Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 6312139, 14 pages
https://doi.org/10.1155/2022/6312139

Research Article

Analysis and Prediction of the Interval Duration between the First and Second Accidents considering the Spatiotemporal Threshold

Fang Liu,[1] **Lanlan Zheng**,[2] **Mingyang Li**,[2] **and Jinjun Tang**[2]

[1]*School of Transportation Engineering, Changsha University of Science and Technology, Changsha 410205, China*
[2]*Smart Transportation Key Laboratory of Hunan Province, School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China*

Correspondence should be addressed to Mingyang Li; 194211046@csu.edu.cn

Received 10 June 2021; Accepted 11 January 2022; Published 7 February 2022

Academic Editor: Alain Lambert

Studying the time interval duration between the first accident and the second accident caused by it can provide decision makers with valuable information on how to effectively deal with high-risk second accidents. This paper is aimed to explore the potential influencing factors of the interval duration between the two accidents and predict it. First, the spatiotemporal definition method is applied to identify the cascaded first accident and the second accident. Then, on the basis of using Kaiser-Meyer-Olkin (KMO) measure and Bartlett's sphere test statistics to ensure the applicability of the data to the factor analysis method, the explanatory variables that can significantly affect the interval duration are obtained through the factor analysis method. Finally, the random forest model (RF), which combines the advantages of machine learning methods, is employed to predict the duration of the interval. Traffic accident data set collected in Los Angeles city from February 2016 to June 2020 is used to validate prediction performance in this study. Bayesian method is applied to optimize the hyperparameters in the RF, while three evaluation indicators, including the Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE), are used to estimate the prediction effect. The test results and comparative experiments confirm that RF is able to predict the interval well and has better prediction performance. This is of great significance for the prediction of the duration of the interval between one accident and the second accident.

1. Introduction

Road traffic accidents can be caused by motor vehicles and nonmotor vehicles [1, 2], and their impacts are also uncertain. Due to the peculiarities of the road, the dangerous traffic conditions caused by the first accident usually expose unattended vehicles and persons to extra risks. This issue may lead to a second accident. The risk of a second accident is estimated to be six times that of the first accident [3]. The huge negative consequences caused by the second accident make it another issue of concern for road traffic accidents to be widely studied.

Raub [4] proposed that any crash that occurred within one mile from the scene of an accident with an event lasting more than 15 minutes is considered to be related to the original event, and this accident is called a second accident cascaded with the first accident. The 15-minute threshold is based on the escape time provided by the related research of Lindley and Tignor [5], that is, the time that may affect the traffic operation after the accident. The one-mile distance is derived from the observation of accidents that occurred during the period of maximum traffic flow. Karlaftis et al. [6] applied the predefined identification parameters of time and distance proposed by Raub to identify secondary accidents. Hirunyanitiwattan and Mattingly [7] considered 60 minutes and 2 miles upstream as the thresholds, but Moore et al. [8] set the thresholds of time and space as 2 hours and 2 miles on the Los Angeles highway. Zhan et al. [9] calculated the thresholds based on the different lane congestion assumptions in the "Expressway Capacity Manual," using accident recovery time of 33.34–52.6 minutes, event dissipation time of 0–21.76 minutes, and maximum queue length of 1.09–1.49 miles as thresholds.

The above studies all used the static spatiotemporal threshold method to identify second accidents. The performance of this static method mainly depended on the thresholds and their applicability to the study area. Sun [10] proposed an improved dynamic threshold method that can extract second accidents from the event database. The dynamic threshold was derived from the initial accident progress curve. The dynamic method described in the study of Sun and Chilukuri [11] improved the existing static method by using the event progress curve to mark the end of the change queue during the entire event. Moreover, some studies have also proposed speed-based methods to determine the time and space range of major events or classify second accidents [12–20].

Not only focus on the identification of second accidents, but also the prevention or rescue of second accidents. Se-Ryong et al. [21] researched the second accident in the tunnel, and they concluded the concrete barriers are suitable to reduce the risk of the second accident. Aoki et al. [22] developed a new robot called "QRoSS." It can replace humans to complete some dangerous second accident rescue missions. Kostikova et al. [23] analyzed the factors of second accidents through data from in-depth accident analysis. Pietila et al. [24] studied determinants of recurring occupational accidents, and it can be found that the substantial reoccurrence of occupational accidents emphasizes the importance of assessing the prevention policies after each accident. Kim et al. [25] propose a road stud system incorporating a wireless control function using RF-based communication with existing solar LED road studs and a system for controlling them. It can be possible to prevent secondary accident after accident.

Although scholars have made a lot of contributions to the study of second accidents, they have not explored the relationship between the time and spatial threshold and the identification of secondary accidents. There is also a lack of research on the prediction of the time between the first and second accidents. In order to solve these problems, this study considers the influence of the first accident spatiotemporal impact threshold on the second accident and proposes a static spatiotemporal threshold definition method based on sensitivity analysis to identify accident pairs. At the same time, consider the influence of the duration of the first accident, explore, and analyze the duration of the interval between two accidents. Thus, it provides more comprehensive and accurate accident information for traffic management and a scientific basis to avoid accidents.

The remaining of this research is organized as follows: Section 2 introduces the data source. Section 3 presents the framework of this work and the methods used. The result and discussion are outlined in Section 4. Section 5 is the conclusion and prospect.

2. Data Description

This study is based on data analysis, which is getting more and more attention to be applied in various research in the field of transportation [26–29]. According to statistics on traffic accidents for five consecutive years (2016–2020) in various cities of the United States, the number of accidents in Los Angeles city ranked first, with 11798, 11388, 11309, 8705, and 1716, respectively, accounting for 29.4%, 29.3%, 29.1%, 30.4%, and 31.3% in five years. Therefore, in order to clarify the potential mechanism of accidents in Los Angeles city, we selected 2016–2020 road traffic accidents in Los Angeles as the data set. The details of the data source are shown in Table 1.

3. Methodology

3.1. Spatiotemporal Definition. The main idea of the spatiotemporal definition method is to treat an accident that occurs within a given time threshold and space threshold from the first accident as a second accident cascaded with it. The mathematical model is described as follows:

$$\mathrm{SC} = \begin{cases} 0, & \text{other,} \\ 1, & \text{if } \left[t_c \in \left(t_p, t_p + \Delta t\right)\right] \text{ and } \left[s_c \in \left(s_p, s_p + \Delta s\right)\right], \end{cases} \tag{1}$$

where t_c is the time point when the first accident occurred, SC is the space point where the first accident occurred, Δt and Δs are the time and space threshold of the spatiotemporal definition method. 1 means that the accident is identified as a second accident; otherwise, it is 0.

3.2. Factor Analysis. In this paper, the factor analysis method is used to identify the influencing factors in the interval duration prediction, analyze the correlation between the respective variables and the dependent variables, and extract common factors. The public factor set formed by this method is applied to represent the accident information without affecting the prediction. On the one hand, the complexity of the model is simplified without affecting the effect of the predictive model. On the other hand, variables that can significantly affect the duration of the interval can be explored.

3.2.1. Correlation Coefficient. This study employs the Pearson coefficient to quantify the closeness of the factors. The coefficient is defined as the quotient of the covariance and standard deviation between two variables. The calculation formula is as follows:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X \sigma_Y} = \frac{E(\mathrm{XY}) - E(X)E(Y)}{\sqrt{E\left(X^2\right) - E^2(X)}\sqrt{E\left(Y^2\right) - E^2(Y)}}, \tag{2}$$

Table 1: Description of data properties.

Property	Description
Number	Unique identifier for incident record
Source	Source of incident report (accident API)
Information channel code	Provide a more detailed description of the event
Collision level	There are 4 levels, the greater the value, the greater the impact on traffic
Start time	Accident start time
End time	Accident end time
Longitude of occurrence	The longitude shown in GPS coordinates at the beginning of the accident
Latitude of occurrence	The latitude shown in GPS coordinates at the beginning of the accident
End longitude	The longitude shown in the GPS coordinates at the end of the accident
End latitude	The latitude shown in the GPS coordinates at the end of the accident
Distance	The length of the road area affected by the accident
Characterization	Natural language description of the accident
Figure	Address bar street number
Street	Street name in the address bar
Side	The address bar shows the opposite side of the street (left/right)
City	City displayed in the address bar
County	County displayed in the address bar
State	State shown in the address bar
Postcode	Postal code displayed in the address bar
Country	Country shown in the address bar
Time zone	The location of the accident shows the time zone
Airport code	Airport weather station closest to the accident site
Weather timestamp	Show the time stamp of the meteorological observation record
Temperature	Temperature (in degrees Fahrenheit)
Wind chill	Wind chill (in degrees Fahrenheit)
Humidity	Humidity (percent)
Pressure	Air pressure (in inches)
Visibility	Visibility (in miles)
Wind direction	Wind direction
Wind speed	Wind speed (in miles per hour)
Precipitation	Precipitation (in inches)
Weather conditions	Weather conditions (rain, snow, thunderstorm, fog, etc.)
Conveniences	The presence comfort indicated by the POI annotation is in a nearby location
Deceleration zone	Speed bump or hump nearby
Intersection	There is an intersection nearby
Yield	Yield nearby
Junction	There is a junction nearby
No exit	There is no exit sign nearby
Railway	There is a railway nearby
Roundabout	There is a roundabout nearby
Station	There is a station nearby
No thoroughfare	There are stops nearby
Traffic light	There are traffic lights nearby
Turn	Turn signs nearby
Sunrise sunset	Show day or night according to sunrise/sunset

where Cov is covariance, σ is the standard deviation. It can be seen from formula (2) that the Pearson coefficient is meaningful if and only if the standard deviations of the two variables are not 0.

3.2.2. Applicability Analysis. If the variables have no correlation or the correlation is low, there is no common factor between these variables. Therefore, only when there is a strong correlation between the variables, the data can use factor analysis with a few false variables instead of objective explanatory variables. In this study, the KMO measure and Bartlett's sphere test are used to test the applicability of factor analysis of data.

(1) KMO Measurement. The KMO measurement is a comprehensive index that takes into account the correlation coefficient and partial correlation coefficient of variables. The calculation formula is as follows:

$$\mathrm{KMO} = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} p_{ij}^2}, \tag{3}$$

where r_{ij} is the correlation coefficient between variables i and j, and p_{ij} is the partial correlation coefficient between variables i and j. When the correlation coefficient is much larger than the partial correlation coefficient, the KMO measure is close to 1; otherwise, the KMO measure is close to 0. That is, the KMO measure is between [0, 1]. The more it is close to 1,

the correlation is stronger, the partial correlation is weaker, and the effect of factor analysis is better; when it is less than 0.5, the correlation is low and factor analysis is not applicable.

(2) Bartlett Sphere Inspection. The Bartlett sphere test judges whether variables are independent based on data correlation and make the null hypothesis that the correlation coefficient matrix is a unit matrix. If the value of the test statistic is large and the corresponding associated probability value is less than the significance level (0.05) given by the study, the null hypothesis is rejected; otherwise, the null hypothesis is accepted and the correlation coefficient matrix is approximately a unit matrix, indicating that the variables may provide some information independently and lack common factors, which is not suitable for factor analysis.

3.2.3. Mechanism of Factor Analysis. The factor analysis method explores the relationship between the original variables [30], converts multiple variables of the original data into several common factors that can express the dependence of the data, eliminates the overlap of information between the variables to a certain extent, and reduces the intrinsic relevance [31]. In the method of factor analysis, factors are abstract concepts and only serve as symbols. The mathematical model is described as follows.

Assuming the original variables x_i $(i = 1, 2, 3, \ldots, p)$ and standardizing them to obtain new variables z_i, the factor analysis model is expressed as follows:

$$z_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{\mathrm{im}}F_m + c_iU_i\,(i = 1, 2, 3, \ldots, p). \tag{4}$$

Among them, F_j $(j = 1, 2, \ldots, m)$ is the common factor; U_i $(j = 1, 2, \ldots, p)$ is only related to the variable z_i and is called the special factor; the coefficient a_{ij} and c_{ij} refer to the factor loading, and $A = (a_{ij})$ is called the factor loading matrix. Then, the above formula can be expressed as the following matrix form:

$$z = \mathrm{AF} + \mathrm{CU}, \tag{5}$$

where $z = (z_1, z_2, \ldots, z_p)^T$, $F = (F_1, F_2, \ldots, F_m)T$, $U = (U_1, U_2, \ldots, U_p)^T$, $A = (a_{ij})_{p*m}$, $C = \mathrm{diag}\,(c_1, c_2, \ldots, c_p)$.

3.2.4. Factor Rotation. The factor loading matrix is not unique, so it is necessary to rotate the factor loading moment [32]. This is helpful that the square value of each column or row of the loading matrix is differentiated to two levels of 0 and 1 and can simplify the factor loading matrix.

This study uses the maximum variance method for factor rotation. On the basis of the initial load matrix, the transformation method of the factor load matrix is obtained according to the simple structure criterion so that the variance of the square value of each column element of the transformed factor load matrix is kept independent of each other. At this time, a few variables have higher loading values on the factors, which can explain the composition of common factors.

3.2.5. Factor Score. After the load matrix is rotated, its factor score function is defined as follows:

$$F_j = \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{\mathrm{jp}}X_p, \quad j = 1, 2, 3, \ldots, m. \tag{6}$$

It can be seen from the above formula that the coefficient of the score function can be calculated to obtain the score of each factor. Since $p > m$, an accurate score cannot be obtained, and only an estimated value of the score can be obtained [33].

Through the Bartlett factor score, use the weighted least squares method to finish estimating. Regarding x_i-u_i as the dependent variable, the factor loading matrix is regarded as the observation of the independent variable, decomposed as follows:

$$\begin{cases} x_{i1} - \mu_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_{i2} - \mu_2 = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \cdots \quad \cdots \\ x_{\mathrm{ip}} - \mu_p = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{\mathrm{pm}}f_m + \varepsilon_p \end{cases}. \tag{7}$$

Because the variances of the special factors are different, the weighted least squares method is used to find the score, so that

$$\frac{\sum_{j=1}^{p}\left[\left(x_{\mathrm{ij}} - \mu_i\right) - \left(a_{i1}\widehat{f}_1 + a_{i2}\widehat{f}_2 + \cdots + a_{\mathrm{im}}\widehat{f_m}\right)\right]^2}{\sigma_i^{\,2}}. \tag{8}$$

Among them, the smallest $\widehat{f}_1, \widehat{f}_1, \ldots, \widehat{f_m}$ is the factor score of the corresponding data.

Expressed as a matrix:

$$x - \mu = AF + \varepsilon. \tag{9}$$

To achieve the minimum $(x - \mu - AF)^T D^{-1}(x - \mu - AF)$, the minimum value F is the factor score of the response data, among them,

$$D = \begin{bmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_p^{-2} \end{bmatrix}. \tag{10}$$

The calculated F is satisfied $A^T D^{-1}F = A^T D^{-1}A\,(x - \mu)$, and the solution is as follows:

$$\widehat{F} = \left(A^T D^{-1}A\right)^{-1}A^T D^{-1}\,(x - \mu). \tag{11}$$

3.3. Random Forest Model. Random forest (RF), proposed by Breiman [34], belongs to the Bagging class of ensemble algorithms. The core idea of Bagging is to use bootstrap to sample randomly, collect the same number of samples for each tree, repeat the process several times to generate several decision trees, train the learners separately, and integrate the training results of the weak learners into strong learning according to the strategy device. For the classification tree, the voting strategy is combined with the result of the weak learner, and the category with the most votes is the final output of the model. For the regression tree, the arithmetic average of the output of the weak learner is used as the final predicted value of the model. The structure diagram is shown in Figure 1.

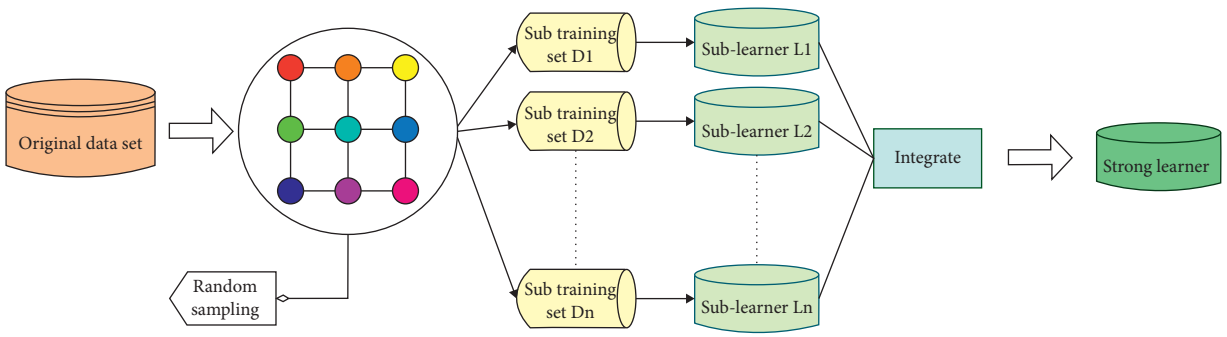

Figure 1: Bagging structure.

The bagging framework that chooses the CART tree as a weak learner is called random forest [35]. When the decision tree grows, it is different from other decision trees. The CART tree is a binary tree and uses the feature with the smallest Gini index as the split point to split to generate two subtrees. The Gini Index, also known as Gini Impurity, is usually used to measure the degree of uncertainty. Because the CART tree is a binary tree, the Gini index can be expressed as follows:

$$\text{Gini}(p) = 2p(1-p). \tag{12}$$

In the formula, p refers to the probability of being classified into this category.

3.4. Bayesian Optimization Algorithm. The Bayesian optimization algorithm, proposed by Snoek et al. [36], is one of the most famous scalable applications of Bayesian networks and is often applied for hyperparameter optimization in machine learning models. The algorithm defines the distribution of the objective function from the input data to the output data and requires that there are several sample points (assuming that the hyperparameters conform to the Gaussian process (GP)). Through the Gaussian regression process, the posterior probability distribution of the known n points is calculated, and the expected mean and variance of each hyperparameter at each value point are obtained. The mean value represents the final expected effect. The larger the mean value, the larger the final index of the model; the variance represents the uncertainty [37]. Therefore, in Gaussian regression, points with large mean and large variance should be selected. The main idea (as shown in Figure 2) is to give an optimized objective function, continuously add sample points provided by the acquisition function (AC) (like Upper confidence board, UCB; Probability of improvement, PI; Expected improvement, EI) to update the posterior distribution of the objective function, and continue to receive the last parameter information to update the current parameters until the posterior distribution Basically fits the real distribution.

3.5. Evaluation Indicator. There are three commonly used regression model evaluation indicators: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE). Their formulations are as follows:

$$\begin{aligned}
\text{MAE} &= \frac{\sum_{i=1}^{N}\left|t_{\text{pi}} - t_{\text{oi}}\right|}{N},\\
\text{MAPE} &= \frac{1}{N}\sum_{i=1}^{n}\left|\frac{t_{\text{pi}} - t_{\text{oi}}}{t_{oi}}\right| \times 100\%,\\
\text{RMSE} &= \sqrt{\frac{\sum_{i=1}^{N}\left(t_{\text{pi}} - t_{\text{oi}}\right)^2}{N}}.
\end{aligned} \tag{13}$$

In the formula, N is the number of data sets, t_{pi} is the prediction value, and t_{oi} is the true value.

3.6. Prediction Framework. Firstly, use the spatiotemporal definition method, identify the secondary accidents cascading with the first accident from the original data, integrate the accident pairs, and verify the accuracy of the accident pair recognition through the accident description. Then, KMO measurement and Bartlett sphere test are introduced to test the applicability of factor analysis to the accident information. Finally, after the verification is passed, the data is analyzed by factor analysis, and a random forest model (RF) is constructed to predict the interval between the first and second accidents of road traffic accidents. The duration model and the framework flow chart shown in Figure 3 are as follows:

Step 1: According to the four dimensions of the accident's start time, end time, the longitude of occurrence, and latitude of occurrence, use the spatiotemporal definition method to process each accident record information in the original data, and extract the secondary accidents cascaded with it.

Step 2: Verify the information extracted in Step 1 based on the accident description information of the accident. If the verification fails, it will be filtered; if the information passes the verification, the accident information will be integrated.

Step 3: Preprocess the data of the new data set integrated for the first accident and the second accident, calculate the Pearson coefficient between the independent variable and the dependent variable, KMO

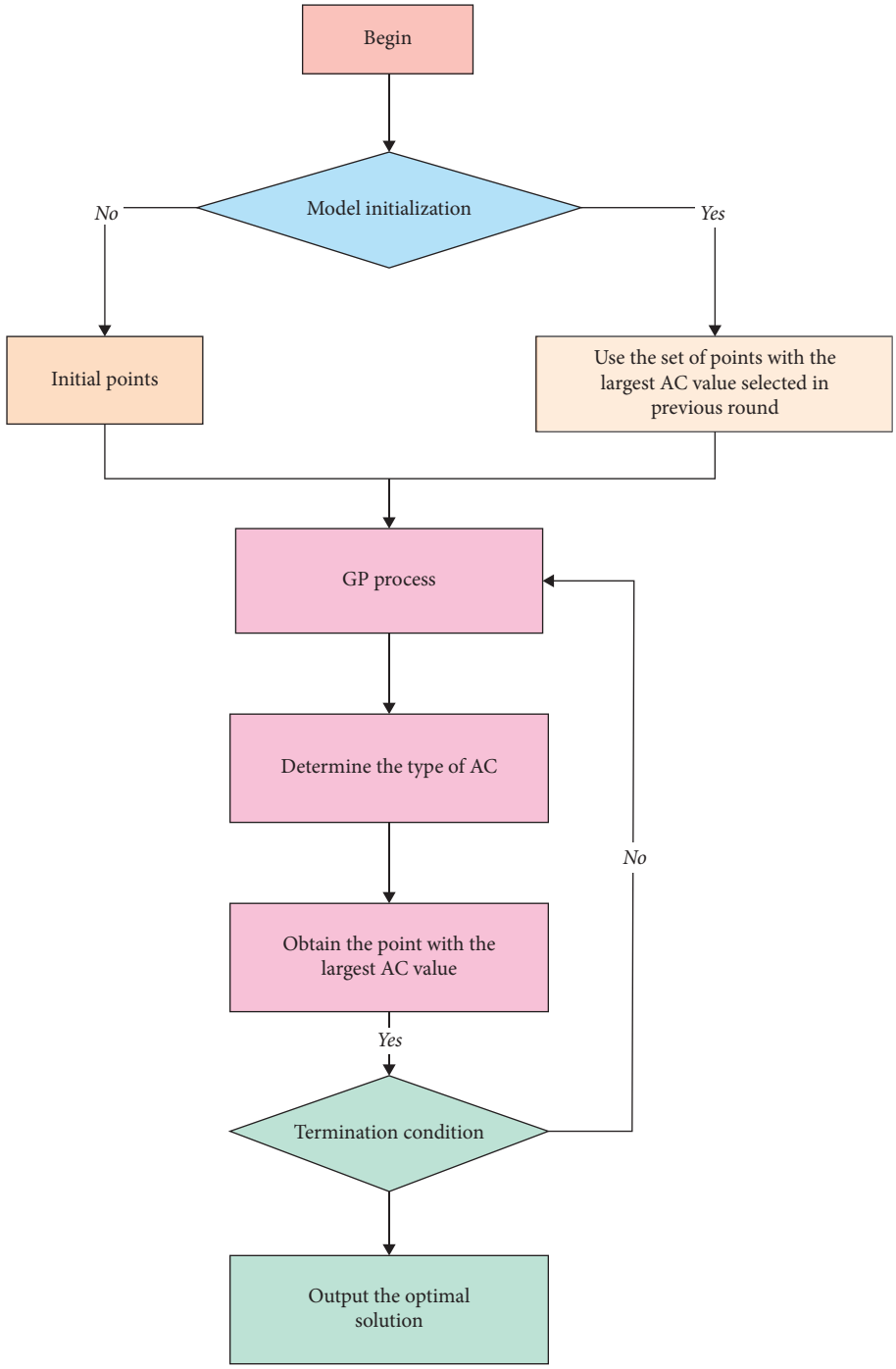

Figure 2: The flowchart of Bayesian optimization.

measurement statistics, and Bartlett's sphere test statistics to analyze whether the data can be applicable for factor analysis method.

Step 4: Use factor analysis to extract the common factors of the data set, stratify the influencing factors, calculate the factor score, calculate the weight of each variable, and feedback the factor score to each sample point to form a new data set.

Step 5: Divide the new data set into a training set and test set according to the ratio of 7 : 3.

Step 6: Use the Bagging method, specifically the Bootstrap self-sampling method, to process the training set, randomly select k samples (k less than the number of samples n) with replacement to form the subtraining set, and repeat this step.

Step 7: Construct a weak decision tree learner for each subtraining set and use the method of randomly selecting features during feature selection.

Step 8: Combine each weak learner to form a strong random forest learner. Input the test set and optimize

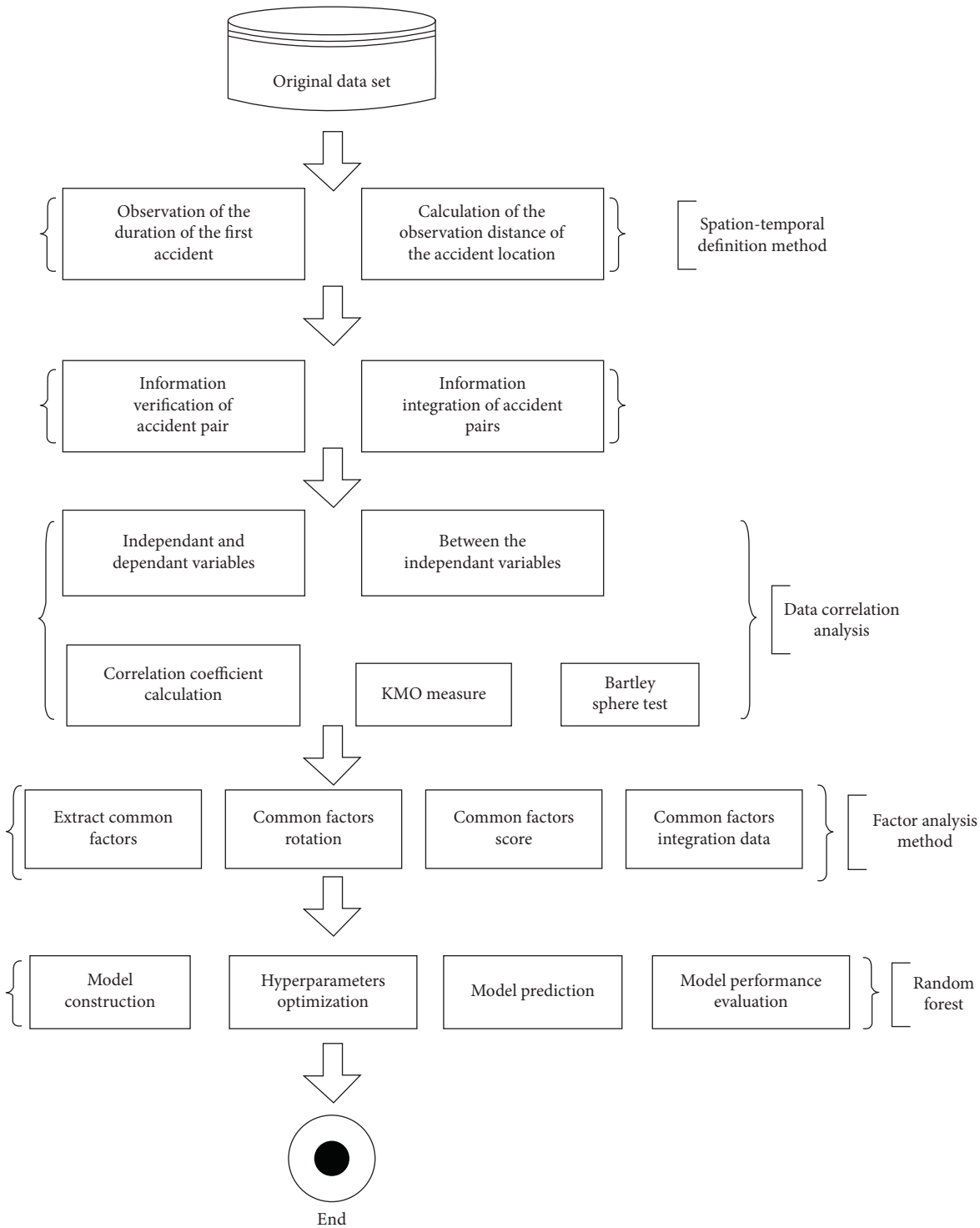

Figure 3: Interval duration prediction framework.

the frame parameters and decision tree parameters of the random forest through the Bayesian algorithm. After the model calibration is completed, input the training set and output the predicted value.

Step 9: Based on the true and predicted values of the test set, the model performance evaluation index is calculated to evaluate the model performance.

4. Results and Discussion

This study has two objects: (1) verifying the performance of the random forest model in predicting the duration of the interval between the first and second accidents and (2) investigating the important factors that affect the duration of the interval. We firstly select the spatiotemporal definition method to identify the cascading first accident and the

second accident and verify the accident pair matching degree through the accident information. Then, calculate the KMO measure and Bartlett's sphere test statistics of the accident pair to judge the applicability of the data to the factor analysis method. At last, the random forest (RF) model combined with factor analysis is applied to analyze and predict the interval duration between the first and second accidents.

4.1. Identification of Second Accidents. A second accident is defined as an accident that occurred within the scope of the initial accident. However, although there are many ways to record accidents in detail, in most accident data sets, accidents do not record the first accident and the second accident separately. It may be because when recording the accident, it is impossible to straightforwardly distinguish the accident as first or second [38]. Therefore, it is necessary to choose an appropriate method to process accident data sets to identify simple accidents and second accidents cascaded with the first accident. This study employs the spatiotemporal definition method to identify cascading accident pairs in the Los Angeles accident data set, and this method is sensitive to time-space thresholds. In order to optimize the effect of this method in identifying second accidents, we use a sensitivity analysis method with different time thresholds (duration to 180 minutes after the duration) and space thresholds (0.5 miles to 3 miles) to analyze the accident identification characteristics of each group of time and space thresholds.

In order to display the changes of the space threshold values in the spatiotemporal definition method to the second accident recognition effect, the time thresholds are set to 1 h after the duration of the first accident, 2 h after the duration of the first accident, 2.5 h after the duration of the first accident, and 3 h after the duration of the first accident. Figure 4 shows different sizes of the transformation space threshold to identify the number of secondary accidents.

By observing Figure 4, we can see that when the set time thresholds are 1 hour after the duration of the first accident, 2 hours after the duration of the first accident, 2.5 hours after the duration of the first accident, and 3 hours after the duration of the first accident, the number of accident pairs recognized through the spatiotemporal definition method remains stable, which means that the spatiotemporal definition method used in this dataset is not sensitive to space threshold.

In order to display the change of the time threshold value to the second accident recognition effect, set the space thresholds as the conditions of 1 mile from the first accident, 2 miles from the first accident, 2.5 miles from the first accident, and 3 miles from the first accident. Figure 5 shows different sizes of the transformation time thresholds to identify the number of second accidents.

By observing Figure 5, we can see that when the time thresholds are set to be 1 mile from the first accident, 2 miles from the first accident, 2.5 miles from the first accident, and 3 miles from the first accident, the number of accidents identified by the spatiotemporal definition method increases in the same trend. This means that the spatiotemporal definition method used in this data set is sensitive to the time threshold. Therefore, in order to determine the time threshold, this study uses an interval of 5 minutes as the duration to identify the number of second accidents.

It can be seen from Figure 6 that in the first 29 intervals, the number of identified accident pairs increases with the increase of the number of intervals. After the 30th interval, the number of identified accident pairs remains stable. This shows that after the 30th intervals of the accident duration, the spatiotemporal definition method is not sensitive to time thresholds. Therefore, the time threshold of the spatiotemporal definition method used in this study is set as 150 minutes (30*5) as the duration of the first accident.

Based on the above time threshold, a total of 767 sets of accident pairs were extracted. On this basis, continue to research and analyze the space threshold. Calculate the spatial distance of each pair of accidents and sort them into the ranges of 0.5 miles, 1 mile, 1.5 miles, 2 miles, and 3 miles. The number of accident pairs that meet the above range conditions is 754 pairs, 3 pairs, 3 pairs, 3 pairs, and 4 pairs, respectively. The results show that about 98.3% of the cascaded first accidents and second accidents in this data set have a space distance of 0.5 miles, so the space threshold is set to 0.5 miles.

Accordingly, we then use accident description fields and street information to verify accident pairs. After the verification is passed, there are a total of 754 valid accident pairs. After preprocessing of accident pairs (removal of redundant features, missing value repair, feature value processing, feature uniqueization, etc.), there are 28 remaining explanatory variables. Their codes and variables are shown in Table 2.

4.2. Applicability Test of Factor Analysis Method. Factor analysis method is introduced to explore the potential internal dependence of the first accident and the second accident; hence, it is necessary to test the applicability of the data factor analysis.

4.2.1. Pearson Coefficient Calculation. Calculate the correlation coefficient between the independent variable and the dependent variable points of the accident-to-data set through the Pearson coefficient. The results show the correlation coefficient between the duration of the first accident and the duration of the interval is 0.883, indicating that the duration of the first accident has a high correlation with the duration of the interval. In addition, the variables that are positively related to the interval duration are collision degree, edge, airport code, visibility, convenience facilities, bus stops, time of occurrence, and peak period; negatively related variables are the first accident impact distance, the second accident impact distance, humidity, air pressure, wind direction, wind speed, intersections, junctions, railways, signal lights, seasons, months, hours, weather, and precipitation.

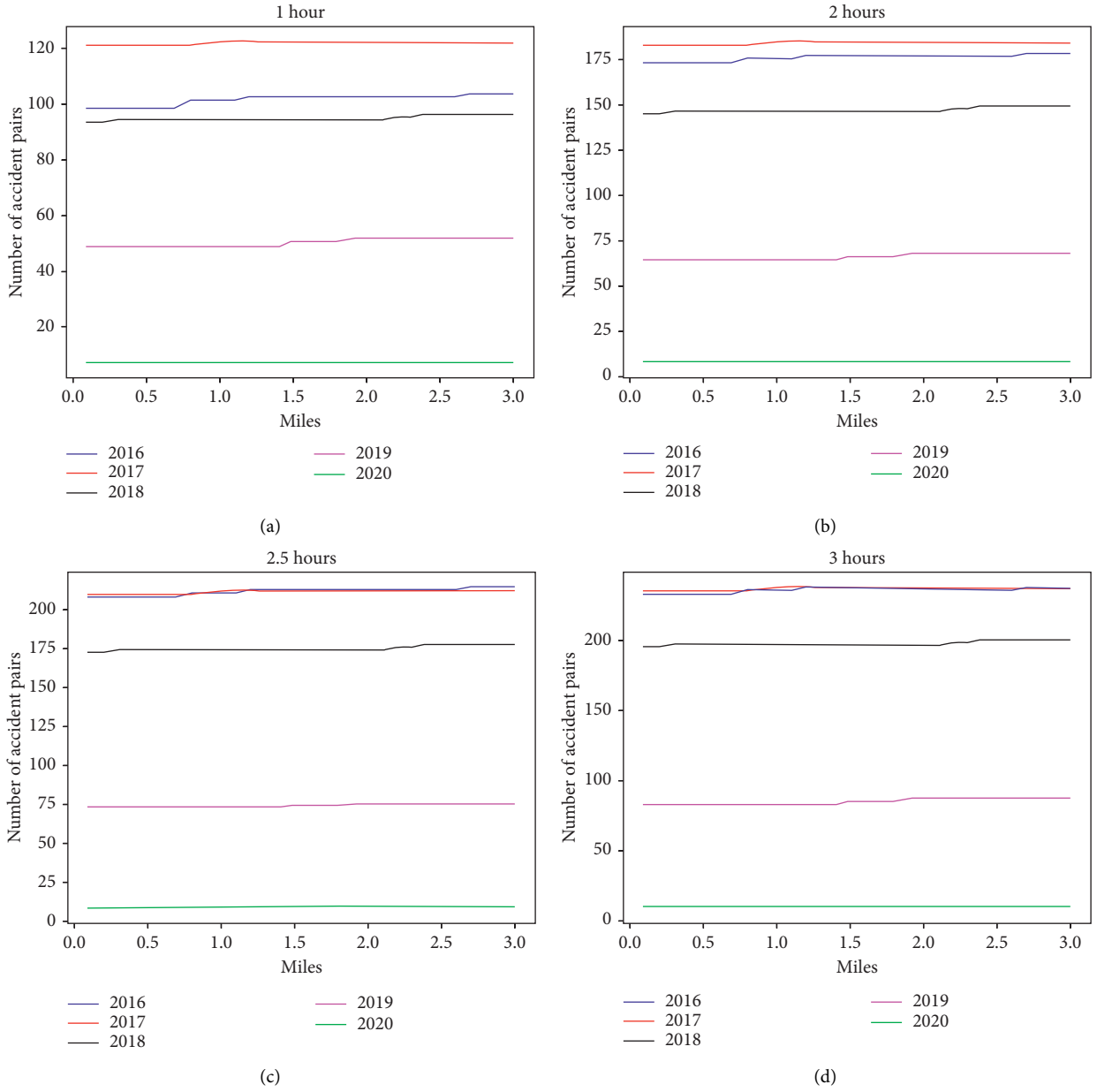

Figure 4: Space threshold sensitivity analysis.

4.2.2. KMO Measurement and Bartley Sphere Test. The effect evaluation corresponding to the calculated statistics [39] of KMO measurement is shown in Table 3.

Calculating the KMO measurement statistics of the integrated accident pair data, the result is 0.661, which is in the range of 0.6 to 0.7. With reference to Table 3, it can be seen that the KMO measurement effect of this data is rated as acceptable. In addition, the Bartlett sphere test is introduced to determine whether the correlation coefficient matrix is a unit matrix. The calculation result is 0.00, which is less than the significance level of 0.05, indicating that the null hypothesis can be rejected and is suitable for factor analysis.

4.3. Factor Analysis Method. After passing the applicability test, we can calculate the total variance explanation table, clarify the number of common factors, and output the factor loading matrix. Judging by the factor loading matrix, whether a reasonable explanation can be made for the variables. Otherwise, an appropriate method is adopted to rotate the factors so that the factor loading presents different characteristics. Determine the common factor corresponding to each explanatory variable according to the factor loading after rotation. Output its factor score coefficient matrix to calculate the weight of each factor and determine the significant influencing factors.

In order to judge the degree of influence of each explanatory variable on the interval duration, this study divides it into three levels based on the weight value of each explanatory variable: "significant influence," "general influence," and "small influence." In order to quantify the grading standard, the grading value needs to be established

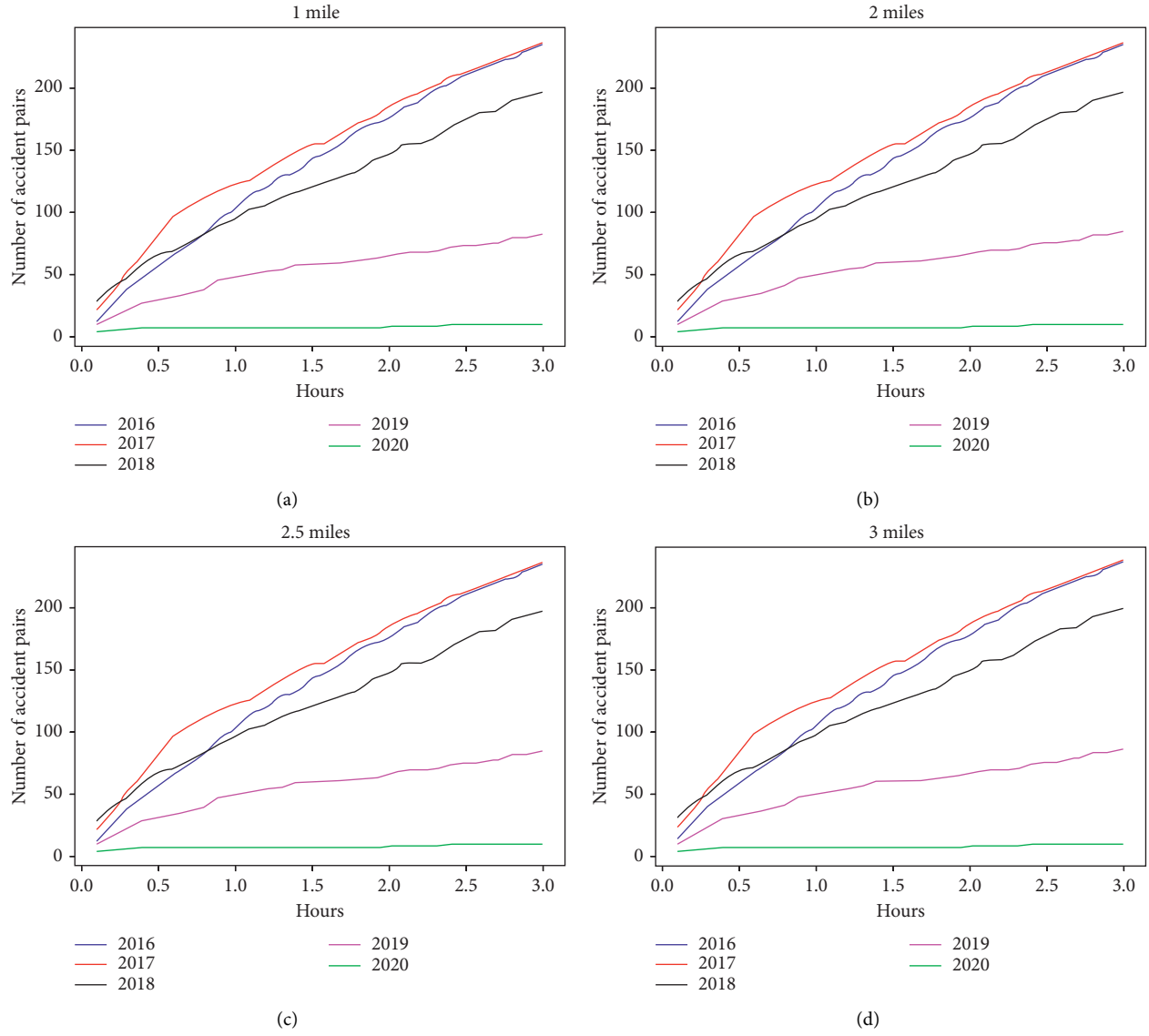

Figure 5: Time threshold sensitivity analysis.

first. As shown in Table 4, the weight values are all small. For the sake of comparison, the calculation results are shown in Table 4 according to $S_j = 100^* |w_j|$ to enlarge the weight value. Afterward, the classification system standards based on the calculated value are as follows:

$$P_j = \begin{cases} 1, & S_j \in [0, 2), \\ 2, & S_j \in [2, 4.5), \\ 3, & S_j \in [4.5, +\infty). \end{cases} \tag{14}$$

In the formula, P_j is the grading standard.

According to the above analysis results, there are 10 explanatory variables that can significantly affect the duration of the interval. They are conveniences, railways, stations, weather, distance affected, and distance affected by the second accident, weather at the time of the second accident, degree of collision, the severity of second accident collision and duration of the first accident; explanatory variables with general effects include wind speed, traffic lights, humidity, sunrise and sunset, visibility, wind direction, no thoroughfare, working day, intersection, junction, no precipitation, and airport postcode. There are six explanatory variables that have a low impact on the duration of the interval: peak period, season, side, peak period of the occurrence of second accidents, temperature, and air pressure.

4.4. Hyperparameter Optimization. The dual randomness of the RF model (random sampling and random selection of feature splits) reduces the variance of the prediction results and makes the model express good fitting results. Therefore, this study optimizes the main hyperparameters of the RF to obtain good prediction performance. The RF model mainly has 4 hyperparameters, which are the

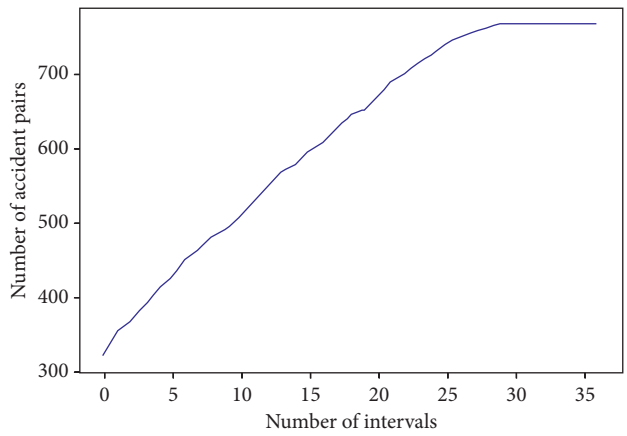

Figure 6: Sensitivity at 5-minute intervals.

Table 2: Correspondence table of variable and codes.

Code	Variable	Code	Variable
X_1	Collision degree	X_{15}	Station
X_2	Influence distance	X_{16}	No thoroughfare
X_3	Side	X_{17}	Traffic light
X_4	Airport postcode	X_{18}	Sunrise sunset
X_5	Temperature	X_{19}	Working day
X_6	Humidity	X_{20}	No precipitation
X_7	Pressure	X_{21}	Season
X_8	Visibility	X_{22}	Peak period
X_9	Wind direction	X_{23}	The weather
X_{10}	Wind speed	X_{24}	Impact distance of second accident
X_{11}	Conveniences	X_{25}	Severity of second accident collision
X_{12}	Intersection	X_{26}	Whether the second accident occurred during the peak period
X_{13}	Junction	X_{27}	Weather at the time of the second accident
X_{14}	Railway	X_{28}	Duration of first accident

number of decision trees, the maximum depth of decision trees, the minimum number of samples required for subdividing internal nodes, and the minimum number of samples for leaf nodes.

The Bayesian optimization algorithm is employed through the Hyperopt module in Python. After setting the objective function, search space, optimization algorithm, and the maximum number of evaluations, the module can automatically search for the optimum according to the given parameters. To avoid overfitting, 10-fold cross-validation is incorporated into the Bayesian optimization algorithm. In this study, RF is the objective function; Parzen tree is the default optimization algorithm; the maximum number of evaluations is set to 4; the search space is the search range of each hyperparameter, and the settings are shown in Table 5. The hyperparameters for the automatic optimization of the Bayesian optimization algorithm of 10-fold cross-validation are shown in the table.

4.5. Interval Duration Prediction and Analysis. The above optimized parameter values are set as hyperparameters of the RF model, the optimal RF model is applied to predict the test set, and the prediction results of the prediction set are analyzed to evaluate the usability of the algorithm.

4.5.1. Results Prediction. By observing from Table 6, the average true value of the test set is 42.605 min, which is slightly lower than the average 42.273 min of the predicted value. This means that the predicted value is generally close to the true value. The standard deviation of the true value is 23.807 min, which is higher than the deviation of 18.010 min of the predicted value, which indicates that the predicted value of the test set is more stable than the true value distribution. The minimum value, 25% quantile, 50% quantile, 75% quantile, and maximum which are of the true value and the predicted value are 26 min, 30 min, 43.5 min, 45 min, 300 min, and 27.717 min, 30.021 min, 43.378 min, 45.064 min, 136.061 min. The distribution range of the true value is 274 and the distribution range of the predicted value is 108.3. That is, the predicted value distribution is more concentrated, which is also consistent with the conclusion of the standard deviation of the true value and the predicted value.

4.5.2. Model Performance Comparison and Evaluation. In order to measure the performance of the RF model in predicting the duration of the interval between the first and second accidents, the *K*-nearest neighbor model (KNN), and the support vector regression model (SVR) were introduced for comparison, and the absolute MAE,

Table 3: KMO measurement statistics effect evaluation.

KMO measurement	>0.9	0.8-0.9	0.7-0.8	0.6-0.7	0.5-0.6	<0.45
Whether to apply factor analysis	Extremely suitable	Very suitable	Suitable	Acceptable	Very bad	Not suitable

Table 4: The importance level of the weight of each variable.

Three-level indicators	w_j	S_j	P_j
Conveniences	0.062	6.2	3
Railway	0.052	5.2	3
Station	0.052	5.2	3
Weather	0.050	5.0	3
Influence distance	0.049	4.9	3
Impact distance of second accident	0.049	4.9	3
Weather at the time of the second accident	0.049	4.9	3
Collision degree	0.047	4.7	3
Severity of second accident collision	0.047	4.7	3
Duration of first accident	0.045	4.5	3
Wind speed	0.041	4.1	2
Traffic light	0.041	4.1	2
Humidity	0.040	4.0	2
Sunrise sunset	0.040	4.0	2
Visibility	-0.037	3.7	2
Wind direction	0.037	3.7	2
No thoroughfare	0.036	3.6	2
Working day	0.033	3.3	2
Intersection	0.033	3.3	2
Junction	0.030	3.0	2
No precipitation	0.030	3.0	2
Airport postcode	0.014	1.4	2
Peak period	0.029	2.9	1
Season	0.028	2.8	1
Side	0.027	2.7	1
Whether the second accident occurred during the peak period	0.027	2.7	1
Temperature	−0.024	2.4	1
Air pressure	−0.018	1.8	1

Table 5: Optimized values of RF hyperparameters.

Hyperparameter	Search space	The optimal value
Number of decision trees	[50, 500]	245
Maximum depth of decision tree	[1, 8]	5
Minimum number of samples required for subdividing internal nodes	[1, 10]	3
Minimum number of samples for leaf nodes	[1, 10]	7

Table 6: Calculation table observed and predicted values of interval duration.

Clustering data set	True value (min)	Predicted value (min)
Sample size	228	228
Average value	42.605	42.273
Standard deviation	23.807	18.010
Minimum	26.000	27.717
25% quantile	30.000	30.021
50% quantile	43.500	43.378
75% quantile	45.000	45.064
Maximum	300.000	136.061

MAPE, and RMSE of the three models were calculated, respectively. The results are shown in Table 7.

As shown in Table 7, the MAPE values of RF, KNN, and SVR in the interval duration prediction are 1.310%, 1.516%, and 1.801%, respectively. The results show that the MAPE value of the RF model is the smallest, indicating that the RF model is more capable than the KNN and SVR models. It is more accurate to predict the duration of the interval.

Table 7: Evaluation index calculation table.

Model	RF	KNN	SVR
MAE (min)	1.500	1.488	2.569
MAPE	1.310%	1.516%	1.801%
RMSE (min)	11.285	11.214	19.675

Specifically, MAE, MAPE, and RMSE of the RF model are 1.689 min, 1.310%, and 11.822 min, respectively. The MAPE value is 1.310%, which is between [0%, 10%], so the RF combined with the factor analysis method can predict the interval duration with higher accuracy.

5. Conclusion and Prospect

In order to explore the potential influencing factors of the interval duration between the first accident and the cascaded second accident and predict it. In this paper, the sensitivity analysis method is applied to determine the time-space thresholds of the spatiotemporal definition, and the cascade of first and second accidents that met the conditions is extracted. Then, after calculating the KMO measure and Bartlett's spherical test statistic to verify that the accident is applicable to the data set for factor analysis, the factor analysis method is carried out to obtain factors that significantly affect the duration of the interval. We divide the processed data into a training set and test set at a ratio of 7 : 3, construct a RF model based on the test set, and select the Bayesian optimization algorithm with tenfold cross-validation to optimize the hyperparameters. Based on the true and predicted values of the test set, the MAE, MAPE, and RMSE are calculated. The main contribution of this work is:

(1) Some studies use the spatiotemporal definition method to identify accident pairs, but the difference of different thresholds in identifying accident pairs is not considered. Therefore, this paper integrates the sensitivity analysis into the spatiotemporal definition method to determine the time-space thresholds. The results show the interval is sensitive to a time threshold and not sensitive to space threshold.

(2) Explore the factors that can significantly affect the duration of the interval between the first accident and the cascaded second accident. In fact, there are many studies related to second accidents, but there is a lack of research on predicting the time of occurrence of second accidents. This paper introduces the factor analysis method, constructs the influencing factor analysis as a three-level indicator, improves the accuracy of the factor analysis, and provides support for traffic managers to prevent second accidents. Finally, it can be concluded that there are 10 explanatory variables that can significantly affect the duration of the interval. They are conveniences, railways, stations, weather, distance affected, distance affected by the second accident, whether at the time of the second accident, degree of collision, the severity of second accident collision, and duration of the first accident.

(3) The test results show the MAPE value of RF is 1.58%, which is within [0, 10%], indicating that the RF model of the fusion factor analysis method can predict the interval duration with high accuracy. Moreover, the comparative experiments confirm the RF outperforms KNN and SVR.

However, there are parts that can be improved in this article. The specific content is as follows:

(1) With the development of intelligent transportation systems, the configuration rate of various detectors used for traffic management on the road is also getting higher and higher, and the types and quality of collected data are becoming more and more abundant, and high-quality data can allow research to better see the nature of the problem through the traffic phenomenon. Therefore, future research will focus on data sets that combine accident data and traffic flow data to obtain more accurate and convincing accident duration predictions.

(2) In order to improve the efficiency and accuracy of secondary accident recognition, we need to incorporate the dynamic changes of time and space thresholds into the recognition method.

(3) This study currently only models the characteristics of the collected accident data. However, some unrecorded or even unobserved potential factors affect the estimation of duration.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

This research was funded in part by the Innovation-Driven Project of Central South University (no. 2020CX041) and the National Natural Science Foundation of China (no. 52172310).

References

[1] J. Y. Wang, H. P. Lu, Z. Y. Sun, T. S. Wang, and K. Wang, "Investigating the impact of various risk factors on victims of traffic accidents," *Sustainability*, vol. 12, no. 9, 2020.

[2] Y. J. Lu, G. T. Lai, and Y. Feng, "The modeling and simulation of collision protection system between the driver of non-motor vehicle and car door," *EURASIP Journal on Wireless Communications and Networking*, vol. 99, 2018.

[3] S. Tedesco, V. Alexiadis, W. Loudon et al., "Development of a 40 model to assess the safety impacts of implementing IVHS user services, moving toward deployment," in *Proceedings of the IVHS America Annual Meeting*, pp. 343–352, Atlanta, GA, USA, 1994.

[4] R. A. Raub, "Occurrence of secondary crashes on urban arterial roadways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1581, no. 1, pp. 53–58, 1997.

[5] J. Lindley and S. Tignor, "Getaway flow rates for freeway incident and geometricbottlenecks," *Public Roads*, vol. 43, pp. 1–7, 1979.

[6] M. G. Karlaftis, S. P. Latoski, N. J. Richards, and K. C. Sinha, "ITS impacts on safety and traffic management: an investigation of secondary crash causes," *ITS Journal-Intelligent Transportation Systems Journal*, vol. 5, no. 1, pp. 39–52, 1999.

[7] W. Hirunyanitiwattana and S. Mattingly, "Identifying secondary crash characteristics for California highway system," in *Proceedings of the Transportation Research Board 85th Annual Meeting*, Washington DC,US, January 2006.

[8] J. E. Moore, G. Giuliano, and S. Cho, "Secondary accident rates on Los Angeles freeways," *Journal of Transportation Engineering*, vol. 130, no. 3, pp. 280–285, 2004.

[9] C. Zhan, A. Gan, and M. Hadi, "Identifying secondary crashes and their contributing factors," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2102, no. 1, pp. 68–75, 2009.

[10] C. Sun, "Secondary accident data fusion for assessing long-term performance of transportation systems," *Midwest Transportation Consortium Rep. MTC Project 2005-04*, Iowa State Univ., Ames, Iowa, 2005.

[11] C. C. Sun and V. Chilukuri, "Dynamic incident progression curve for classifying secondary traffic crashes," *Journal of Transportation Engineering*, vol. 136, no. 12, pp. 1153–1158, 2010.

[12] Y. Chung and W. Recker, "A methodological approach for estimating temporal and spatial extent of delays caused by freeway accidents," *IEEE Transactions on Intelligent*, vol. 13, no. 3, pp. 1454–1461, 2010.

[13] Y. Chung, "Quantification of nonrecurrent congestion delay caused by freeway accidents and analysis of causal factors," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2229, no. 1, pp. 8–18, 2011.

[14] Y. Chung, "Identifying primary and secondary crashes from spatiotemporal crash impact analysis," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2386, no. 1, pp. 62–71, 2013.

[15] H. Yang, B. Bartin, and K. Ozbay, "Identifying secondary crashes on freeways using sensor data," *Transportation Research Record*, vol. 2396, no. 10, pp. 82–92, 2013.

[16] H. Yang, B. Bartin, and K. Ozbay, "Mining the characteristics of secondary crashes on highways," *Journal of Transportation Engineering*, vol. 140, no. 4, pp. 4–24, 2014.

[17] A. Haghani, D. Iliescu, M. Hamedi et al., "Methodology for Quantify Hecost Effective of Freeway Service Patrol Program," *Hudson Valley Highway Emergency Local Patrol (HELP) Program Report*, University of Maryland, College Park, MD, USA, 2006.

[18] C. Chou and E. Miller-Hooks, "Simulation-based secondary incident filtering method," *Journal of Transportation Engineering*, vol. 136, no. 8, pp. 746–754, 2009.

[19] D. Chimba and B. Kutela, "Scanning secondary derived crashes from disabled and abandoned vehicle incidents on uninterrupted flow highways," *Journal of Safety Research*, vol. 50, pp. 109–116, 2014.

[20] J. Wang, B. Liu, T. Fu, S. Liu, and J. Stipancic, "Modeling when and where a secondary accident occurs," *Accident Analysis & Prevention*, vol. 130, pp. 160–166, 2019.

[21] B. Se-Ryong, Y. J. Kyu, and L. J. Han, "A study on the installation of a barrier to prevent large-scale traffic accidents in tunnel," *The International Journal of Advanced Smart Convergence*, vol. 8, no. 4, pp. 161–168, 2019.

[22] T. Aoki, K. Asami, S. Ito, and S. Waki, "Development of quadruped walking robot with spherical shell: improvement of climbing over a step," *Robomech Journal*, vol. 7, no. 1, 2020.

[23] M. Kostikova, K. Bucsuhazy, P. Moravcova et al., "Comparative analysis of two almost identical traffic accidents. Připadova studie: komplexni porovnani dvou teměř identickych nehod," *Česko-Slovenská Patologie*, vol. 51, no. 1, pp. 6–10, 2021.

[24] J. Pietilä, T. Räsänen, A. Reiman, H. Ratilainen, and E. Helander, "Characteristics and determinants of recurrent occupational accidents," *Safety Science*, vol. 108, pp. 269–277, 2018.

[25] H. Kim, J. Jeon, H. Kim, and J. Ahn, "A study on smart road stud system with RF wireless control," *Journal of Korea Institute of Information, Electronics, and Communication Technology*, vol. 12, no. 3, pp. 282–289, 2019.

[26] H. Zhang, C. X. Zhuge, J. M. Jia, B. Y. Shi, and W. Wang, "Green travel mobility of dockless bike-sharing based on trip data in big cities: a spatial network analysis," *Journal of Cleaner Production*, vol. 313, 2021.

[27] H. Zhang, B. Shi, C. Zhuge, and W. Wang, "Detecting taxi travel patterns using GPS trajectory data: a case study of Beijing," *KSCE Journal of Civil Engineering*, vol. 23, no. 4, pp. 1797–1805, 2019.

[28] X. Chen, L. Qi, Y. Yang et al., "Video-based detection infrastructure enhancement for automated ship recognition and behavior analysis," *Journal of Advanced Transportation*, vol. 2020, pp. 1–12, 2020.

[29] X. Chen, Z. Li, L. Qi, and Y. Yang, "High-resolution vehicle trajectory extraction and denoising from aerial videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 1–12, 2021.

[30] R. Gorsuch, "Factor analysis," *Handbook of Psychology*, Wiley, New York, NY, USA, 1983.

[31] Y. Li, *Analysis of Influencing Factors of Road Traffic Accidents in China and Prediction of the Severity of Accidents*, Jilin University, Changchun,China, 2018.

[32] S. Bing, *Research on Satisfaction Evaluation of Sustainable Development Level of Urban Human Settlement Environment*, Chongqing University, Changchun, China, 2008.

[33] C. Ke, *The Evaluation Model of College Students' Physical Fitness Based on Factor Analysis*, The Guide of Science and Education, Washington, DC, USA, 2015.

[34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] W. Teng, "Analysis of random forest data sentiment mining method," *Communication World*, vol. 27, no. 1, pp. 7–9, 2020.

[36] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," *Advances in Neural Information Processing Systems*, vol. 4, pp. 2951–2959, 2012.

[37] X. Mu, *Research on the Fixed Point of Nonlinear Mapping in Probabilistic Metric Space*, Nanchang University, Nanchang, China, 2016.

[38] L. Zhao, *Prediction of the Duration of Highway Traffic Accidents*, Beijing Jiaotong University, Beijing, China, 2017.

[39] B. Liu, *Population Environmental Health Survey Evaluation and Application Research*, China University of Petroleum, Dongying, China, 2013.