Traffic & Transportation

Research on Risk Factors Affecting High-Speed Railway Delays in China Based on Association Rules

Aidi WANG[1], Yingying XING[2], Hong LANG[3], Hongwei WANG[4], Jian John LU[5]

Original Scientific Paper
Submitted: 24 Dec 2024
Accepted: 22 Apr 2025
Published: 24 Feb 2026

[1] aidiwang@tongji.edu.cn, Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, China
[2] Corresponding author, yingying199004@tongji.edu.cn, Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, China
[3] honglang@tongji.edu.cn, Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, China
[4] wang_hongwei@ihpc.a-star.edu.sg, Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
[5] jianjohnlu@tongji.edu.cn, Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, China

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT

This study aimed to investigate the contributing factors to high-speed railway (HSR) delays and their interdependency in China. A total of 420 records of high-speed railway delays in China were collected, and 15 risk factors related to high-speed railway delays were extracted. Descriptive statistics were used to illustrate the causes of HSR delays in terms of device, personnel and environmental conditions. The association rule mining technique was further applied to explore contributing factors that cause HSR delays, revealing the underlying mechanisms of delay occurrence. The results show that HSR delays of more than 1 hour are most likely caused by foreign objects hanging on the catenary. Moreover, 65% of HSR delays are within the range of 9–30 minutes. Among delay events with durations of 9 to 30 minutes, about 9% of are caused by the fault of on-board train control equipment, mainly the Automatic Train Protection (ATP). Regarding short delays of HSR within 8 minutes, the most likely cause is platform screen door faults, followed by traveller's misconduct and train door faults. This study offers transportation agencies insights into HSR delay causes and aids in developing policies and engineering measures to reduce delays.

KEYWORDS
high-speed railway; delay; data mining; association rule; risk factors.

1. INTRODUCTION

The high-speed railway has experienced rapid development in recent years, driven by its safety, speed and convenience. Delays in high-speed railway operations occur when trains fail to arrive at stations according to the scheduled operating timetable due to accidents. Once high-speed trains experience delays, the exclusive and competitive nature of railway transportation resources amplifies the horizontal and vertical propagation effects along the operational lines. This is particularly significant in sections with high train density, as high-speed railway delays disrupt the established order of train schedules, increase railway operation costs and pose challenges in the optimal utilisation of transportation resources. The negative impact on the reliability and punctuality of high-speed railway operations is substantial. Therefore, the prompt and accurate identification of various emergencies during high-speed train operations, as well as the scientific and efficient response to these emergencies, are urgent issues faced by high-speed railway management [1].

Current research on high-speed railway delay focuses on the discovery of the delay time law of high-speed railway and the prediction of delay propagation/recovery time. Wen et al. [2] investigated the distribution law of initial delay points of high-speed railway based on actual performance data, they found that the lognormal distribution could fit the initial delay point distribution well, and their inverse model fitted the distribution of the number of trains affected by the initial delay points optimally. However, there was no detailed study on the fitting of the distribution of the initial delay points and the number of trains affected by each type of contributing factor. Lessan et al. [3] studied the delay distribution law from the perspective of train interval running time. The case results showed that the log-logistic Stee distribution works best for the distribution of train interval running time. Huang et al. [4] analysed the initial delay recovery mechanisms of high-speed trains by utilising real-world performance data from the Wuhan-Guangzhou high-speed railway. They employed a random forest model to forecast the recovery of initial delays but did not specify the impact of various contributing factors. Steven et al. [5] developed a mathematical model of delay propagation based on interval travel and tracking interval redundancy, and achieved theoretical prediction of cumulative delay duration, delay propagation boundary and delay settling time based on the input initial delay points under the assumption of redundancy time homogeneity. Barbour et al. [6] used Support Vector Regression (SVR) to estimate freight delays and utilised origin-specific characteristics, train priority and number of trains as influencing factors. The results show a valid improvement relative to the historical average forecast baseline. Marković et al. [7] applied SVR to train delay analysis for the first time, extracting the factors affecting train delays from actual train operation performance data as input to the SVR. The results were compared with those of commonly used artificial neural networks, which demonstrated the high accuracy performance of the model. Corman and Kecman [8] modelled the variation of train delays with time as a stochastic process and proposed a Bayesian network-based delay propagation prediction model. The results showed that the method was effective in predicting train delays within 30 minutes. Nair et al. [9] proposed a large-scale integrated prediction model for predicting train delays. This integrated model used two statistical models and a simulation model to generate train delay predictions. Huang et al. [10] proposed a train delay prediction model based on a combination of the Fully Connected Neural Network (FCNN) and the Long Short-Term Memory (LSTM) artificial neural network. The model also used weather-related factors as feature inputs, which is of great significance to the study of train delay prediction models. Ilalokhoin et al. [11] proposed a performance metric centred on train and passenger delay minutes. Their findings revealed that a failure in one of London's most critical electricity traction power grids could lead to disruptions affecting 75% of trains in the southern region and 25% of trains nationwide.

Overall, many studies have focused on a limited set of common factors and have not systematically considered several other important factors. Additionally, the interactions between various risk factors have not been thoroughly investigated. In terms of understanding the delay patterns of high-speed railways, most current research relies on statistical models to fit existing operational data and predict delays. However, a deeper analysis of the underlying causes of these delays remains insufficient.

High-speed rail is a highly complex and integrated system involving various types of equipment and facilities, and it is always under the harsh working conditions of high speed, high density, high intensity and heavy load in daily operation. Moreover, it frequently contends with external environmental factors, such as natural disasters and human errors, further adding to its challenges. The interplay of these factors introduces considerable uncertainty to the safe and punctual operation of high-speed trains.

Statistics on the final punctuality rate of passenger trains in European railways for 2005 indicated that equipment, transportation organisation and human factors were the primary contributors to delays in train operations [12]. Zhuang et al. [13] classified the causes of high-speed railway delays into seven categories: vehicle failure, ATP failure, track failure, pantograph and signalling system malfunctions, foreign object intrusion, adverse weather conditions, and issues related to organisation and management. Wang et al. [14] proposed a Bayesian network that incorporated expert knowledge using Dempster-Shafer evidence theory. The network structure was subsequently refined through a test for conditional independence. Goverde [15] conducted a systematic study on the factors causing train delays and pointed out that factors such as railway infrastructure equipment, traveling environment and dispatching command can affect the normal operation of trains and are the most important causes of train delays. Yang [16] analysed that the main influencing factors of high-speed railway delays can be divided into three categories: device factors, personnel factors and environmental factors. Among these, device factors account for 86% of all delay events and are the primary cause of train delays.

Given that current research on HSR delay mainly focuses on the distribution law and prediction of delay. There is limited research on the contributing factors of HSR delay. Moreover, existing studies of HSR delay have explored the most likely factors or combinations that contribute to delay. However, most of them are based on independent analyses that ignore the correlations and interactions among factors contributing to HSR delay. This study employs the Apriori-based association rule mining algorithm to analyse the risk factors affecting high-speed railway delays. It aims to explore potential factors or combinations that lead to delays, identify interdependencies between different factors and the resulting high-speed railway delays, and further explain the mechanism of delay occurrences. Compared to methods such as fault tree analysis, accident tree analysis and Petri nets, association rule mining not only offers a comprehensive description of the dependencies among various influencing factors and identifies key system components, but also enables a quantitative analysis of the correlations between these factors. Finally, effective countermeasures and recommendations to reduce the delay rate of the high-speed railway are proposed for the high-speed rail department to formulate corresponding policies and measures.

2. DATA

This section introduces the data sources and processing methods used for HSR delay incidents, and presents a descriptive statistical analysis of the risk factors identified as contributing to HSR delays.

2.1 Data collection

A total of 550 original fault records related to high-speed railways, covering the period from 1 November 2018 to 31 March 2019 were collected from the Guangzhou Railway Group. These records included detailed information on fault conditions, handling processes, the number of affected trains and other related data. After eliminating invalid and missing records, 420 valid records about high-speed railway delay were extracted. All the records were collated and analysed to extract the influencing factors of high-speed railway delay. In this paper, the direct factors of HSR delay are divided into personnel, device and environment. The direct factors, also referred to as parent factors, each encompass a set of related child factors. *Table 1* presents the descriptions and value sets for these factors. In addition to the above factors, other equipment faults or alarms, staff errors and other contingencies are grouped separately in the other factors category. Moreover, the indirect factor considers the slope condition of the train stopping location, which affects the delay probability by influencing the device status.

Table 1 – Attribute value encoding

Category	Factors	Designation	Attribute value
Facility and equipment factor	On-board train control equipment (mainly ATP system fault)	Fac_1	—
	Signal equipment fault	Fac_2	—
	Platform screen door fault	Fac_3	
	Catenary blackout (trip)	Fac_4	
	Objects hanging on catenary	Fac_5	
	Pantograph fault	Fac_6	
	Automatic lowering pantograph during operation	Fac_7	
	Traction converter fault	Fac_8	
	Train door fault	Fac_9	
	Carbody’s vibration	Fac_10	
Human factor	Passenger misconduct or sudden illness	Fac_11	—
	Personnel intrusion of railway clearance	Fac_12	
Environment factor	Driving in heavy rain or snow	Fac_13	—
	Railway clearance intrusion by objects	Fac_14	
	Train collision against objects	Fac_15	
Other factor	Other faults	Other_Fac	other equipment fault or alarm (1), staff error (2), other circumstances (mainly abnormal sound) (3)
	Vehicle stop position slope	Slope	up (1), down (2)
Delay time		Time	≤ 8min (1), 9–30min (2), 31–60min (3), ＞1h (4)
Number of trains affected		Number	1 (1), 2–10 (2), ＞10 (3)

2.2 Descriptive statistical analysis of contributory factors

Factors associated with facilities and equipment encompass ten specific categories, as illustrated in *Figure 1*. These factors contribute to 61.5% of the overall delays, significantly higher than those caused by human and environment factors. Among them, ATP, platform screen doors, and communication and signalling system faults were more influential than other equipment factors, accounting for 16.9%, 11.0%, and 8.1% of all delay events, respectively.

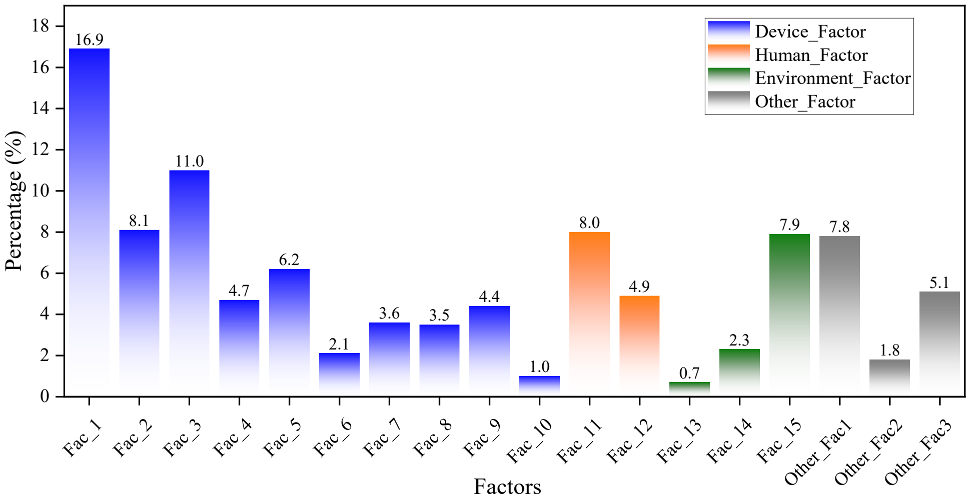


Figure 1 – Distribution of high-speed railway delays by different factors

Delays caused by personnel factors account for 12.9% of the total. Two significant personnel factors are passenger misconduct or sudden illness (FAC_11) and personnel intrusion into railway clearances (FAC_12). *Figure 1* shows that FAC_11 contributes to 8.0% of delays, which result from improper or even illegal behaviours of passengers, such as smoking triggering smoke alarms. This factor also includes situations where passengers with sudden illnesses need to disembark for treatment. FAC_12 refers to instances where residents living along the railway line sometimes encroach upon railway clearance, posing potential dangers to traffic safety. Notably, as staff management improves and high-speed railways become more automated, the number of delays caused by operator errors continues to decline [14].

Delays led by environment factors account for 10.9% of the total. It is mainly divided into three factors: bad weather, foreign object intrusion and foreign object collision, among which the probability of foreign object collision is the highest, reaching 7.9%, and the impact of bad weather is lower, only 0.7%. Owing to the inherent stability and safety features of high-speed railways, weather variations generally have minimal impact on their operation. However, bad weather may indirectly cause other equipment fault. For example, thunderstorms may cause catenary tripping and blackout, thus causing train delays. Although high-speed railway mostly uses elevated bridge bases, which have better drainage, excessive rainfall can also cause trains to run at limited speed. Usually, high-speed railway is installed with high wind alarm system and rainfall overrun alarm system to ensure the safe operation of trains. Railway clearance, specifically referring to structural clearance here, represents a cross-sectional profile perpendicular to the railway's centreline. In this study, railway clearance intrusions by objects (such as fallen rocks or small animals) are considered only when they do not result in a collision with trains (FAC_14). In cases where a collision occurs, the incident is classified as a train collision with objects (FAC_15).

Delays attributed to other factors rank second only to device factors. This category is subdivided into three distinct causes: other equipment fault or alarm (7.8%), staff error (1.8%) and other circumstances (mainly abnormal sound) (5.1%). Among them, the "other circumstances (mainly abnormal sound)" reflects incidents where trains experienced abnormal sounds during operation, often requiring temporary stops for inspection. These abnormal sounds typically originate from the train's undercarriage, running gear or driver's cab. In most cases, mechanics confirmed no actual faults after inspection, and the train resumed normal operation. However, in rare instances, factors such as foreign objects, ice build-up or traces like bloodstains on obstacle removers were identified. Despite not involving a major structural or mechanical failure, these incidents resulted in delays as a precautionary measure to ensure passenger safety.

The most direct consequences of high-speed railway accidents are the delay of time (DOT) and the influence on subsequent trains (IOST). These two effects are easy to observe and record. It is important to note that no accidents resulting in injuries or fatalities were recorded. As illustrated in *Figure 2*, the dataset shows a high frequency of events associated with shorter delays, and a distinct critical threshold is identified at approximately 8 minutes. Specifically, delay events occurring within 8 minutes are highly frequent, whereas the number of events drops markedly beyond this duration. Therefore, 8 minutes are classified as a critical cut-off point for short delays. In *Figure* 2, the distribution of delay times reveals a more nuanced pattern. The delays in the range of 0–4 minutes (18.8%) and 4–8 minutes (16.4%) exhibit the highest proportions. Furthermore, delays exceeding 60 minutes, though less common individually, collectively contribute to a substantial portion of the data, emphasising the impact of prolonged disruptions. Similarly, in the revised *Figure 3*, while a single train delay is most prevalent (48.7%), the presence of clusters involving more than 10 affected trains underscores the occasional occurrence of severe incidents that significantly impact the network. This refined breakdown allows for a better understanding of both minor and major delay events.

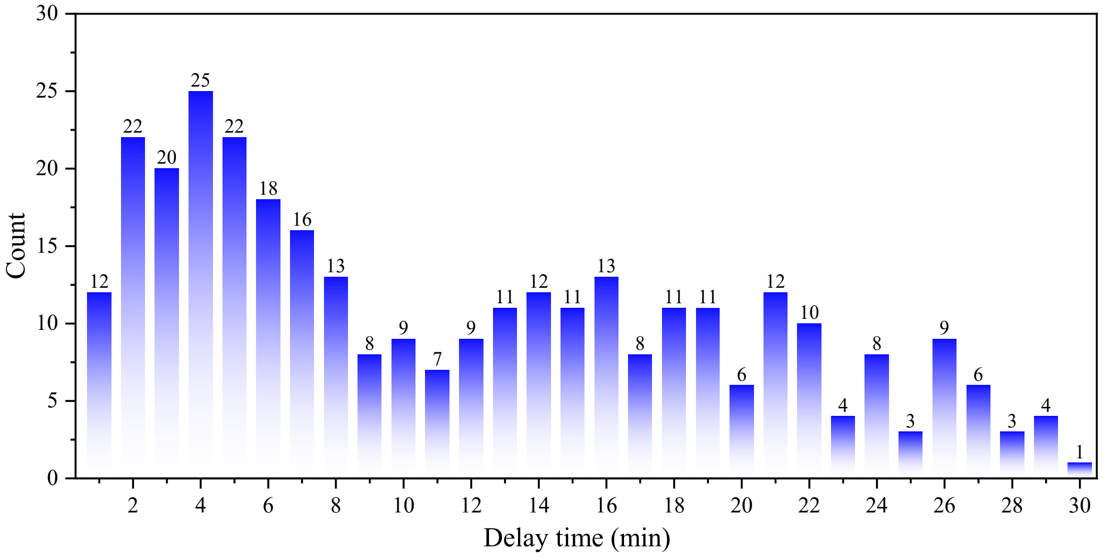


Figure 2 – Delay distribution for events included in the dataset

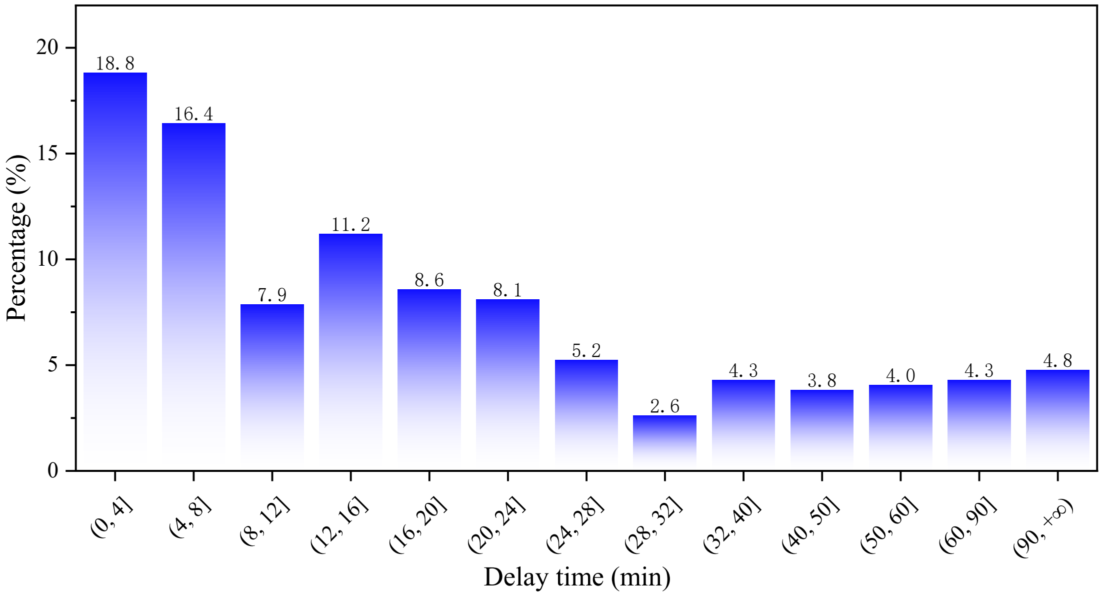


Figure 3 – Distribution for each delay period

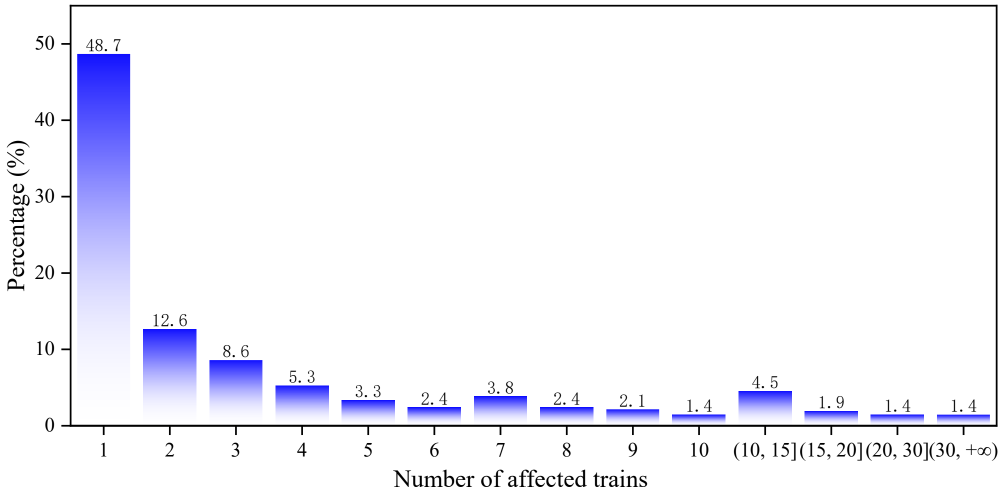


Figure 4 – Distribution of the number of affected trains

3. DATA

Previous literature has shown that correlations between independent variables seriously hinder the statistical analysis of accident data and may lead to erroneous conclusions [17]. While data mining techniques such as Classification and Regression Trees (CART) can circumvent the problems caused by correlations, they do not provide quantitative analysis for these correlations. The association rule mining used in this paper is an important data mining method, which aims to find the correlations between items in the data. If k denotes the number of categories, the process can be intuitively understood as identifying k distinct sets of correlations among items. Compared to traditional statistical methods and artificial intelligent techniques, association analysis offers the advantage of not requiring prior designation of dependent and independent variables. This flexibility enables the identification of valuable relationships that might otherwise remain hidden. In recent years, association rule analysis has gained widespread application across various fields, including road traffic safety research [18, 19], occupational injury analysis in the construction industry [20], injury analysis in steel plants, etc. [21]. Association rule mining, with interpretable results and the ability to mine all potential relationships in a dataset, is more suitable for analysing observational data collected outside the scope of designed experiments, such as high-speed railway delay data. The association rule in the paper is defined as follows: suppose I_m is the m^{th} record in the HSR delay database; $I=\{I_1,I_2,...,I_m\}$ is the set of all items in the database [22]; X and Y are any set of k-items, where, $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \varnothing$. The minimum support is denoted as minsup, the minimum confidence is denoted as minconf and the minimum lift is denoted as minlift. A rule is considered a strong association rule if it satisfies the following two conditions:

1) The support of the term set $X \cup Y$ is at least minsup;
2) The rule $X \to Y$ has a confidence of at least minconf and a lift of at least minlift.

The above mentioned support, confidence and lift are the three important metrics used to mine association rules. Their definitions and calculation formulas are as follows:

1) Support. The support $sup(X \cup Y)$ of an item set $X \cup Y$ is the proportion of records containing the subset $X \cup Y$ of all HSR delay event records in the HSR delay event database and can be calculated by *Equation 1*.

$$\sup(X \cup Y) = \frac{X \cup Y}{N} \tag{1}$$

where N is the total number of records in the HSR delay database; $X \cup Y$ is the number of records in the database where both item set X and item set Y occur.

2) Confidence. The confidence $\text{conf}(X \to Y)$ of rule $X \to Y$ is the conditional probability of $X \cup Y$ in the records containing X and can be calculated by *Equation 2*.

$$\text{conf}(X \to Y) = \frac{\sup(X \cup Y)}{\sup(X)} \tag{2}$$

3) Lift. The lift $\mathrm{lift}(X \to Y)$ of the rule $X \to Y$ is the ratio of the confidence $conf(X \to Y)$ to the support $\sup(Y)$. See *Equation 3*.

$$\mathrm{lift}(X \to Y) = \frac{\sup(X \cup Y)}{\sup(X) \cdot \sup(Y)} \tag{3}$$

The set of terms X and Y are the antecedent and the consequent of this rule, respectively. The lift can be understood as the ratio of the probability of X under the condition of Y to the probability of Y. A lift greater than 1 indicates that the antecedent X has a facilitating effect on the consequent Y; a lift less than 1 indicates the opposite; a lift equal to 1 indicates that X and Y are independent of each other. The higher the lift, the stronger the relevance of the rule.

The general framework for generating association rules has two stages, each corresponding to the two conditions that must be satisfied by the strong association rules mentioned above. The first stage is much more complex and time-consuming than the second stage, which generates all frequent item sets. Suppose that the HSR delay database $T = \{T_1, T_2, ..., T_n\}$ is the set of all HSR delay event records, where each record T_i is a non-empty subset of the item space I, corresponding to a unique identifier T_D. The objective is to find all item sets $U = X \cup Y$ in T such that the proportion of records including subset U to all HSR delay records is no less than a predetermined minsup [23].

This study uses the Apriori algorithm for frequent pattern mining proposed by Agrawal and Srikant [23] as the fundamental algorithm for mining the contributing factors of high-speed railway delay. The Apriori algorithm is one of the most fundamental and widely used algorithms in association rule mining. Its core idea is to mine the set of frequent items in a transaction, use several iterations to calculate the set of frequent items in the database and discover strong associations among them. The Apriori algorithm involves two critical processes: concatenation and pruning. In the concatenation step, the frequent (k-1)-items set L_{k-1} is concatenated with itself to generate the candidate k-items set $C_k = \{C_{k1}, C_{k2}, ..., C_{kn}\}$, where each $C_{ki} (i = 1, 2..., n)$ represents a candidate set of k-items. The pruning step leverages an essential property known as "a priori property," which states that any infrequent k-items set is not a subset of the frequent k-items set, and thus removes $C_{kj} (j \in i)$ with infrequent subsets from C_k. Based on this, the k-items set that satisfies minsup, denoted as the frequent item set L_k, is further filtered from C_k. The complete Apriori algorithm is illustrated in *Figure 5*.

```
Algorithm Apriori (HSR Delay Event Database: $T$, Minimum support: minsup)
begin
        $k = 1$
        $L_1 = \{C_{1i} \in C_1 \mid C_{1i}$.count≥minsup}; # Generate frequent 1-item set;
        while $L_{k-1}$ !=Null, do begin
                $C_k =$ sc_candidate($L_{k-1}$); # $L_{k-1}$ is concatenated with itself to generate $C_k$;
                for each $T_i$ in $T$
                        $C_k = \{C_{ki} \in C_k \mid i$ !=$j$ }; # Remove $C_{kj}(j \in i)$ with infrequent subsets
from $C_k$ according to the priori property;
                  $L_k = \{C_{ki} \in C_k \mid C_{ki}$.count≥minsup}; # Retain the $k$-terms set in $C_k$ with
                  support at least minsup, constituting $L_k$;
                $k = k+1$;
        end
        return $(\bigcup_{i=1}^{k} L_i)$ ;
end
```

Figure 5 – Apriori algorithm

As shown in *Figure 5*, first, the Apriori algorithm scans the database *T*, and items that satisfy the minimum support are retained to produce the frequent itemset of length 1. Next, a new candidate *k*-items set is generated by concatenating the frequent *k*-1-items set L_{k-1} with itself. According to the priori property, if the *k*-items subset of a candidate k-items set is not in L_{k-1}, then the candidate cannot be frequent either. Thus, for each candidate $C_{kj} \in C_k$, check whether all its subsets are in L_{k-1}. If not, prune this itemset C_{kj} from C_k. After the candidate itemset C_k of length *k* is generated, the support of each itemset can be determined by counting the number of occurrences of each candidate in the HSR delay event database d. Only candidates that satisfy the minimum support are retained to form the set $L_k \subseteq C_k$ of frequent *k*-items set. The algorithm repeats the above steps iteratively to find frequent item sets of different lengths until the set L_k is empty, at which point the algorithm terminates. After the algorithm terminates, the frequent item sets of different lengths are calculated as the final output of the algorithm [23].

After the frequent item sets are found using the Apriori algorithm, all the true subsets of the algorithm output $\bigcup_{i=1}^{k} L_i$ are formed into alternative association rules in a permutation and combination manner. The confidence and the lift of the rules are calculated to filter out the final eligible strong association rules.

4. RESULT AND DISCUSSION

When applying association rule analysis to investigate the combinations of factors that commonly occur together in high-speed railway delay events, threshold values for support (S), confidence (C) and lift (L) need to be pre-specified. Considering the actual situation of this study, the three metrics were set as follows: support ≥ 2%, confidence ≥ 35% and lift ≥ 1.2. To facilitate the analysis, some important association rule mining results after filtering are sorted according to the lift from largest to smallest. The results are presented in *Table 2*, while *Figure 6a* and *6b* illustrates the causal chains of the representative findings.

Table 2 – Association rules

Rule ID	Association rules		Sup	Conf	Lift
	Antecedent	Consequent			
1	Fac_5	Time_4	0.02	0.35	3.83
2	Fac_15	Slope_2	0.03	0.39	3.31
3	Fac_15	Slope_1	0.03	0.42	3.02
4	Fac_3	Time_1	0.10	0.93	2.65
5	Other_Fac_3	Number_2	0.05	0.91	2.17
6	Fac_11	Time_1	0.06	0.74	2.09
7	Fac_3	Number_1	0.10	0.96	1.96
8	Fac_11	Number_1	0.08	0.94	1.93
9	Fac_15	Time_2	0.06	0.76	1.80
10	Fac_9	Time_1	0.03	0.63	1.79
11	Slope_2	Time_2	0.09	0.74	1.76
12	Fac_7	Number_2	0.03	0.71	1.68
13	Time 1	Number_1	0.29	0.82	1.68
14	Slope_2	Number_2	0.08	0.70	1.67
15	Other_Fac_3	Time_2	0.04	0.68	1.62
16	Fac_12	Number_2	0.03	0.67	1.59
17	Fac_8	Number_1	0.03	0.75	1.54
18	Fac_7	Time_2	0.03	0.65	1.54
19	Fac_15	Number_2	0.05	0.64	1.52

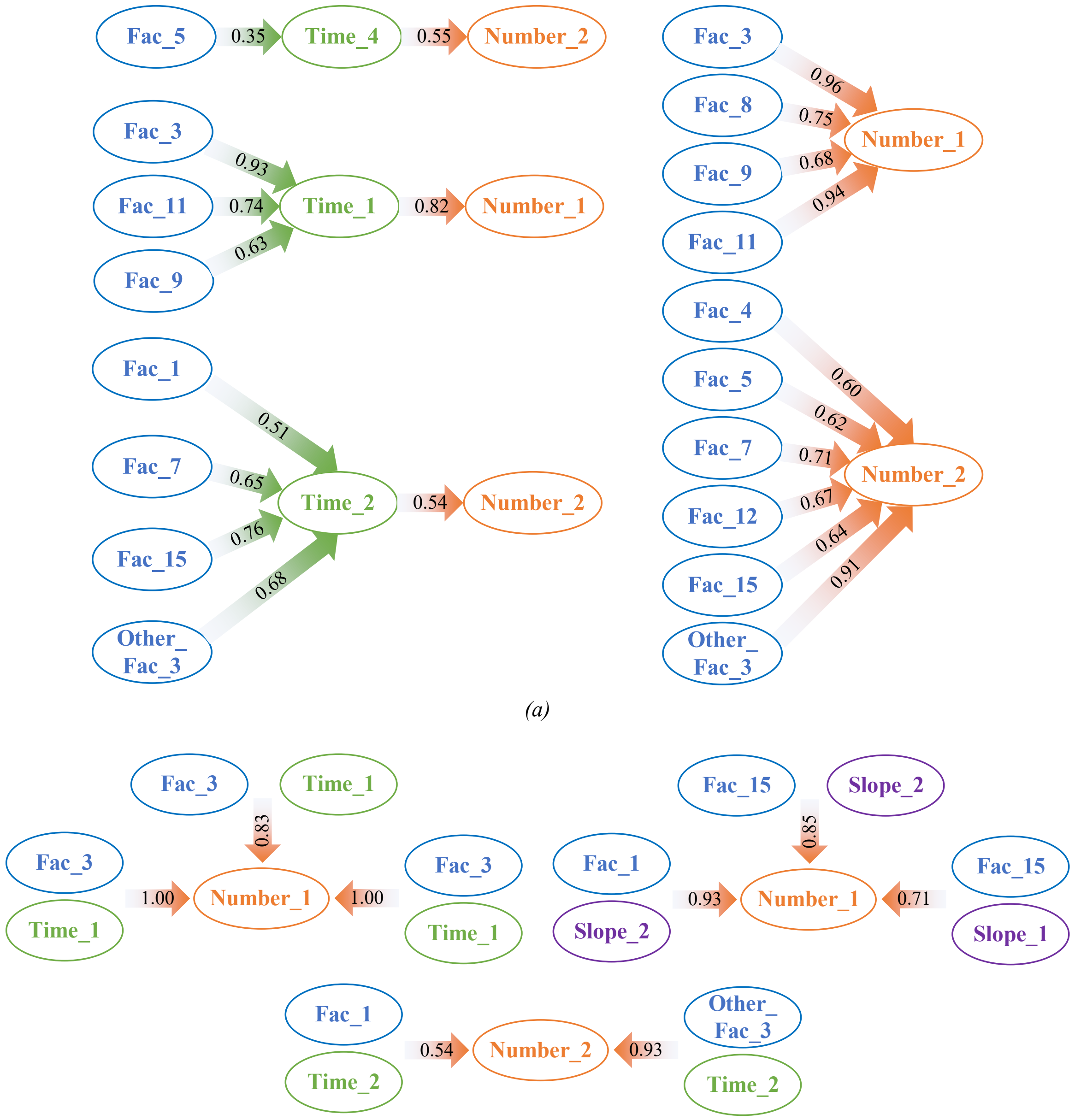
Rule ID	Association rules		Sup	Conf	Lift
	Antecedent	Consequent			
20	Slope_1	Number_2	0.09	0.63	1.50
21	Fac_5	Number_2	0.04	0.62	1.47
22	Fac_4	Number_2	0.03	0.60	1.43
23	Time_3	Number_2	0.08	0.60	1.42
24	Fac_9	Number_1	0.03	0.68	1.40
25	Number_3	Time_2	0.05	0.56	1.34
26	Slope_1	Time_2	0.08	0.56	1.33
27	Time_4	Number_2	0.05	0.55	1.32
28	Time_2	Number_2	0.23	0.54	1.28
29	Fac_1	Time_2	0.09	0.51	1.22
30	Fac_15 & Number_2	Slope_1	0.03	0.52	3.73
31	Fac_15 & Time_2	Slope_2	0.03	0.44	3.70
32	Fac_3 & Number_1	Time_1	0.10	0.98	2.77
33	Time_2 & Other_Fac_3	Number_2	0.03	0.93	2.23
34	Fac_11 & Number_1	Time_1	0.06	0.78	2.22
35	Fac_1 & Slope_2	Time_2	0.03	0.93	2.20
36	Number_1 & Fac_9	Time_1	0.02	0.77	2.18
37	Fac_3 & Time_1	Number_1	0.10	1.00	2.05
38	Fac_11 & Time_1	Number_1	0.06	1.00	2.05
39	Fac_15 & Slope_2	Time_2	0.03	0.85	2.01
40	Fac_15 & Slope_1	Number_2	0.03	0.79	1.88
41	Fac_9 & Time_1	Number_1	0.02	0.83	1.71
42	Fac_15 & Slope_1	Time_2	0.02	0.71	1.69
43	Number_2 & Slope_2	Time_2	0.06	0.71	1.69
44	Slope_1 & Time_2	Number_2	0.05	0.70	1.66
45	Other_Fac_3 & Number_2	Time_2	0.03	0.70	1.66
46	Time_2 & Slope_2	Number_2	0.06	0.68	1.61
47	Number_2 & Fac_1	Time_2	0.05	0.67	1.58
48	Fac_1 & Number_1	Time_1	0.05	0.55	1.57
49	Fac_1 & Time_2	Number_2	0.05	0.54	1.29

Several association rules in *Table 2* are related to the traction power supply system faults. Rule 1 has the highest lift among all rules at 3.83, indicating that foreign objects hanging on the catenary are the most probable cause of delays exceeding one hour on the high-speed railway. Rule 21 and Rule 22 indicate a high likelihood that delays will occur due to foreign objects hanging on the catenary or catenary blackout. Moreover, the interaction between trains may disrupt the established operating schedule of adjacent trains. In addition, the confidences of Rule 12 and Rule 18 are 0.71 and 0.65, respectively, which indicates that in high-speed rail delays caused by pantograph malfunctions, 71% of the incidents will affect two or more trains, resulting in cascading delays. Moreover, 65% of the delays last between 9 to 30 minutes. Rule 17 reveals that 75% of high-speed railway delays, caused by the power supply system faults, such as traction converter faults, affect only the train with faults. Analysing the reasons for the aforementioned association rules, the following conclusions can be drawn: the catenary plays a crucial role in the power supply of the high-speed railway. Since the catenary is at a high distance above the tracks, it is susceptible to external influences and is easily entangled by objects such as balloons, kites and plastic bags, leading to tripping or pantograph fault. In the case of particularly small foreign objects that have minimal impact on train operation, the driver can lower the pantograph and decelerate to avoid them, thereby reducing the impact on railway transport. However, for larger foreign objects, it is necessary to apply for temporary blockade: instruct the trains on the line to stop, disconnect the catenary power supply equipment, conduct manual cleaning by maintenance personnel, and in some cases, mobilise the catenary operation vehicle for assistance. This significantly increases the delay time of the high-speed railway. If the pantograph carbon sliding plate breaks or wears to the limit, the device will trigger an automatic pantograph lowering fault. In such cases, the train should be stopped immediately, and the mechanic should be notified to handle the situation. The faulty pantograph should be removed, and the train should proceed to the nearest station for further inspection. This series of troubleshooting work will significantly disrupt the high-speed railway operation plan, leading to HSR delays. It is worth noting that the support of Rule 28 reaches 0.23, indicating that high-speed railway delay events with train delays of 9-30 minutes and affecting 2-10 trains account for 23% of all delay events. In order to avoid potential traveling safety accidents caused by disturbances from the train ahead, trailing trains must track at a limited speed, resulting in collateral delays. This interaction between trains becomes more obvious, particularly for less flexible train operation diagrams with insufficient reserved redundancy time. In comparison to catenary and pantograph faults, the severity of high-speed railway delay events caused by traction converter faults is lower. One possible explanation is that the high-speed railway system in China is distributed, meaning that traction converters are distributed across multiple cars in the train. When a traction converter fault occurs, it can be promptly switched to another one, resulting in little impact on the entire train.

Rule 13 indicates that 29% of HSR delay events are single train delays lasting 8 minutes or less. The confidence of this rule is 0.82, indicating that 82% of these short time delay events within 8 minutes involve only the train with faults and do not affect the operation of other trains on the line. There are many factors leading to short delays on HSR. Rule 4 reveals that 10% of delay events on HSR are caused by platform screen door faults with delays of 8 minutes or less. 93% of delay events caused by platform screen door faults result in short delays of 8 minutes or less. This is likely due to the linkage failure between the platform screen door and the train door, causing the faulty platform door to send out a locking alarm, which leads to the failure of outbound signal to open. This fault is easily fixed by simply bypassing the platform screen door, and the outbound signal will be reopened, allowing the train to resume operation in a very short time. Rule 6 indicates that high-speed railway delay events caused by traveller's misconduct or sudden illness, such as a passenger blocking the train door, causing the door to fail to close, or opening a locked platform door by themselves, smoking on the high-speed railway, or mistakenly pulling the emergency brake valve, are likely to result in short delays. The most common measure to deal with sudden illness is to contact medical personnel or police at the next station and transfer passengers to them, which causes little delay and minimal damage. Similar to the linkage failure of the platform and train door, traveller's misconduct can cause the outbound signal to shut down. However, as long as the staff detects and stops it in time, the signal can be quickly reopened, allowing the train to resume normal operation after a brief treatment. The confidences of Rule 7 and Rule 8 are 0.96 and 0.94, respectively, indicating that high-speed railway delay events caused by platform screen door faults and traveller's misconduct may usually only affect the train with faults. Furthermore, from Rule 37 and Rule 38 of the 3-items association rules, it can be found that the confidence reaches 1 when the delay time within 8 minutes is satisfied simultaneously. The opening and closing of platform screen doors involve a series of facilities and equipment, including station infrastructure, train control equipment, signal equipment, and so on. Moreover, platform screen doors are frequently used and are prone to failure due to aging or lack of timely maintenance. While platform screen door faults and traveller's misconducts are common in high-speed railway delay events, effective measures can usually be promptly taken to eliminate the failures, thus avoiding their impact on subsequent trains on the line. Rule 10 and Rule 24 indicate that train door faults are likely to cause short delays for the train with faults. Additionally, the lifts of Rule 4, Rule 6 and Rule 10 are 2.65, 2.09 and 1.79 in descending order, indicating that the most likely cause of short delays within 8 minutes is platform door faults, followed by traveller's misconduct and train door faults.

In addition to on-board equipment faults and personnel factors, the operation of high-speed railways is also susceptible to interference from the external environment, such as foreign object intrusion or collision, extreme weather, etc. The occurrence of such events is random and uncontrollable, seriously affecting the train operation plan and causing widespread impact. The confidences of Rule 2 and Rule 3 are 0.39 and 0.42, respectively, indicating that high-speed railway delay events caused by foreign object collision are likely to occur at the ramp. The lifts of both rules are greater than 3, indicating that high-speed railway delays occurring at the ramp are more likely to be caused by trains colliding against foreign objects. One possible explanation is that sections of track lines with undulating slopes are usually in mountainous areas with complex terrain, where wild animals are sometimes found invading the high-speed rail tracks and posing a safety hazard to passing trains. Rule 39 and Rule 40 provide further clarification: Train collisions against foreign objects on a slope are likely to cause delays of 9–30 minutes and affect subsequent trains. Other_FAC_3 is mainly related to trains making abnormal sounds, which in most cases are still caused by the train colliding against foreign objects during travel. Therefore, it can be analysed together with FAC_15. Rule 9 and Rule 15 indicate that train collisions with foreign objects or trains making abnormal noise are likely to result in delays of 9–30 minutes. Rule 5 and Rule 19 further indicate that they will also result in collateral delays for subsequent trains. One possible reason is that high-speed trains cannot avoid collisions by emergency braking. To protect the safety of the train and prevent derailing or even overturning due to emergency braking, the train will not stop immediately. Instead, it will gradually decelerate and stop smoothly, even if the driver finds foreign objects ahead (mainly birds). Subsequently, the mechanical engineer will check the seriousness of the collision, which takes some time and leads to longer delays. As a result, subsequent trains have to limit their speed to ensure safety when passing through the collision area, thus causing a wide range of train delays.

Several other important association rules are also worth mentioning. Rule 29 has a support of 0.09, indicating that 9% of HSR delay events are caused by on-board train control equipment faults (mainly ATP) with delays of 9–30 minutes. Due to the involvement of a large number of electronic components, on-board devices are susceptible to temperature, humidity, dust, electromagnetic waves, etc., resulting in issues such as poor contact, abnormal transmission, key failure. Additionally, the difficulty in detecting and promptly dealing with internal problems with electronic components from the outside makes them prone to failure. During the operation process, the driver inputs or retrieves stored information through the on-board equipment to obtain the train's braking characteristics, receive line information and travel permissions from the ground equipment, and monitor the train's running speed in real-time. Once the fault occurs, the train will be unable to receive and transmit train and line information normally, leading to incorrect judgement of the front running environment. As a consequence, speeding and crossing the line may occur, posing a threat to the safe operation of train 0. Rule 35 and Rule 29 show that the confidence increases from 0.51 to 0.93 with the additional Slope2 factor, indicating that if the on-board train control system fails during a downhill slope, the probability of delay will significantly increase. One possible explanation is that if the ATP system fails during a downhill, the train is more likely to overspeed and lose control than on a flat road. Moreover, the driver will generally take stricter speed limit measures to decelerate to a safe range and stop at the next station to restart the ATP system, thus significantly increasing the likelihood of longer delay. Additionally, Rule 16 indicates that HSR delay events caused by personnel infringement are likely to have an impact on other trains on the line. When the train driver finds personnel intruding into the railway clearance, he needs to report to the dispatch centre and decelerate to a limited level. The train speed will return to normal only after the dispatch centre confirms that the unrelated persons have evacuated the scene [24]. China's high-speed railway is characterised by high running density, and the actual train tracking intervals are small. Once a train is disturbed by abnormal events during operation and cannot resume normal operation in time, the delay may exceed the buffer and adjustment capacity of the planned train running schedule. Consequently, the actual interval running time of the train may overlap with the planned interval running time of the trailing train. In such cases, the trailing train must take a series of measures, such as speed-limited tracking operation, to avoid conflicts with the preceding train, resulting in delays for the trailing train [1]. Due to the delay propagation phenomenon, these delays may affect multiple trains on the same line, leading to collateral delays.



(a)

(b)

Figure 6 – Chain of causality for high-speed railway delay: a) 2-item association rules; b) 3-item association rules

5. CONCLUSIONS

This study collected original fault records of high-speed railway faults in China from November 2018 to March 2019 and screened 420 valid HSR delay records. Descriptive statistics were used to illustrate the causes of high-speed railway delay in terms of device, personnel and environmental conditions. The Apriori-based association rule mining algorithm was further applied to research contributory factors that cause HSR delay and investigate the underlying mechanisms of delay occurrence.

To present the hierarchical relationship of the proposed measures for delays and their causes more clearly, a hierarchical structure of the proposed measures is illustrated, as shown in *Figure 7*. According to the results of descriptive statistics, the percentage of device-related factors in HSR delay events was 61.5%, significantly higher than that of personnel (12.9%) and environment (10.9%). This indicates that device-related factors are the most important contributors to HSR delay. The reliability and stability of the equipment are the key factors for improving the punctuality of high-speed railway. Related departments should fully utilise the skylight repair time for key equipment and prioritise overhaul and maintenance activities to ensure the normal functioning of the equipment. Among all the device-related factors, ATP (16.9%) had the highest impact, followed by platform screen door fault (11.0%) and communication and signalling system (8.1%). Therefore, reducing ATP faults is crucial for reducing the rate of HSR delay. Personnel factors represent the second most significant contributor, with 8.1% of delays caused by traveller misconduct or sudden illness. Although traveller's misconduct belongs to unexpected situations, preventive measures should be implemented to reduce its occurrence. For instance, high-speed rail crews should clearly communicate travel requirements to passengers and improve management. Additionally, railway transportation authorities should impose stricter penalties for inappropriate passenger behaviour. Moreover, related departments should enhance inspections of high-speed railway lines and roads to prevent personnel-related incidents. Environmental conditions had the third highest percentage, with foreign object collision (7.9%), foreign object invasion (2.3%) and bad weather (0.7%) being the influencing factors in descending order. Additionally, the percentage of other factors such as other device faults and staff errors was 15.0%. Regarding staff, emphasis should be placed on improving the relevant staff's expertise to reduce work errors and, consequently, the occurrence of faults and accidents. Besides, it is necessary to enhance the operational skills and emergency handling ability of field staff to address temporary situations during operation, thus reducing the likelihood of train delays.

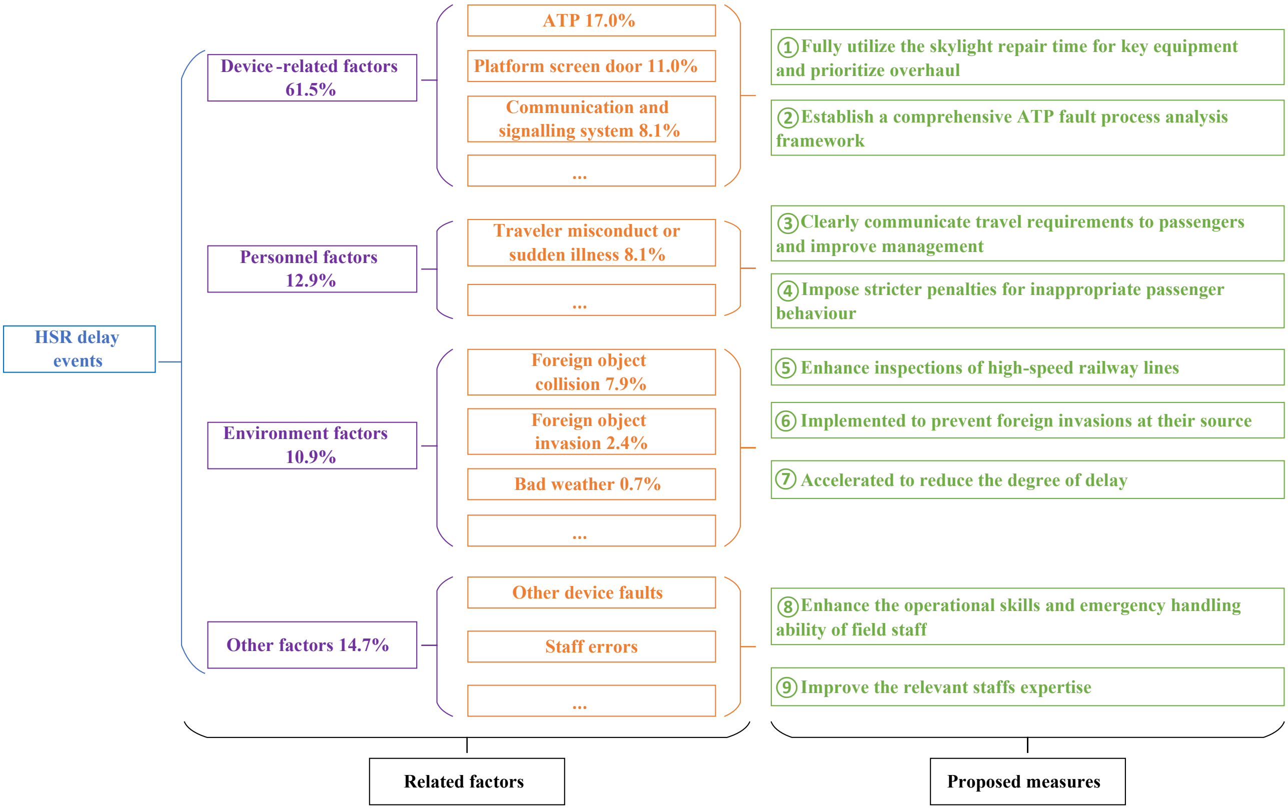

ATP 17.0%
Platform screen door 11.0%
Communication and signalling system 8.1%
...
Device -related factors 61.5%
①Fully utilize the skylight repair time for key equipment and prioritize overhaul
②Establish a comprehensive ATP fault process analysis framework
Personnel factors 12.9%
Traveler misconduct or sudden illness 8.1%
...
③Clearly communicate travel requirements to passengers and improve management
④Impose stricter penalties for inappropriate passenger behaviour
HSR delay events
Environment factors 10.9%
Foreign object collision 7.9%
Foreign object invasion 2.4%
Bad weather 0.7%
...
⑤ Enhance inspections of high-speed railway lines
⑥ Implemented to prevent foreign invasions at their source
⑦ Accelerated to reduce the degree of delay
Other factors 14.7%
Other device faults
Staff errors
...
⑧Enhance the operational skills and emergency handling ability of field staff
⑨Improve the relevant staffs expertise
Related factors
Proposed measures

Figure 7 – Hierarchical structure for mitigation measures

Additionally, analysis of the high-speed railway delay event data revealed three important association rules related to traction power system faults: (1) High-speed railways are delayed for more than one hour, most likely because of foreign objects hanging on catenary, and due to the interaction between trains, disrupts the established operating schedule of adjacent trains. (2) Of the high-speed railway delay events caused by pantograph faults, 71% will affect two or more trains, generating cascading delays, with 65% of delays ranging from 9 to 30 minutes. (3) 75% of the high-speed railway delays caused by traction converter faults only affect the train with faults. These results reveal the cause of high-speed railway delay events related with traction power system fault, offer valuable insights for developing policies aimed at preventing HSR delays.

The important association rules related to short delays are as follows: 93% of delay events caused by platform screen door fault result in short delays of 8 minutes or less, and 82% of short high-speed railway delay events of 8 minutes or less involve only the train with faults and do not affect the operation of other trains on the line. High-speed railway delay events caused by traveller's misconduct or sudden illness, such as passenger blocking the train doors, causing the doors to close, or opening a locked platform door by themselves, are likely to result in short delays for a single train. To prevent passenger misconduct events, the crew should not only strengthen publicity and education to improve passenger understanding, but also provide proper reminders and guidance. Additionally, the association rule indicates that the most likely causes of short delays of up to 8 minutes for HSR are platform door failure, followed by traveller's misconduct and train door faults.

The important association rules related to the environment are as follows: (1) High-speed railway delay events caused by collisions with objects or making strange noises are likely to occur on slopes, often resulting in delays of 9–30 minutes and leading to cascading delays of subsequent trains. (2) HSR delay events caused by personnel infringement are likely to impact other trains on the line. Therefore, it is essential to enhance management during daytime operations. On the one hand, measures should be implemented to prevent foreign invasions at their source. On the other hand, the handling process should be accelerated to reduce the degree of delay.

In addition, the following two important association rules are worth mentioning: (1) 9% of HSR delay events are caused by faults in on-board train control equipment (mainly ATP) and result in delays of 9–30 minutes. (2) A fault of the on-board train control system during a downhill slope will significantly increase the probability of delay. ATP is one of the most critical components in train control vehicles, ensuring the safe operation of high-speed railways. However, ATP faults are characterised by randomness, sudden occurrence and diversity, leading to a relatively high rate of false or missed diagnoses. To reduce potential faults and enhance the accuracy of ATP fault diagnosis, it is essential to establish a comprehensive fault process analysis framework.

ACKNOWLEDGEMENTS

This work was supported in part by the Natural Science Foundation of Shanghai (Grant NO. 22ZR1465500).

REFERENCES

[1] Chen S. Research on the method of propagation delays of high-speed railway trains based on maximum algebra and ensemble learning. PhD thesis. Beijing Jiao Tong University; 2021.

[2] Wen C, et al. Statistical investigation on train primary delay based on real records: Evidence from Wuhan-Guangzhou HSR. *International Journal of Rail Transportation.* 2017;5(3):170-189. DOI: 10.1080/23248378.2017.1307144

[3] Lessan J, et al. Stochastic model of train running time and arrival delay: A case study of Wuhan–Guangzhou high-speed rail. *Transportation Research Record*. 2018;863081277. DOI: 10.1177/0361198118780830

[4] Huang P, Peng Q, Wen C, Yang Y. Random forest prediction model for Wuhan-Guangzhou HSR primary train delays recovery. *Journal of China Railway Society*. 2018;40(7):1–9. DOI: 10.3969/j.issn.1001-8360.2018.07.001

[5] Steven H, Fabrizio C, Otto A. A closed form railway line delay propagation model. *Transportation Research Part C: Emerging Technologies*. 2019;102:189-209. DOI: 10.1016/j.trc.2019.02.022

[6] Barbour W, Mori J, Kuppa S, Daniel B. Prediction of arrival times of freight traffic on US railroads using support vector regression. *Transportation Research Part C: Emerging Technologies*. 2018;93:211-227. DOI: 10.1016/j.trc.2018.05.019

[7] Marković N, Milinković S, Tikhonov K, Schonfeld P. Analyzing passenger train arrival delays with support vector regression. *Transportation Research Part C: Emerging Technologies*. 2015;56:251-262. DOI: 10.1016/j.trc.2015.04.004

[8] Corman F, Kecman P. Stochastic prediction of train delays in real-time using Bayesian networks. *Transportation Research Part C: Emerging Technologies*. 2015;95:599-615. DOI: 10.1016/j.trc.2018.08.003

[9] Nair R, et al. An ensemble prediction model for train delays. *Transportation Research Part C: Emerging Technologies*. 2019;104:196-209. DOI: 10.1016/j.trc.2019.04.026

[10] Huang P, et al. Modeling train operation as sequences: A study of delay prediction with operation and weather data. *Transportation Research Part E: Logistics and Transportation Review*. 2020;141(102022). DOI: 10.1016/j.tre.2020.102022

[11] Ilalokhoin O, Pant R, Hall J. A model and methodology for resilience assessment of interdependent rail networks – Case study of Great Britain's rail network. *Reliability Engineering and System Safety*. 2023;229(108895). DOI: 10.1016/j.ress.2022.108895

[12] Preston J, et al. Impact of delays on passenger train services: Evidence from Great Britain. *Transportation Research Record*. 2009;2117(1):14-23. DOI: 10.3141/2117-03

[13] Zhuang H, et al. Cause based primary delay distribution models of high-speed trains on account of operation records. *Journal of China Railway Society*. 2017;39(9):25-31. DOI: 10.3969/j.issn.1001-8360.2017.09.004

[14] Wang J, Peng Y, Lu J, Jiang Y. Analysis of risk factors affecting delay of high-speed railway in China based on Bayesian network modeling. *Journal of Transportation Safety & Security*. 2021;3:1-22. DOI: 10.1080/19439962.2021.1890290

[15] Goverde R. A delay propagation algorithm for large-scale railway traffic networks. *Transportation Research Part C: Emerging Technologies*. 2010;18(3):269-287. DOI: 10.1016/j.trc.2010.01.002

[16] Yang B. Study on countermeasures of increasing punctuality of high-speed trains. *Railway Transport and Economy*. 2012;34(12):53-57.

[17] Nabian M, Alemazkoor N, Meidani H. Predicting near-term train schedule performance and delay using Bi-level random forests. *Transportation Research Record*. 2019;2673(5):564-573. DOI: 10.1177/0361198119840339

[18] Mirabadi A, Sharifian S. Application of association rules in Iranian railways (RAI) accident data analysis. *Safety Science*. 2010;48:1427-1435. DOI: 10.1016/j.ssci.2010.06.006

[19] Montella A. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. *Accident Analysis & Prevention*. 2011;43:1451-1463. DOI: 10.1016/j.aap.2011.02.023

[20] Cheng C, Lin C, Leu S. Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction Industry. *Safety Science*. 2010;48:436-444. DOI: 10.1016/j.ssci.2009.12.005

[21] Verma A, Khan S, Maiti J, Krishna O. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. *Safety Science*. 2014;70:89-98. DOI: 10.1016/j.ssci.2014.05.007

[22] Han J, Kamber M, Pei J. *Data mining: Concepts and techniques*. Beijing, China: Machinery Industry Press; 2012.

[23] Aggarwal C. *Data mining: The textbook*. Beijing, China: Machinery Industry Press; 2021.

[24] Wang J, Wang Y, Peng Y, Lu J. Examining partial proportional odds model in analyzing severity of high-speed railway accident. *Smart and Resilient Transportation*. 2020;3(1):1-13. DOI: 10.1108/SRT-10-2020-0022