Cause Analysis of Railway Traffic Accidents Based on Random Forest

Minxuan Wang and Di Liu*
College of Traffic and Transportation Engineering, Dalian Jiaotong University, Dalian, Liaoning, China
Email: liudi@djtu.edu.cn

Abstract. With the rapid growth of the railway operation scale, all kinds of railway traffic accidents happen from time to time, so it is great significant to accurately identify the main influencing factors and their influence degree. In this paper, random forest model is proposed to analyze the cause of railway traffic accidents. Considering people, equipment, environment and other aspects, 11 influencing factors were extracted from 491 accident data. The influence degree of different factors on the severity of the accident is judged through variable importance measures of the random forest. On the basis of this, some suggestions are put forward for raising the safety level of railway transportation. The results show that the random forest model is accurate for analyzing the causes of railway traffic accidents, which can provide decision support for railway transportation safety management.

Keywords. Railway traffic accidents, Random forest, Variable importance measures, Cause analysis.

1. Introduction
As the backbone of comprehensive transportation system, railway occupies an important proportion of transportation system because of its characteristics of low cost, large volume and long distance. In recent years, with the development of economy and technology, China's railway transportation has entered a period of rapid development. In 2019 the railway mileage exceeded 139000 kilometers 3.66 billion passengers and 4.389 billion tons of goods were sent. The rapid growth of railway operation scale brings great pressure to transportation safety management and all kinds of railway traffic accidents occur frequently. The cause of the accident is usually not a single factor, often a combination of multiple factors. So how to accurately identify the main factors influencing the accident severity, advances some corresponding management ideas, this will help reduce the incidence of railway traffic accidents and improve the level of safety management of railway transportation is of great significance.

On the cause analysis of railway traffic accidents, Y Qin [1] proposed the disturbance-safety domain model for accident cause analysis of complex system is which can not only describe the dynamic process of multi-factor disturbance in accident development, but also provide scientific and precise guidance and operation basis for accident prevention. H W Xin [2] constructed the cause-related network of railway accidents and established the dynamic weight model. Through the simulation results, it was proved that strengthening the prevention and control of nodes with large degree value could effectively cut off the propagation of the cause chain of accidents and improve the safety of the whole network. Q J Zhan [3] respectively established HFACs-RAS model to conduct modeling and quantitative analysis of railway traffic accidents, identify the main causes of accidents,
and narrow the scope of problem solving, so as to effectively improve the existing safety defects. C Y Lam [4] used the network topology method to analyze the railway traffic accidents in Japan, described the causal relationship in the accident chain, and revealed the important factors causing railway accidents. With the development and improvement of computer language function, machine learning has been better applied in the research method of accident cause. T Chai and S X Wang [5-6] used the support vector machine model in machine learning to predict road traffic accident duration and railway accident respectively. Ci Liang [7] established Bayesian networks to predict railway level crossing accidents, and determined the most dangerous factors by quantifying the influencing factors to improve the safety of crossings. But these machine learning algorithms are all single classifiers, which have the bottleneck of performance improvement and the problem of overfitting. So the method of integrating multiple classifiers to improve the prediction performance arises. Random forest is one of the typical machine learning integration algorithms. X Y Zhou [8] has already applied the random forest model to the safety problem of highway level crossing, and the accuracy of accident prediction is improved.

Based on the analysis of the influencing factors of railway traffic accidents, this paper uses the random forest method in machine learning to analyze the importance of the cause of accidents and improves the accuracy of the model by adjusting the parameters, which provides a reference for the safety management suggestions.

2. Basic Theory of Railway Traffic Accident

2.1. Classification of Railway Traffic Accidents
According to the regulations of Railway Traffic Accident Investigation and Handling Rules, the grade of railway traffic accidents is divided into four grades: special major accidents, major accidents, larger accidents and general accidents. The general accident, which does not constitute a larger accident, is divided into general A-class accidents, general B-class accidents, general C-class accidents and general D-class accidents. The severity of accidents decreases from A-class to D-class.

2.2. Factors Affecting Railway Traffic Accidents
The factors influencing railway traffic accidents are complex. Accidents can occur due to personnel and equipment failures. Therefore, it is divided into four parts: human factors, equipment factors, environmental factors and other factors.

2.2.1. Human Factors. The main influencing factors of people in railway traffic accidents are the unsafe factors of railway operators, which are considered from two aspects: professional accomplishment and physical quality of operators. Professional accomplishment refers to the operational level of operators, their mastery of the business of their positions and proficiency, as well as whether there are illegal operations. Among them, operation is the main factor affecting the severity of the accident results. When operators violate the rules, it may cause serious casualties and other irreversible consequences. Physical fitness refers to the degree of health of the operator, whether there is a serious disease affecting the work of physical and mental health problems. The normal operation will be greatly affected when the operator's health condition appears abnormal and sudden disease during the operation.

2.2.2. Human Equipment Factors. Among the influencing factors of railway traffic accident equipment, the paper mainly considers locomotive equipment and electrical equipment. As the operation center of train operation, the failure of driving equipment may cause the train to get out of control and cause serious accident. Electrical equipment mainly includes two aspects: catenary pantograph and underground cable. When the catenary pantograph fails or the underground cable is damaged, it will affect the normal operation of the train and lead to the accident.
2.2.3. Environmental Factors. Environmental factors have an impact on both human and equipment, including the natural and operational environments. The natural environment includes the climate, temperature and weather conditions at the time of the accident. Extreme weather conditions are not conducive to the operation, there is a hidden danger of accidents. For railway traffic accidents, the operating environment includes crossing environment, whether pedestrians and motor vehicles intrude. When there are other traffic modes at the crossing, the train cannot avoid collision, which will lead to serious accidents and resulting in casualties. The operating environment also includes the external environment, whether there is a plant or building invasion line. Because the plant or building invasion line will cause damage.

2.2.4. Other Factors. In addition to the above three aspects, time, place and type of accidents also have an impact on the severity of railway traffic accidents. Due to the continuous operation of the railway 24 hours, the accident occurred in the day shift or night shift. Station or section, driving, off-road or personal accident is one of the factors affecting the severity of the accident. Therefore, the time, place and type of accidents are selected for a more comprehensive analysis.

3. Random Forest Models

3.1. Random Forest Classifiers
Random forest is an integrated model containing multiple decision trees, which is composed of decision tree and autonomous sampling. The classification process of random forests is divided into four steps:

(1) Generate m training subsets from the training set using Bootstrap sampling;
(2) K variables were randomly selected from all attributes and a decision tree was constructed with M training subsets;
(3) Repeat the first two steps until L decision trees are constructed;
(4) The test sets are classified and predicted using L decision trees, the results of each random build are summarized, and the final output is obtained by weighting or voting.

Random forest classifies the selected data into multiple decision trees and integrates the advantages of autonomous sampling and random attribute selection. Random forest can be used for dichotomy and multi-classification problems, and it has good generalization ability and generally does not appear overfitting. On the basis of the known data classification results, the data is divided into training set and test set. The training set establishes the classification model according to the data characteristics. The test set is the model established according to the data characteristics of the training set to test and judge whether the model meets the requirements. The random forest model can be tested on the test set after learning the data classification results of the training set. The random forest model can be used to test the training set after learning the data classification results, and the model accuracy can be improved through parameter adjustment. It is also possible to calculate the importance score of each characteristic factor to the degree of influence of the classification results.

It can be seen that the random forest is suitable for the analysis of factors affecting the severity of railway traffic accidents. The variable importance measures can be used to determine the degree of influence of each factor on the severity of the accident results.

3.2. Input and Output Variables
In the random forest, the input variable refers to the characteristic variable of the data, and the output variable refers to the classification result of the data, all of which are expressed in the form of vectors.

In the railway traffic accident data, the specific process of the accident and the type of accident are recorded in detail. To the big data of railway traffic accident, which is divided into m types of accidents, n influencing factors are extracted, and the input characteristic variables and output variables of the model are constructed. Therefore, the influencing factors and classification results of text attributes
should be assigned in advance. The expressions are:

\[ X = (x_1, x_2, ..., x_n), n = 1, 2, 3, ... \]  \( (1) \)

\[ Y = (y_1, y_2, y_3, ..., y_m), m = 1, 2, 3, ... \]  \( (2) \)

### 3.3. Variable Importance Measures

Variable importance measures are one of the main results obtained by random forest, which is used to explain the importance of input variables to the occurrence of results. It is used as the basis for classifying output variables. Approximately 1/3 of samples at a time during random forest Bootstrap sampling do not appear in the resulting sample set. These become out-of-bag data (OOB).

The process of variable importance measures is divided into three steps:

1. Using the OOB of each decision tree in a random forest to calculate the corresponding out-of-bag data error, recorded as errOOB1;
2. Noise interference is added to the characteristic X of all OOB samples, and the out-of-bag error is calculated again and recorded as errOOB2;
3. Suppose there are N trees in a random forest, the importance score of the characteristic X is:

\[ \sum_{i=1}^{N} \left( err_{OOB2} - err_{OOB1} \right) / N \]  \( (3) \)

In the random forest model, the higher the importance score of the input variable, the greater the influence of this variable on the classification of the results. Input variable of railway traffic accident is the n influence factor extracted, when the importance score is higher, it means that the factor has greater influence on railway accident classification.

### 4. Example Applications

This paper selects a total of 491 railway transportation accidents (general A-class accidents, general B-class accidents, general C-class accidents, general D-class accidents) which occurred in 2019 in a certain railway transportation enterprise area in northeast China.

#### 4.1. Data Description and Statistical Analysis

According to the specific contents of the accident and the analysis of the influencing factors of the accident in 2.2, 11 factors are selected for more detailed classification. Statistical analysis of the relationship between individual variables and the severity of railway traffic accidents is helpful to intuitively understand the causes of traffic accidents and help decision makers to make decisions in railway transportation safety management. Table 1 shows the percentage of all types of traffic accidents caused by 11 traffic accident related factors.
**Table 1.** Statistics of railway traffic accidents.

| Factors                      | Content                  | A-class | B-class | C-class | D-class |
|------------------------------|--------------------------|---------|---------|---------|---------|
| Accident time                | Day shift                | 0%      | 34.2%   | 11.3%   | 54.5%   |
|                              | Night shift              | 0.9%    | 33.1%   | 11.3%   | 54.7%   |
| Location of accident         | Station                  | 0%      | 17.2%   | 2.5%    | 80.3%   |
|                              | Interval                 | 0.7%    | 41.5%   | 15.6%   | 42.2%   |
| Type of accident             | Driving                  | 0.3%    | 1.2%    | 1.6%    | 96.9%   |
|                              | Road out                 | 0.5%    | 70.2%   | 24.7%   | 4.6%    |
|                              | Personal status          | 0%      | 100%    | 0%      | 0%      |
| Driving equipment status     | Normal                   | 0.6%    | 41.4%   | 14.2%   | 43.8%   |
|                              | Failure                  | 0%      | 0%      | 0%      | 100%    |
| External equipment status    | Normal                   | 0.4%    | 35.2%   | 11.9%   | 52.5%   |
|                              | Failure                  | 0%      | 7.4%    | 7.4%    | 85.2%   |
| Underground cable status     | Normal                   | 0.4%    | 34.6%   | 11.9%   | 53.1%   |
|                              | Failure                  | 0%      | 0%      | 0%      | 100%    |
| Health status of operators   | Health                   | 0.4%    | 32.4%   | 11.8%   | 55.4%   |
|                              | Sudden illness           | 0%      | 100%    | 0%      | 0%      |
| Operations                   | Operate according to regulations | 0.5%    | 45.4%   | 15.4%   | 38.7%   |
|                              | Illegal operation        | 0%      | 3.6%    | 2.2%    | 94.2%   |
| Weather                      | Normal                   | 0.2%    | 33.9%   | 11.2%   | 54.7%   |
|                              | Abnormal                 | 14.3%   | 14.3%   | 42.9%   | 28.5%   |
| Crossing Environment         | Normal                   | 0.3%    | 5.4%    | 3.6%    | 90.7%   |
|                              | Pedestrian appearance    | 0%      | 100%    | 0%      | 0%      |
|                              | Motor vehicles           | 3.2%    | 9.3%    | 53.2%   | 34.3%   |
|                              | Non-motor vehicles       | 0%      | 0%      | 100%    | 0%      |
|                              | Livestock                | 0%      | 0%      | 100%    | 0%      |
| Operating External Environment | Normal           | 0.4%    | 34.4%   | 10.3%   | 54.9%   |
|                              | Abnormal                 | 0%      | 0%      | 70%     | 30%     |

First of all, from the point of view of the severity of the accident, no matter what factors under the railway traffic accidents, the largest number of general D-class accidents. This is in line with people's general feelings, the number of traffic accidents should be inversely proportional to its severity.

Secondly, in terms of time, location and type of accidents, the number of accidents occurred on day shift and night shift is basically the same. However, the accidents occurring in the interval are more serious than those occurring in the station, and the severity of off-road accidents and personal accidents is much higher than that of driving accidents.

For equipment, whether driving equipment, external equipment or underground cable, once the accident occurs due to equipment failure, the consequences are inescapable. Because the paper selects the general accident data, the accident consequence caused by equipment failure in this range is relatively light, most of them are general D-class accidents.

Human behavior is also a major factor affecting transport safety. Once the accident caused by the sudden illness of the operators themselves, the consequences are more serious. In contrast, the consequences of accidents due to irregularities are lighter.

For the external environment, abnormal weather relative to good weather will cause a high
proportion of general A-class accidents. This is in line with the realistic conditions, extreme weather is indeed the uncontrollable force affecting railway transportation, which will cause serious consequences. Railway crossing is also an accident-prone area, when the train passes, other traffic behavior at the crossing and the train conflict caused by the accident consequences are more serious. Whether with vehicles, pedestrians, even livestock collision, it will lead to emergency train braking, and even cause casualties. The accident result caused by the abnormal external environment of operation is light, only general C-class and D-class accidents.

4.2. Data Processing and Modelling
For experimental data sets, each piece of data contains 11 attributes, corresponding to 11 factors in table 1. Assignment of input and output variables, accident time "Day shift" is "1", "night shift" is "2"; The location of the accident "station" is "1","interval" is "2"; The type of Accident "driving" is "1", "road out" means "2", "Personal status" is "3"; The crossing environment "normal" is "0", "pedestrian appearance" means "1", "motor vehicles" is "2", "non-motor vehicle" is "3" and "livestock" is "4"; Health status of operators, equipment, operation, weather, external environment, etc. "Normal" means "0", "Abnormal" is "1". Encodes the output variable accident level "A-D" class in turn as the number "1-4", namely:

\[ Y = (1,2,3,4) \]  

After the data processing is completed, 491 accident information descriptions are transformed into vector forms jointly expressed by 11 dimensional input variables and 1 dimensional output variables, forming a matrix of 491×12, and establishing a mathematical model.

4.3. Parameter Adjustment
The random forest is implemented using Python3.7. The sample data set is divided into the training set and the test set according to the ratio of 7:3. The random forest was used to conduct the classification experiment, and the classification results were the same as the final results. The ratio of the sample size to the total sample size of the same result is used as the accuracy of its classification.[9] The accuracy of the model is 94.59%.

The random forest model can improve the accuracy of the model by adjusting the parameters. The complexity of the model can be adjusted by different parameter combinations to make its generalization error reach the minimum value. The article selects two parameters which have great influence on the accuracy of random forest model: n estimators and max features.

(1) n estimators adjustments
N estimators is the frame parameter of random forest model. It refers to the number of sub-data sets generated by putting back sampling on the original data set. N estimators is the number of decision trees. If n estimators is too small to make the model underfitting easily, increasing the n estimators value will improve the stability of the model, but it will not affect the complexity of a single model. So it is the preferred parameter for tuning parameters of random forest models.

In the initial calculations, n estimators is calculated according to the default value 10 in the scikit-learn packet. Depending on the size of the data set, with the same parameters, n estimators set to cycle every 10 times between 0-200. That is, 1,11,21,......output the corresponding model accuracy value and compare. Edit the code and run it to get results: when n estimators=21, the highest accuracy was 95.76%. An increase of 1.17% compared to the original accuracy rate of 94.59%.

(2) max features adjustments
Max features is the decision tree parameter of random forest model. It refers to the maximum number of features considered in constructing the optimal model of decision tree. It limits the number of features that can be considered when branching, and anything beyond that will be discarded. Increasing max features will increase the model complexity and vice versa.

In the initial calculations, the value max features is its default auto value. It is the square of the total number of features. The total number of features in this dataset is 11, and its auto value is 3.
to the number of features, if the other parameters are constant and n estimators=21, max features set to cycle every 1 between 3-11, output corresponding model accuracy values for comparison. Edit the code and run it to get results: when max features=5, the highest accuracy of the model was 95.93%. Added 1.34% accuracy after adjusting the n estimators.

Thus the optimal combination of parameters corresponding to the random forest model in this dataset is: n estimators=21, max features=5, which can make the model accuracy reach 95.93%.

4.4. Model Results
By adjusting the random forest parameters, set the parameters n estimators=21, max features=5, and the remaining parameters are the default values. The variable importance score of the data set was calculated under the condition that the highest accuracy of the model was 95.93%. the results of the variable importance score of each factor from large to small are shown in table 2.

| Factors                  | Importance |
|--------------------------|------------|
| Crossing Environment     | 0.4764     |
| Type of accident         | 0.3502     |
| Operation                | 0.0404     |
| Location of accident     | 0.0391     |
| Driving equipment status | 0.0295     |
| Accident time            | 0.0191     |
| Weather                  | 0.0163     |
| Operating External Environment | 0.0139 |
| Health status of operators | 0.0126 |
| External equipment status | 0.0013 |
| Underground cable status | 0.0012     |

In the classification process, the two most important indicators of classification effect are crossing environment (0.4764) and type of accident (0.3502). Followed by operation (0.0404), location of accident (0.0391). The higher the variable importance measures, the more obvious the effect on the classification of random forest, which indicates that it has more influence on the occurrence of accidents and the final consequences than other factors.

According to the calculation results, the four most important factors affecting the severity of railway traffic accidents are crossing environment, type of accident, operation and driving equipment status. Therefore, in railway transport operations we should pay special attention to these factors and strengthen management in these four aspects. The environment of the crossing should be guaranteed to be normal, and other traffic modes should be strictly prevented from intruding into the crossing when the train passes. For the frequent crossing of accidents, the number of staff can be reasonably increased and the accident outside the road can be avoided. In addition to regular maintenance of the driving equipment, more attention should be paid to the health status of operators, regular health checks. Workers with potential disease should be properly arranged and transferred from front-line production posts to prevent personal accidents. In the aspect of operation, safety training and operation training for operators are carried out to enhance safety awareness from the root causes. And to strengthen supervision and management of the operation process, so as to restrain the violation of rules and regulations in the bud.

5. Conclusions
There are many factors that affect the severity of railway traffic accidents. The article selects 491 general accident data from a railway transportation enterprise in 2019, extracts 11 different factors that affect the severity of accidents. The random forest was used to calculate the variable importance measures of each factor and analyze the importance degree of the factors affecting railway traffic
accidents. The validity of random forest in railway traffic accident causation analysis is verified.

Random forests are actually constructed as a black box model, which is less interpretable than traditional methods. Because it is not possible to give specific functional expressions. As a new machine learning model, random forest has the advantages of more accurate prediction, higher tolerance to outliers and noise, and higher accuracy in analyzing railway traffic accidents. However, random forest has its own defects. In the future, random forest can be combined with other models to form a new model to analyze railway traffic accidents and improve its accuracy.

Through the analysis of the cause of railway traffic accident by random forest, the calculation results show that the four most important factors affecting the severity of railway traffic accident are crossing environment, the type of accident, operation and driving equipment. In view of these four factors, relevant suggestions are made in railway transportation operation. It is of great significance to reduce the accident rate and improve the safety management level of railway transportation.

References
[1] Qin Y, Ma H and Jia L M 2014 The theoretical model of accident cause based on P-SR and its application in rail transit Chinese Railways 1213-8
[2] Xin H W 2016 Research on Railway Accident Causation Modeling Analysis (Beijing: Beijing Jiaotong University)
[3] Zhan Q J 2017 HFACS-RAs Based Railway Accident Casual Factor Modeling and Hybrid Learning Approach (Beijing: Beijing Jiaotong University)
[4] Lam C Y and Tai K 2020 Network topological approach to modeling accident causations and characteristics: Analysis of railway incidents in Japan Reliability Engineering and System Safety 193
[5] Chai T 2018 The Research of Traffic Accident Duration Prediction Based on Machine Learning (Harbin: Harbin Institute of Technology)
[6] Wang S X 2017 The Analysis and Prediction of Railway Accident Based on Data-Driven Method (Beijing: Beijing Jiaotong University)
[7] Liang C, Ghazel M and Cazier O 2018 Using bayesian networks for the purpose of risk analysis at railway level crossings IFAC Papers OnLine 51 (9)
[8] Zhou X Y, Lu P, Zheng Z J, Tolliver D and Keramati A 2020 Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree reliability Engineering and System Safety 200
[9] Xu X K and Wang J 2018 Research on classification model of emergencies based on random forest Journal of Safety Science and Technology 14 77-81