A Hybrid Approach for Urban Expressway Traffic Incident Duration Prediction with Cox Regression and Random Survival Forests Models*

Axiang Ke[1], Zhen Gao[1*], Rongjie Yu[2*], Min Wang[1], Xuesong Wang[2]
1.School of Software Engineering, Tongji University, Shanghai 201804, China
2. Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China

***Abstract*— Traffic incidents such as crashes have significant impacts on urban expressway operation. The roadside service and operational efficiency of urban expressways could be improved based on a well-developed incident duration prediction model. In this study, a hybrid approach that combines Cox regression and random survival forests algorithm is developed to establish incident duration analysis model. The study is conducted based on traffic incident data from Shanghai urban expressways. For each traffic incident, information about the road geometry, traffic operation, and weather conditions was collected for experiments, where 80% of sample is used for training and the rest 20% for validation. In the hybrid model, a Cox regression model is predeveloped to investigate and identify the significant contributing factors of incident duration. Then, these identified significant factors are used as inputs for the random survival forests model. Finally, the statistical measurements including mean absolute error (MAE) and normalized mean square error (NMS) are used to measure the model performance and compare with other models. The analysis results show that incident type, location, affected lane numbers and other attributes have significant impacts on incident duration, and the hybrid approach model provides better prediction accuracy over traditional traffic incident duration prediction methods.**

Keywords— Traffic operation management; Traffic incident duration prediction; Cox regression; Random survival forests; Urban expressway

I. Introduction

Rapid development of city transportation has brought great changes to the urban traffic infrastructure and urban expressway system has become an important means to meet the long-distance traffic demands. However, compared to the general urban roads and highways, the standard Chinese urban expressway design does not include emergency lanes. Without emergency lanes, the urban expressway tends to have more severe traffic congestion when traffic incidents occur[1]. There were in total 16,869 traffic incidents in the first half of 2013 in urban expressway of Shanghai, which means 92.8 average traffic incidents occurred daily[2]. Study shows that there is about 50-75% of the congestion caused by sudden events[3]. Therefore, developing an effective accidents impact analysis system is one of the important ways to improve the efficiency of urban expressway traffic operation.

As a quantitative indicator, traffic incident duration reflects the impact of traffic incidents on the extent of traffic objectively. Traffic incident duration prediction models have been widely adopted to produce important information for traffic operation management system, as necessary measures such as driver reroute and ramp control could be used to improve the efficiency of road operation[4]. Therefore, there's an urgent requirement to develop a traffic incident duration prediction model based on real-time traffic data, combined with road geometry and traffic event characteristics.

In terms of traffic incident duration analyses, various methodological and statistical modeling techniques have been developed and applied. Park and Haghani (2016) introduced Bayesian learning to neural networks for accurate prediction of incident duration[5]. Yang and Wang (2013) proposed a new forecasting model using Bayesian network and nonparametric regression for traffic incident duration[6].Smith (2002) presented and applied nonparametric regression and classification trees as models to predict the clearance time of a freeway accident[7]. Wei and Ying (2007) created an adaptive procedure for sequential forecasting of incident duration using ANN(Artificial Neural Network) models[8]. JiYang BB (2008) presented a prediction method of traffic incident duration of expressway, grounded on the Bayesian method-based decision tree classification algorithm[9]. The advantage of these models is that data can be directly used for model learning, without having to understand the details of the event. However, such models also have their own shortcomings, such as the neural network model requires a large number of parameters and learning time is longer, the decision tree model is easy to over-fitting.

Studies show that the distribution of traffic incident duration is not normally distributed, and often contains truncated data. Therefore, the traditional multiple regression analysis method is not applicable[10].Recently, survival analysis methods including hazard-based analysis, accelerated failure time(AFT) metric analysis and Cox regression analysis are gaining popularity in the transportation field. Survival analysis is a branch of statistics for analyzing the expected duration of time before consequent events happen, such as death of biological organisms[11] or failure in mechanical systems[12].Survival analysis has been widely used in engineering, economics and sociology. Hensher (1994) firstly proposed the hazard-based duration models in

*Corresponding authors,
Email: gaozhen@tongji.edu.cn; yurongjie@tongji.edu.cn

978-1-5090-5507-4/17/$31.00 ©2017 IEEE
ICIS 2017, May 24-26, 2017, Wuhan, China

transportation[13]. Nam (2000) applied hazard-based duration models to statistically evaluate the time it took to detect/report, respond to, and clear incidents[14]. Lin (2016) proposed a novel approach for accident duration prediction. In the approach the original M5P tree algorithm is improved through the construction of a M5P-HBDM(hazard-based duration model) model, in which the leaves of the M5P tree model are HBDMs instead of linear regression models[15]. Lee and Fazio(2005) and Yang (2014) used a Cox regression model to analyze the influential factors to the traffic incident duration[16],[17]. Alkaabi (2011)， Hojati (2013) and Li (2015) utilized the AFT metric model for accident duration prediction and investigated the effects of traffic accident characteristics on the accident duration[18],[19],[20].These studies provided a comprehensive and insightful understanding of survival analysis on traffic incident duration. These traditional survival analysis models quantitatively analyzed the effect of features on the probability of duration. However, they all relied on the restrictive assumptions including proportional hazards and may lead to biased prediction[21]. This present study aims to address these shortcomings by developing a hybrid approach which combine Cox regression and Random survival forests methods. Here we choose Cox regression because as an overwhelming survival analysis method it can handle multivariate analysis, regardless of the distribution of traffic incident duration, and can effectively use the censored data[10][22]. Random survival forests was first proposed by Ishwaran in 2008[23]. As an extension of random forests and survival analysis, it is not established based on any certain assumptions, and hence avoids the problems of high variance and bias. Random survival forests model has already become a prevalent method for various problems in medicine[24] and economics[25], while it is rarely used in transportation despite traditional random forests model has been used for predicting incident duration and got a good reputation [26].

In this study, a hybrid approach is developed to combine Cox regression and random survival forests methods for comprehensively analyzing traffic incident duration based on traffic incident data of Shanghai urban expressways. The experiments results demonstrate that the proposed hybrid approach performs reasonably well.

This paper is organized as follows. The data description and pre-processing procedure are provided in Section 2, followed by the methodology description with de-tailed model structures in Section 3. The model estimation and discussions are presented in Section 4, and the research is concluded in Section 5.

II. Data Description

This study is conducted based on one-month traffic incident records collected from Shanghai urban expressways in April 2014. Four major datasets are used in this study: incident data, road geometry data, traffic operation data, and weather conditions data, including incident types, locations, occurrence time, the ramp type, the length between two ramps of road, roadway geometric characteristics, weather conditions and so on. The incident data is obtained from Shanghai urban expressway monitoring center; the road geometry data is got from Shanghai urban expressway GIS maps; the traffic operation data is aggregately computed from loop detectors and weather conditions is searched from public website. Finally, a total of 1,931 sample records are sorted out for model development with incomplete, duplicate and abnormal data filtered out to enhance the data quality.

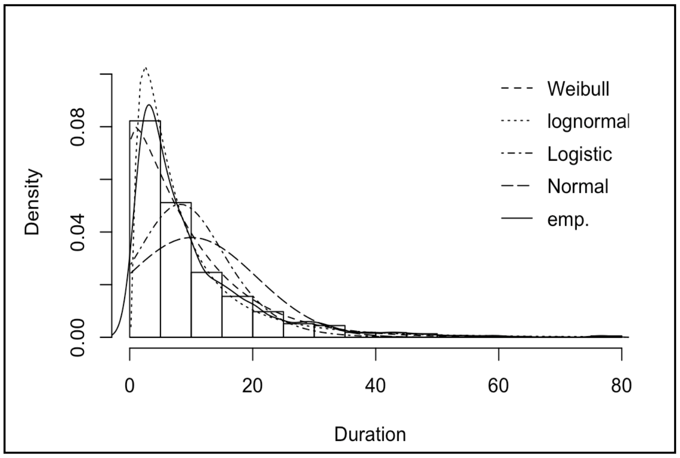
The sample is divided randomly into training data (80% of sample) and validation data (20% of sample). The descriptive statistics for training data and validation data are provided in Table I from which it is clear that training data and validation data have similar distribution. The distribution of traffic incident duration is provided in Figure 1. The emp. mean that the empirical density of traffic incident duration.

TABLE I. DESCRIPTIVE STATISTICS OF TRAFFIC INCIDENT

Index	Training data	Validation data
Number of incident	1545	386
Average Duration	10.03	10.18
Max Duration[a]	78.07	76.47
Min Duration[a]	0.2	0.25
25%Quantile Duration[a]	3.18	3.18
50%Quantile Duration[a]	6.35	6.71
75%Quantile Duration[a]	13.33	12.8

[a] The unit of duration is minute.

Fig. 1. Number of Variables for Each Node Split



III. Methodology

Firstly, Cox regression model is developed to investigate and identify significant contributing factors to traffic incident duration using the training data. Subsequently, the identified factors are utilized to establish a random survival forests model. Finally, the accuracy of the model is validated against the validation data based on the statistical measurements including MAE, and NMSE.

A. Cox Regression and Features Selection

Cox regression was proposed by Cox D R in 1971[22]. It is the most commonly used multivariate approach for analyzing survival time. In a Cox regression model, the unique effect of an unit increase in a covariate is multiplicative with respect to the hazard rate, which is the risk of failure, given that the event has persisted to a specific time[27]. Unlike other statistical models, the survival time is not assumed to follow a particular statistical distribution in Cox regression model. What's more, it can quantitatively analyze the intensity and direction of the hazard factors.

The hazard function for the Cox regression model is listed in equation (1),

$$h(t) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m\right) \quad (1)$$

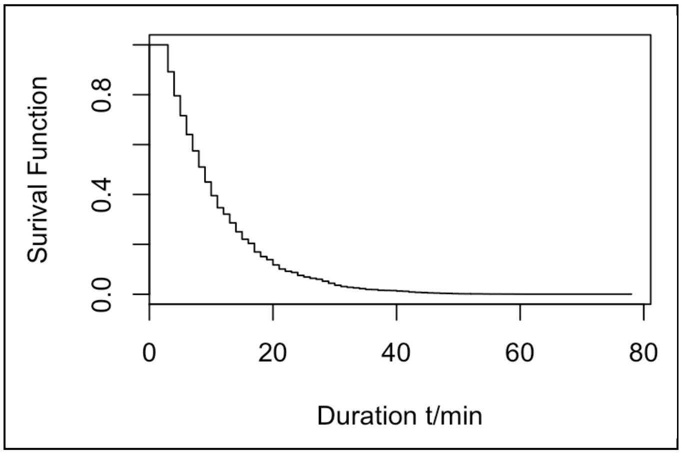
where the hazard function *h(t)* is dependent on a set of m covariates($x_1, x_2, \ldots, x_m$), whose impact is measured by the size of respective coefficients($\beta_1, \beta_2, \ldots, \beta_m$). The term h_0 is called the baseline hazard, and is the value of the hazard when all the xi equals zero (the quantity exp(0) equals 1). The '*t*' in *h(t)* reminds us that the hazard may vary over time. And the baseline hazard function is estimated non-parametrically. The survival function and hazard function for traffic incident duration are presented in Figure 2 and Figure 3.

The log of the hazard ratio(HR), i.e. the hazard function divided by the baseline hazard function at time *t*, is a linear combination of parameters and regressors, i.e.,

$$ln\frac{h(t)}{h_0(t)} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m \quad (2)$$

In this study, the Cox regression model is adopted to identify significant features. All the significant variables are identified based on their P-values at the significant level of P=0.05. These significant features will be used for random survival forests model creation.

Fig. 2. Survival function for traffic incident duration



B. Random Survival Forests for Regression

Random survival forests model is employed to predict incident duration using the significant factors identified in the Cox regression model. The model uses a set of bootstrap samples, developing an independent tree model on each sub-sample of the population. Each tree is developed by recursively partitioning the population based on optimization of a splitting rule over the p-dimensional covariate space. At each split, a subset of m⩽p candidate variables are tested for the split rule optimization, dividing each node into two children nodes. Each children node is then split again until the process reaches the stopping criteria of either node purity or node member size, which defines the set of terminal nodes for the tree[28]. In regression tress, the split rule is based on minimizing the MSE(Mean Squared Error) calculated as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i\text{-}y_i)^2 \quad (3)$$

where p_i represents the prediction value, y_i is the actual value.

One of the advantages of Random Survival Forests is that it has a built in generalization error estimate. Each bootstrap sample selects approximately 63.2% of the population on average. The remaining 36.8% of observations, called the OOB(Out of Bag) samples, can be used as a hold out testing set for each of the trees in the forests. An OOB prediction error estimate can be calculated for each observation by predicting the response over the set of trees which were not trained with that particular observation. Mathematically, the OOB is calculated as

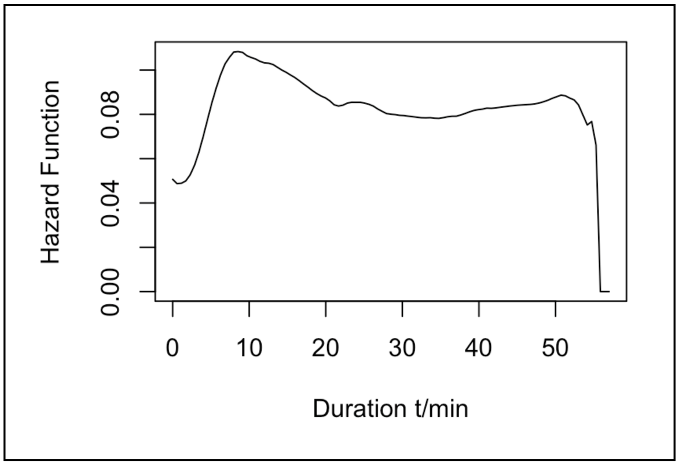
$$P = (1 - \tfrac{1}{N})^N \quad (4)$$

If $N \to \infty$, then

$$\lim_{N\to\infty} P = \lim_{N\to\infty}(1 - \tfrac{1}{N})^N = e^{-1} \approx 0.368 \quad (5)$$

where N is the number of original data.

Fig. 3. Hazard function for traffic incident duration



C. *Model Performance Evaluation*

The statistical measurements including MAE and NMSE are used to measure the model performance. These statistical measurements are shown in formula (6) and (7).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |p_i\text{-}y_i| \tag{6}$$

$$NMSE = \frac{\frac{1}{n}\sum_{i=1}^{n}(p_i\text{-}y_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{n}\sum_{i=1}^{n}y_i\text{-}y_i)^2} \tag{7}$$

IV. MODELING RESULT AND DISCUSSION

A. *Significant Features Selection*

The significant features are identified using the Cox regression model. Based on AIC estimation method, the significant features are identified based on their P-values at the significant level of P≤0.05. Table III shows the effect sizes, 95% confidence intervals (CI), regression coefficients, HR, P-values and so on. The selected significant features are shown in Table III. There are seven factors including incident type, location, number of lanes affected, weather, length between two ramps of the road where incident occurred, coefficient of variation of speed, and road curvature that significantly affect the incident duration.

These factors are analyzed in detail as below. For the variable IncidentType, as seen in Table II, compared to two vehicles collision type and other incident type, single car broken down condition incurs the longest duration. The variable, RelatedLanes2, denoting the number of lanes affected, has a negative coefficient, showing that the duration of traffic incident that involves more than 1 lane is longer than that involves only 1 lane. For environmental conditions, variable weather1 has a negative coefficient, indicating that the duration in raining days is longer than that in other weather conditions, it may due to slower processing of incidents in the rain. For traffic conditions, the coefficient of variable of speed has a negative coefficient, meaning that high speed deviation is more prone to resulting in high accident severity, and the duration will be prolonged accordingly. For expressway geometric characteristics, variable Length denoting the length between two ramps of the road where incident occurred, has a negative coefficient, which means the duration becomes longer when the length is longer. Reason might be that more traffic vehicles will accumulate in the longer road segment.

B. *Comparison of Hybrid Model and Other Popular Models*

In our study, the incident duration prediction model (named hybrid model) is built using random survival forests algorithm with the features selected by Cox regression model. The model prediction is validated respectively on training data and validation data. The results are shown in Table IV, in which the prediction results of pure Cox regression model. Here pure model is trained using all the features which cover the features of the hybrid model.

In Table III, the accuracy of the hybrid model with five minutes tolerance is 64.79% for training data and 77.20% for validation data respectively. Compared to the accuracy of the pure Cox regression model, the hybrid model has yields significant improvements. for validation data, the MAE is 3.66 for the hybrid model but 6.38 for the pure Cox regression model. The NMSE is 0.28 for the hybrid model but 0.98 for pure Cox regression model. The smaller value the statistical measurements MAE and NMSE are, the greater the model performs. All these measurements prove the quality of the hybrid model is much greater than the pure Cox regression model.

V. CONCLUSION

In this study, a hybrid approach is proposed to integrate the Cox regression model and the random survival forests model to predict traffic incident duration. Firstly, the Cox regression model is developed to identify significant features and then the random survival forests model is employed to conduct prediction. Two statistical performance measures including MAE and NMSE are used to quantify the model performance. The results demonstrate that the proposed hybrid approach performs much better than the pure Cox regression model[10]. Furthermore, the excellent generalization ability of the hybrid model is proved using ten-fold cross-validation. The hybrid model is unexpectedly more accurate to predict traffic incident duration for previously unseen data.

An inference analysis is conducted to quantify the feature’s contributions to traffic incident duration, and the selected features conclude incident type, location, number of lanes affected, weather, length between two ramps of the road where incident occurred, coefficient of variation of speed, and road curvature based on Cox regression model. It is found that the four features including incident type, number of lanes affected, weather, length between two ramps of the road where incident occurred have the most significant impact on the traffic incident duration. Besides, the coefficient of variation of speed is discovered to have a great impact on the duration, so the traffic management department may effectively shorten the traffic incident duration by real-time observation of traffic conditions and accordingly inducing traffic behavior via changing velocity limit.

The proposed methodology and research findings provide insights for developing effective countermeasures such as driver reroute and ramp control to reduce the congestion caused by traffic incidents and as a result improve the efficiency and safety of urban expressways.

ACKNOWLEDGMENT

This study was jointly sponsored by the National Natural Science Foundation of China (No. 71401127) and Science and Technology Commission of Shanghai Municipality (No. 15DZ1204800).

TABLE II. COEFFICIENTS ESTIMATION OF COX REGRESSION MODEL

Variables	Description	Coefficients	HR	95%Confidence Interval	P-value	Count	Average/Standard Deviation
CV_Speed	Coefficient of variation of speed	-0.6445	0.5249	(0.3411, 0.8078)	0.0034	-	0.1758/0.1125
IncidentType2	Incident type is other	-0.35	0.7047	(0.605, 0.8208)	<0.0001	212	-
IncidentType3	Single car broke down	-0.6363	0.5292	(0.476, 0.5884)	<0.0001	571	-
Weather1	Rainy	-0.1726	0.8414	(0.7258, 0.9755)	0.0221	205	-
RelatedLanes2	More than 1 lane	-0.6197	0.5381	(0.4234, 0.6838)	<0.0001	75	-
Place2	Middle ring	-0.2682	0.7647	(0.6601, 0.886)	0.0004	430	-
Length	Length between two ramps of road	-0.0002	0.9998	(0.9997, 0.9999)	<0.0001	-	1095.299/511.8952
Curve1	Road has curve	0.1257	1.134	(1.0343, 1.2433)	0.0074	1038	-

TABLE III. Performance evaluation of the model

Evaluation	Training dataset(1545)				Validation dataset(386)			
	Cox regression model		Hybrid model		Cox regression model		Hybrid model	
	Value	Percentage	Value	Percentage	Value	Percentage	Value	Percentage
MAE	6.36	-	5.07	-	6.38	-	3.66	-
NMSE	0.94	-	0.47	-	0.98	-	0.28	-
Prediction error≤5min	938	60.71%	1001	64.79%	234	60.62%	298	77.20%
Prediction error ≤10min	1278	82.71%	1388	89.84%	322	83.42%	367	95.08%
Prediction error ≤15min	1509	97.67%	1478	95.66%	349	90.41%	378	97.93%

REFERENCES

[1] Beijing General Municipal Engineering Design and Research Institute. Specification for Design of Urban Expressway. Beiging: Ministry of Housing and Urban-Rural De-velopment of the People's Republic of China, 2009.

[2] Sun W S. Cause and Suggestion of Frequency Congestion in Shanghai Expressway. China Municipal Engineering, 2014, (3): 9-11

[3] Giuliano G. Incident characteristics, frequency, and duration on a high volume urban freeway. Transportation Research Part A General, 1989, 23(5):387-396.

[4] Jiyang B B, Zhang X N, Sun L J. A Review of the Traffic Incident Duration Prediction Methods. Highway Engineering, 2008, 33(3): 72-79.

[5] Park H, Haghani A, Zhang X. Interpretation of bayesian neural networks for predicting the duration of detected incidents[J]. Journal of Intelligent Transportation Systems, 2016.

[6] Yang C, Wang C. Traffic Incident Duration Forecast Model of Expressway. Journal of Tongji University (Natural Science), 2013, 41(7): 1015-1019.

[7] Smith K W, Smith B L. Forecasting the Clearance Time of Freeway Accidents. Traffic Congestion, 2002.

[8] Wei C H, Ying L. Sequential forecast of incident duration using Artificial Neural Network models. Accident Analysis & Prevention, 2007, 39(5):944-54.

[9] Jiyang B B, Zhang X N, Sun L J. Trafic IncidentDuration Prediction Grounded on Bayesian Decision Method-Based Tree Algorithm. Journal of Tongji University (Natural Science), 2008, 36(3): 319-324.

[10] Kang Guoxiang, FANG Songen. Application of Cox Regression Model in Traffic Incident Duration[J]. Journal of Transport Information and Safety, 2011, 29(2):104-106.

[11] Györffy B, Lanczky A, Eklund A C, Denkert C, Budczies J, & Li Q, et al.An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Research & Treatment, 2010, 123(3):725-731.

[12] Luoma M, Laitinen E K. Survival analysis as a tool for company failure prediction. Omega, 1991, 19(6):673-678.

[13] David A. Hensher, Fred L. Mannering. Hazard‐based duration models and their application to transport analysis. Transport Reviews, 1994, 14(1):63-82.

[14] Nam D, Mannering F. An exploratory hazard-based analysis of highway incident duration. Transportation Research Part A Policy & Practice, 2000, 34(2):85-102.

[15] Lin L, Wang Q, Sadek A W. A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations.[J]. Accident Analysis & Prevention, 2016, 91:114-126.

[16] Lee J T, Fazio J. Influential factors in freeway crash response and clearance times by emergency management services in peak periods. Traffic Injury Prevention, 2005, 6(4):331-339.

[17] Yang W C, Zhang L, Shi Y C, Yang T. Survival analysis of traffic incident duration for urban expressway. Transportation system engineering and information technology, 2014, 14(5): 168-174.

[18] Saeed A M, Dissanayake D, Bird R N. Analysing clearance time of urban traffic accidents in Abu Dhabi using hazard-based duration modelling method. 2011.

[19] [19] Hojati A T, Ferreira L, Washington S, Charles P. Hazard based models for freeway traffic incident duration. Accident; analysis and prevention, 2013, 52C(12):171-181.

[20] Li R. Traffic incident duration analysis and prediction models based on the survival analysis approach[J]. Iet Intelligent Transport Systems, 2015, 9(4):351-358.

[21] Dinse G E, Jusko T A, Ho L A, Annam K, Graubard B I, Hertzpicciotto I, et al. (2014). Accommodating measurements below a limit of detection: a novel application of cox regression. American Journal of Epidemiology, 179(8), 1018-1024.

[22] Cox D R. Regression Models and Life-Tables. 1972, 34(2):527-541.

[23] Ishwaran H, Kogalur U B, Blackstone E H, Lauer M S. Random Survival Forests. The Annals of Applied Statistics, 2008, 2(3): 841-860.

[24] Hsich E, Gorodeski E Z, Blackstone E H, Ishwaran H, Lauer M S (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circulation Cardiovascular Quality & Outcomes, 4(1), 39-45.

[25] Fantazzini D, Figini S. Random Survival Forests Models for SME Credit Risk Measurement. Methodology & Computing in Applied Probability, 2009, 11(1):29-45.

[26] YANG Chao, LI Haixia. Estimation of the Duration of the Incidents at Urban Expressways Using Random Forest[J]. Journal of Transport Information and Safety, 2015, 33(6): 72-76

[27] Bradburn M J, Clark T G, Love S B, Altman D G.Survival Analysis Part II: Multivariate Data Analysis - an Introduction to Concept and Methods. British Journal of Cancer, 2003, 89(3):431-6.

[28] Song Q Q, Wu X Y, Hou Y, et al. Survival Analysis of High Dimensional Genomic Data Using Random Survival Forests. Chinese Journal of Health Statistics, 2013, 30(6): 786-789