_i have significant impacts on the impact range of the accident. Table 3 shows the results of the intersubjective effect test, it can be seen that the P values of weather, peak time and intersection are less than 0.05, H=0, which indicates that weather, peak time and intersection have significant impacts on the impact range of the accident. The P value of the intersection is equal to 0.001, which statistically indicates that the impact of the intersection is extremely significant, because the intersection is a necessary place for vehicles to gather, turn and evacuate, so the accident at the intersection has a great impact on the traffic flow. Day/night, weekend, and season have no significant effects on the impact range of the accident.

Table 3. The intersubjective effect test

Number	F1	F2	F3	F4	F5	F6
P	0.016	0.035	0.625	0.155	0.113	0.001
H	0	0	1	1	1	0

Post-hoc test is to further determine the significances of differences among different categories within the factors that have a significant impact on the intersubjective effect test. Since the number of categories within the factors of post-hoc test is not less than 3, only the weather and peak hours are tested to observe their significances of the differences.

The least significant difference method (LSD) is the most sensitive post-hoc multiple comparison method with high test efficiency. The LSD method was used to conduct the post-hoc test for weather and peak hours. Set the significance level to 0.05, null hypothesis: H0—no difference; H1—there is a difference. The results of the post-hoc test in Table 4 show that there are significant difference between sunny day and rainy day, significant difference between sunny day and haze, no significant difference between sunny day and cloudy day, no significant difference between cloudy and haze, significant difference between cloudy day and rainy day, no significant difference between haze and rainy day, significant difference between off-peak time and morning peak time, no significant difference between evening peak and morning peak, and no significant difference between evening peak and non-peak. Generally speaking, the weather such as rain and fog will cause poor road traffic conditions, and the impact range of the accident of rain and fog weather will be larger than that of sunny and cloudy weather; and the traffic flow during the morning rush hour is large, the impact range of the accident during the morning rush hour is greater than that during the off-peak hour.

4 CONCLUSION

(1) The distribution of the impact range of the accident shows a phenomenon of many low values and right skewness, and could be well fitted by lognormal distribution, inverse Gaussian distribution and gamma distribution, and the lognormal distribution fitting effect is the best.

(2) Intersection, weather and morning peak have significant influences on the impact range of accident, and the influence of intersection is the most significant; Season, weekend, and day/night have no significant influences on the impact range of the accident.

(3) The effects of sunny day and rainy day on the impact range of the accident are significantly different, the effects of sunny day and haze on the impact range of the accident are significantly different, the effects of cloudy day and rainy day on the impact range of the accident are significantly different, while the effects of sunny day and cloudy day on the impact range of the accident are not significantly different, the effects of cloudy day and haze on the impact range of the accident are not significantly different, and the effects of haze and rainy day on the impact range of the accident are not significantly different. The effects of morning peak and off-peak on the impact range of the accident are significantly different, while the effects of evening peak and morning peak on the impact range of the accident are not significantly different, and the effects of evening peak and off-peak on the impact range of the accident are not significantly different.

(4) The research focus of this paper is on the influences of spatiotemporal and external conditions on the impact range of the accident, while the factors such as the type of accident vehicle and the traffic flow of the accident section have not yet been considered, and the subsequent research will consider other publicly available data sources and include them in the analysis of the impact range of the accident.

Table 4. Post-hoc test

Factor	Categories	Subcategories	P	H
Weather condition	sunny	cloudy	0.531	0
		haze	0.043	1
		rainy	0.020	1
	cloudy	sunny	0.531	0
		haze	0.140	0
		rainy	0.032	1
	haze	sunny	0.043	1
		cloudy	0.140	0
		rainy	0.203	0
	rainy	sunny	0.020	1
		cloudy	0.032	1
		haze	0.203	0
Peak time	off-peak	morning peak	0.004	1
		evening peak	0.135	0
	morning peak	off-peak	0.004	1
		evening peak	0.070	0
	evening peak	off-peak	0.135	0
		morning peak	0.070	0

Acknowledgment. we would like to thank the center for balance architecture of zhejiang university for funding this project (grant no. k-heng 20212792).

REFERENCES

[1] Wu, C. E, Luo, S. K., Wu L. M., and Xu, J.L., “Forecasting the severity of Freeway traffic accidents considering the spatio-temporal effect of traffic accidents,” Highways & automotive Applications (4), 22-27, 34(2023).

[2] Han, T. Y., Tian, S., Lv, K. G., Li, X., Zhang, J. T., and Wei, L., “Network analysis on causes for serious traffic accidents based on text mining,” China Safety Science Journal 31(9), 150-156(2021).

[3] Yu, G. Z., Liu, Y. M., Jin, M. J., and Wang, Y. P., “Traffic impact analysis of highway accident based on the shockwave theory,” Journal of Beijing University of Aeronautics and Astronautics 38(10), 1420-1424(2012).

[4] Cao, Z. Y., Guo, Z. Y., Zhang, Q. S., and Zha, X. D., “Research on time and spatial extent of terrible traffic accident on highway,” Highway Engineering 36(6), 55-58, 73(2011).

[5] Zhang, W. T., Wang, J. J., and Li, W. J., “Analysis of expressway traffic accident influence scope based on route choice,” Journal of Chang'an University (Natural Science Edition) 38(4), 87-94(2018).

[6] Chung, Y. and Recker, W. W., “Frailty Models for the Estimation of Spatiotemporally Maximum Congested Impact Information on Freeway Accidents,” IEEE Transactions on Intelligent Transportation Systems 16(4), 2104-2112(2015).

[7] Sun, C. S., Pei, X., Hao, J. H, Wang, Y. W., Zhang, Z., and Wong, S. C., “Role of road network features in the evaluation of incident impacts on urban traffic mobility,” Transportation Research Part B, Methodological 117, 101-116(2018).

[8] Sun, J. P., Guo, J. F., Zhang, X., and Xu, C. L., “Spatial and temporal distribution of occasional congestion based on speed variation,” Journal of Transportation Systems Engineering and Information Technology 19(2), 196-201, 215(2019).

[9] Tang, J. J., Liu, X. Y., Ji, K., and Ye, J. Q., “Estimation of traffic accidents impact on urban road network considering lane queuing characteristics,” Journal of Railway Science and Engineering 19(9), 2541-2551(2022).

[10] Eboli, L., Forciniti, C., and Mazzulla, G., “Factors influencing accident severity: an analysis by road accident type,” Transportation Research Procedia 47, 449–456(2020).

[11] AlKheder, S., AlRukaibi, F., Aiash, A., and Kader, A. A., “Weather risk contribution to traffic accidents types in Gulf Cooperation Council (GCC) countries,” Natural Hazards 114(2), 2177–2187(2022).

[12] Retallack, A. E. and Ostendorf, B., “Relationship between traffic volume and accident frequency at Intersections,” International Journal of Environmental Research and Public Health 17(4), article number 1393, 1–22(2020).

[13] Moosavi, S., Samavatian, M. H., Parthasarathy, S., and Ramnath, R., “A countrywide traffic accident dataset,” arXiv:1906.05409(2019).

[14] Chen, J., Wang, X. Y., Luo, L. L., and Cui, J. J.., “Comparison of machine learning and statistical learning in the imputation of missing values,” Statistics and Decision 36(17), 28-32(2020).