ISSN: 2688-8645

Research on Prediction of Urban Traffic Congestion Time and Construction of Traffic System based on Survival Analysis

Yindi Xu*

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China

*xyd2664736992@163.com

Abstract

With the development of economy and society, the number of cars in major cities in China has been rising sharply in recent years, and the traffic congestion problem has become increasingly serious. Aiming at the prediction of traffic congestion time, multiple linear regression and survival analysis methods are used to establish multiple linear model, Kaplan-Meier nonparametric regression model, Cox regression model and other prediction models. SPSS, MATLAB and other programming software are used to make the congestion index expression, and the internal relations among the influencing variables are discussed by combining the correlation coefficient matrix. Based on this, the probability distribution of congestion time is explored, and the results obtained have certain reference significance for formulating traffic policies, improving traffic trips and achieving the urbanization development goal based on smart growth.

Keywords

Traffic Congestion; Multiple Linear Regression; Time Prediction; Cox Regression; Traffic System.

1. Introduction

Traffic congestion is gradually evolving from single road sections or intersections to regional network congestion. The essence of traffic congestion is the imbalance of traffic demand, and the travel time can intuitively reflect the travel cost from the start point to the end point, which is an intuitive and effective traffic flow parameter to characterize the traffic congestion state of road sections. Time prediction is one of the core issues of modern transportation, and the accurate prediction results of time are the important basis for users to choose a reasonable travel route, which can effectively improve travel efficiency.

At present, in the research of traffic congestion, most scholars pay attention to the law of congestion occurrence, formation and spread, congestion charging and other issues. Jinsoo You and others established a mixed model of travel time prediction by using GIS technology [1]. Ramezani used Markov chain to accurately obtain the travel time distribution characteristics of trunk lines, so as to provide countermeasures and suggestions for alleviating traffic congestion [2]. Neila Bhouri used the traffic flow and traffic time of Paris Expressway as explanatory variables to construct traffic impact assessment indicators to study the impact of travel time variability [3]. Liu aimed at the uncertainty of users' travel time, based on degradable transportation network, constructed a multi-level time equilibrium model of users' demand, and priced according to users' elastic demand [4]. John Zaki defined the traffic state in the peak hours of two-dimensional space based on the hidden Markov model and the new model of contrast and used the average speed and contrast to predict the traffic congestion time [5].

To sum up, there are few researches on the characteristic distribution of congestion duration at present. However, the duration and distribution characteristics of traffic congestion will

ISSN: 2688-8645

directly affect citizens' travel route choice. Therefore, studying the distribution characteristics of traffic congestion duration can provide an important basis for travelers' route selection and traffic managers' traffic guidance. In this paper, multiple linear regression prediction and survival analysis model are used to construct traffic congestion prediction model, and the distribution characteristics of traffic congestion duration are studied in order to improve the speed and accuracy of traffic congestion prediction and provide more accurate services for travelers.

2. Data Acquisition and Hypothesis

The data used in this paper are all from the big data of population positioning and traffic flow released by Baidu Map Intelligent Traffic Business Center. In order to facilitate the analysis of the problem, this paper needs to make the following assumptions on the data used: (1) Assume that traffic congestion is caused by natural factors, and exclude vehicle congestion caused by uncontrollable factors such as natural disasters, traffic accidents and bad weather; (2) It is assumed that the external factors such as the size and performance of congested vehicles in the traffic flow are consistent, and the vehicles are running normally according to the rules and regulations, and the influence of other less influential vehicles is ignored; (3) It is assumed that all roads are straight one-way streets of main roads; (4) It is assumed that the traffic speed is all the main roads in Bengbu under the condition of smooth traffic, and the maximum speed limit is 40km/h.

3. Research on Congestion Time Prediction based on Multiple Linear Regression Model

3.1. Theoretical Basis

Multiple regression model is used to study the quantity problem between dependent variables (explained variables) and multiple independent variables (explained variables), and it can be used to predict and control the relationship between multiple variables [6]. In this paper, congestion index is selected as dependent variable, average travel speed, average delay time, morning peak, late peak and working day as independent variables to construct multiple linear regression model.

3.2. Research Methods and Ideas

In this paper, an empirical congestion time prediction model is obtained through statistical analysis, and the congestion time is quantitatively analyzed with congestion index as an index [7]. The optimization goal is to improve the completion degree and control the fitting degree of congestion index. Firstly, the valid data related to the variables to be analyzed are selected from the acquired data sets, then the index characteristics are analyzed and used as explanatory variables for multiple regression. Finally, the regression coefficients are analyzed and the above multiple linear regression equations are revised to further optimize the model.

Simplify the data of whether it is morning or evening peak or working day into variables of 0 and 1, that is, morning or evening peak (0), not morning or evening peak (1), working day (0) and not working day (1).

If the dependent variable is y and the independent variable is, the multivariate logarithmic linear regression model can be expressed as: $x_1, x_2, \cdots, x_m$, then this multiple log-linear regression model can be expressed as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_m x_m$$

ISSN: 2688-8645

Where b0 is a constant; Bi is the partial regression coefficient, which represents the independent variable when controlling the linear influence of other variables on the dependent variable. The linear regression model can be expressed as: $x_i(i=1,2,3\cdots m)$, the linear regression model can be expressed as

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_m x_{1m} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_m x_{2m} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_n x_{nm} + \varepsilon_n \end{cases}$$

The matrix form is, where, $y = \beta x + \varepsilon$,

$$y = \begin{bmatrix} \mathrm{y}_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, , , \ \mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{\mathrm{n}} \end{bmatrix}, \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{\mathrm{n}} \end{bmatrix}$$

Congestion index Y and average travel speed x_1, average delay time x_2, whether it's morning peak x_3, whether it's late peak x_4, whether it's working day x_5. Because of the low sampling rate of the existing GPS system, and because the microwave signal transmitted by the satellite to the terminal passes through the atmospheric ionosphere, troposphere and decoding the satellite positioning signal, the reported information of the GPS terminal will be delayed, so it's difficult and inaccurate to measure x_1. In order to avoid further influence on the accuracy of the model, the interaction between x_2 and x_5 is added to the basic model, and a regression model is established. After the interaction between and is added, the model is recorded as:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + a_5 x_5 + a_6 x_2 x_5 + \varepsilon$$

3.3. Analysis of Results

In this paper, the dependent variable congestion index and the independent variable average travel speed, average delay time, whether it's morning peak, whether it's late peak and whether it's working day are regressed, and the following congestion model is obtained:

$$y = -1.4197 + 0.0714\ x_1 + 1.5516\ x_2 - 0.0404\ x_3 + 0.0729\ x_4 + 0.0045\ x_5 - 0.0074\ x_6 .$$

The results obtained by using the statistical toolbox of are shown in Table 1:

Table 1. Fitting results of each parameter after the improved model and corresponding confidence interval table

Parameter	Parameter estimation	Parameter confidence interval
a0	-1.4197	[-1.5660, -1.2734]
a1	0.0714	[0.0675, 0.0754]
a2	1.5516	[1.4960, 1.6072]
a3	-0.0404	[-0.0610, -0.0198]
a4	0.0792	[0.0632, 0.0952]
a5	0.0045	[0.0002, 0.0089]
a6	-0.0074	[-0.0126, -0.0023]
R^2 =0.9550 F=3269.8523 p=0.0000 s^2=0.4730×10^{15}		

ISSN: 2688-8645

From the table, which indicates that the model has good applicability. According to the model, it can be estimated that during the weekend morning rush hour, the average travel speed measured by GPS in real time is 20km/h, and the congestion index when the average delay time is 2min is as follows:

$$\hat{y} = \hat{a}_0 + \hat{a}_1 \times 20 + \hat{a}_2 \times 2 + \hat{a}_3 \times 1 + \hat{a}_4 \times 0 + \hat{a}_5 \times 0 + \hat{a}_6 \times 2 \times 0 = 3.0307$$

In order to verify the prediction accuracy of the model, the fitting degree of all samples is analyzed, which is = 0.9550 and the fitting degree is 0.9550. It can be seen that the predicted value of the congestion index prediction model based on multiple linear regression in this paper has a good fitting result with the actual value of the congestion index.

4. Congestion Duration Model based on Survival Analysis

4.1. Theoretical Basis

The congestion duration model based on survival analysis follows the basic concept of survival analysis, and the survival time of traffic congestion refers to the time from the occurrence to the end of traffic congestion [7]; The data of traffic congestion duration refers to the incomplete data that the traffic congestion event occurred earlier than the start time of the study or the congestion continued after the end of the study time, or could not be accurately recorded due to some factors [8]; Traffic congestion survival function means that the traffic congestion survival function refers to the probability distribution of samples from the beginning of traffic congestion to the time t, and let t represent the congestion duration of roads or sections, which is a non-negative random variable.s(t).

The analysis of influencing factors of congestion duration in this paper mainly focuses on three aspects: parameter symbol, significance of influencing factors and relative risk ratio: (1) parameter symbol: when the significance test value is less than, the influencing factor is significant, which indicates that the influencing factor has obvious influence on congestion duration; When the significance test value is greater than, the influencing factor is not significant, indicating that the influencing factor has no obvious influence on the congestion duration. (2) Significance of influencing factors: when the coefficient is positive, the influencing factor is a risk factor. The greater the covariate value, the greater the risk rate and the greater the probability of congestion ending; When the coefficient is negative, the influencing factor is a protective factor. The bigger the covariate, the smaller the risk rate and the smaller the probability of congestion ending. (3) Relative risk rate ratio: it indicates the influence of variables on risk rate in favorable conditions relative to unfavorable conditions.

4.2. Research Methods and Ideas

In order to facilitate the establishment of the model, the morning peak time is set at 7: 20 ~ 8: 20, and the evening peak time is set at 17: 20 ~ 18: 20. After finding out the main influencing factors of congestion index, the probability distribution characteristics of congestion duration are further discussed by survival analysis. Firstly, the single value and numerical range of the research object are redefined, and the overall comparison is made. Secondly, Kaplan-Meier nonparametric regression model is used to directly estimate the danger function and survival function of traffic congestion, which reflects the influence of a single variable on the congestion duration. Then, Log-rank method is used to test the significance of the influencing factors, and the accurate distribution of congestion duration is obtained.

Let *f(t)* denote the probability density function of T, P represents the probability and the distribution function of T is:

ISSN: 2688-8645

$$F(t) = P(T \le t) = \int_0^t f(x)dx$$

Survival function *s(t)*, which represents the probability that the congestion duration is greater than T, and the expression is:

$$s(t) = P(T > T) = \int_t^{\infty} f(x)dx = 1 - F(t)$$

F(t) represents the distribution function; P represents probability; T represents the duration of traffic congestion; F(x) represents the probability density of t taking the value of time *x*. When the survival probability is low, the survival curve s(t) is steep, and when the survival probability is high, the survival curve s(t) is flat.

In the analysis of survival, T is described by the risk rate function h(t). The hazard function can be expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \frac{f(t)}{s(t)} = \frac{-d \ln s(t)}{dt}$$

The distribution of congestion duration will be different under different time and space factors, and the morning and evening peaks, working days and weekends will have a certain impact on the congestion duration. In order to explore the factors that affect the duration and distribution of congestion, the proportional risk rate of congestion duration is further established, and the relationship between the factors that affect the duration of congestion is discussed.

For the event that ends in time congestion on the risk rate set, the ending probability as the observed event is:

$$\frac{\exp[\sum_{j=1}^{p} \beta_j X_{j(i)}]}{\sum_{l \in R(t_{(i)})} \exp[\sum_{j=1}^{p} \beta_j X_{ji}} = \frac{\exp(\beta' X_{(i)})}{\sum_{l \in R(t_{(i)})} \exp[(\beta' X_l)]}$$

Multiply the conditional probabilities of all congestion durations to obtain a partial likelihood function:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp[\sum_{j=1}^{p} \beta_j X_{j(i)}]}{\sum_{l \in R(t_{(i)})} \exp[\sum_{j=1}^{p} \beta_j X_{ji}]} = \prod_{i=1}^{k} \frac{\exp(\beta' X_{(i)})}{\sum_{l \in R(t_{(i)})} \exp(\beta' X_i)}$$

$$LL(\beta) = \ln L(\beta) = \sum_{i=1}^{k} \sum_{j=1}^{p} \beta_j X_{ji} - \sum_{i=1}^{k} [\sum_{l \in R(t_j)} \exp(\sum_{j=1}^{p} \beta_j X_{ji})] = \sum_{i=1}^{k} \{\beta' X_{(i)} - \ln[\sum_{l \in R(t_{(i)})} \exp(\beta' X_i)]\}$$

The maximum likelihood estimation can be obtained from the above formula to get the estimated value of the parameters, and then a conclusion can be drawn.

4.3. Analysis of Results

The congestion time of Bengbu's main roads in the morning peak hours on weekends and weekdays is obviously longer than that in the off-morning peak hours. However, the difference between the congestion time at the late peak and that at the non-late peak is not high.

In the regular congestion data, there is a significant difference in the distribution of congestion duration between the flat peak and the morning and evening peak, followed by the weekday and weekend. Analyze the specific weekend working day variables and the distribution of the duration of the average peak, morning and evening peak.

ISSN: 2688-8645

(1) Comparison of the distribution of the duration of frequent congestion between the flat peak and the morning and evening peak.

Table 2. Quantile Table of Survival Rate of Frequent Congestion in Three Congestion Samples

	Broad congestion			General congestion			Strict congestion		
Survival quantile	25	50	75	25	50	75	25	50	75
Pingfeng	2	4	10	2	4	12	2	4	12
morning rush hour	2	6	12	2	4	10	2	4	10
evening peak	2	6	16	2	4	10	2	4	8

Comparing the survival rates of 25, 50 and 75 quantiles, the biggest difference between the flat peak and the morning and evening peak is the wide congestion event. In a broad sample of congestion, the duration of occasional congestion in the flat peak period is shorter than that in the morning and evening peaks, and that in the regular congestion in the morning and evening peaks is shorter than that in the flat peak period. In the broad congestion, the survival rate of congestion in the flat peak period is the smallest, while in the strict congestion sample, the survival rate of frequent congestion in the flat peak period is the largest.

(2) Comparison of the duration of frequent congestion on weekdays and weekends

Table 3. Quantile Table of Survival Rate of Frequent Congestion in Three Congestion Samples

	Broad congestion			General congestion			Strict congestion		
Survival quantile	25	50	75	25	50	75	25	50	75
Working day	2	4	10	2	4	10	2	4	10
weekend	2	4	10	2	4	10	2	4	8

In the chronic congestion, only in the strict congestion sample, there is a difference in the survival rate of 75 quantiles between weekdays and weekends.

Occasional congestion, because of its suddenness, contingency and unpredictability, will cause a series of changes in traffic flow that are different from the frequent congestion. As the occasional time often causes high-level congestion, the occasional congestion data in general congestion samples are used to study the characteristics of the occasional congestion time. In sporadic general congestion, there is no obvious difference between morning and evening peak and flat peak, while in sporadic strict congestion, there is no obvious difference between weekends and weekdays. Analyze the variables with different congestion durations.

(3) Comparison of occasional congestion time between weekend and weekday groups.

And the duration of occasional congestion is different between weekend group and weekday group. In the two congestion samples, the survival rate of occasional congestion on weekends is smaller than that on weekdays. When occasional incidents lead to congestion, it takes a process to deal with the incidents, and it takes different time to deal with the incidents with different severity. Obviously, the incidents that lead to moderate congestion and severe congestion are more severe, and the duration of occasional widespread congestion is different from that of occasional general congestion.

(4) Comparison of the duration of occasional congestion between flat peak and morning and evening peak.

There are significant differences in the duration of sporadic congestion between the peak and the morning and evening peaks between the broad congestion sample and the strict congestion sample. The survival rate of occasional congestion in the peak period is much smaller than that in the morning and evening peak periods. There is a huge traffic demand in the morning and

ISSN: 2688-8645

evening rush hour, and long-term serious traffic congestion is not allowed on the road. After the congestion occurs, corresponding measures will be taken in time to evacuate the traffic, so the congestion will end in time.

(5) The distribution difference of the duration of occasional congestion between the inner and outer rings.

Compare the sporadic inner and outer rings in general congestion samples and strict congestion samples. Congestion in the inner ring is more difficult to evacuate, and the inner ring has a higher survival rate.

The congestion duration model based on survival analysis is tested, and the results are shown in Table 4.

Table 4. Test Table of Model Coefficient

step	-2 times log-likelihood value	chi-square	significance level	chi-square	Sig.	chi-square	Sig.
1	5749.541	160.336	0.000	188.157	0.000	188.157	0.000
2	5737.257	177.946	0.000	12.284	0.000	200.441	0.000

In the above analysis, the survival time is taken as the response variable, and the Cox regression model with time-dependent covariates is used to solve the problem of time limitation. The significant difference is 0.000, and the results show that the model has a high explanatory degree.

5. Conclusion

The duration and distribution characteristics of traffic congestion will directly affect citizens' choice of travel routes. According to the traffic flow data of Bengbu, this paper adopts multiple linear regression prediction and survival analysis model, and establishes a traffic congestion prediction model by nonparametric method in survival analysis. The distribution law of road congestion duration in Bengbu is analyzed, and the main conclusions are as follows: the non-morning peak risk function on weekends and weekdays is greater than the early peak risk function, and the non-evening peak risk function is greater than the late peak risk function. The probability of peak danger function fluctuates greatly in the morning and evening. In the same duration of traffic congestion, the possibility of ending congestion in the morning and evening peaks is smaller than that in the non-morning and evening peaks. This paper studies the distribution characteristics of traffic congestion duration in order to improve the speed and accuracy of traffic congestion prediction and provide more accurate services for travelers. Due to the huge investment in urban construction and people's livelihood in Bengbu, the expenditure is high, and the output can't reach the expectation for a while, so the scale benefit has declined. With the passage of time, the income brought by these investments has been obviously improved, and the scale benefit has been continuously increased, and the urbanization development shows a good trend.

Acknowledgments

The research idea of this paper originated from the Undergraduate Research Innovation Fund Project of Anhui University of Finance and Economics (number: XSKY22233).

ISSN: 2688-8645

References

[1] Jinsoo You, Tschangho John Kim. Development and evaluation of a hybrid travel time forecasting model [J]. Transportation Research Part C,2000, 8(1).

[2] Mohsen Ramezani, Nikolas Geroliminis. On the estimation of arterial route travel time distribution with Markov chains [J]. Transportation Research Part B,2012,46(10).

[3] Neila Bhouri, Habib Haj-Salem, Jari Kauppila. Isolated versus coordinated ramp metering: Field evaluation results of travel time reliability and traffic impact[J]. Transportation Research Part C, 2013, 28.

[4] Bing-quan Liu, Chong-chao Huang. Multi-class time reliability-based congestion pricing model based on a degradable transportation network[J]. Applied Mathematical Modelling,2016,40(5-6).

[5] John F. Zaki, Amr Ali-Eldin, Sherif E. Hussein, Sabry F. Saraya, Fayez F. Areed. Traffic congestion prediction based on Hidden Markov Models and contrast measure[J]. Ain Shams Engineering Journal, 2019.

[6] wangxin, Wang Haiyi, Jiang Yuxuan, etc. Analysis of influencing factors of regional economic vitality based on multiple linear regression [J]. China Science and Technology Information, 2020 (09): 94-95.

[7] Liu Jiaqian, Zhu Jiaming. Dynamic optimal path solving algorithm of traffic network under uncertainty [J]. Journal of Bengbu University of Engineering Technology, 2016,30(03):246-251.

[8] Liu Menghan. Hierarchical traffic congestion evaluation model and algorithm for megacities [D]. Beijing Jiaotong University, 2009.