Research Article

Modeling the Influence of Disturbances in High-Speed Railway Systems

Ping Huang,1,2,3 Chao Wen,1,2,3 Qiyuan Peng,1,2 Chaozhe Jiang,1,2 Yuxiang Yang,1,2 and Zhuan Fu4

1 School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, China
2 National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 610031, China
3 Railway Research Center, University of Waterloo, Waterloo, Canada N2L 3G1
4Haikou Train Operation Depot, Hainan Railway Co., Ltd., Haikou 570100, China

Correspondence should be addressed to Chao Wen; wenchao@swjtu.cn and Chaozhe Jiang; jiangchaozhe@163.com

Received 7 November 2018; Accepted 28 January 2019; Published 19 March 2019

Academic Editor: Antonino Vitetta

Copyright © 2019 Ping Huang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Accurately forecasting the influence of disturbances in High-Speed Railways (HSR) has great significance for improving real-time train dispatching and operation management. In this paper, we show how to use historical train operation records to estimate the influence of high-speed train disturbances (HSTD), including the number of affected trains (NAT) and total delayed time (TDT), considering the timetable and disturbance characteristics. We first extracted data about the disturbances and their affected train groups from historical train operation records of Wuhan-Guangzhou (W-G) HSR in China. Then, in order to recognize the concatenations and differences of disturbances, we used a K-Means clustering algorithm to classify them into four categories. Next, parametric and nonparametric density estimation approaches were applied to fit the distributions of NAT and TDT of each clustered category, and the goodness-of-fit testing results showed that Log-normal and Gamma distribution probability densities are the best functions to approximate the distribution of NAT and TDT of different disturbance clusters. Specifically, the validation results show that the proposed models accurately revealed the characteristics of HSTD and that these models can be used in real-time dispatch to predict the NAT and TDT, once the basic features of disturbances are known.

1. Introduction

An operating train may encounter various unexpected disturbances such as bad weather, power outage, facility failures, and so on [1], which can lead to considerable losses for both railway managers and travelers. For example, in the Dutch railway network, statistics show that there are approximately 22 infrastructure-related delays per day, lasting on average for 1.7 hours [2]. According to the statistics of the China Railway Corporation, the average departure punctuality in origin stations for China’s 23,000km high-speed rail (HSR) network was as high as 98.8% in 2016. However, due to various disturbances during the trip, the average punctuality at final destination stations was less than 90%, even though delays smaller than five minutes are considered punctual [3]. When disturbances occur, dispatchers need to anticipate the potential influences of a specific delay. They need to estimate the number of affected trains (NAT) and total delayed time (TDT) before rescheduling the timetable. Modeling the high-speed train disturbances (HSTD) will be helpful and of great significance, although it is extremely challenging due to the following two aspects:

(1) Various influencing factors. The influence of railway disturbances is related to various factors, for example, timetable structure, facility conditions, and experience and preference of dispatchers which makes it difficult for these factors to be interpreted through functional relationships.

(2) Complex train interactions. Due to resources occupation conflicts and the continuity of train operation,
trains are interactive, which makes mathematical model incapable of modeling train interaction.

In practice, some skilled dispatchers usually predict HSTD empirically, which leads to differences in dispatching even for the same dispatcher when working in different situations. Data-driven approaches have recently gained more attention due to their better understanding of train delay concatenation and the fact they are more supportive of robust timetables and real-time dispatching [4]. In addition to the availability of train operation records, advanced data-mining techniques enable us to address these problems from a data-analysis perspective. Train operation records are therefore assumed to be the interactive consequences of all influencing factors. Therefore, mining train operation records provides us with a brand-new way of examining train interactions arising from heterogeneous factors.

To bridge the gap between the empirical and mathematical models, this paper aims to establish data-driven models of the NAT and TDT caused by different types of disturbances using the train operation data of Wuhan-Guangzhou (W-G) HSR in China. To this end, we used a K-Means algorithm to categorize the extracted disturbances according to their influencing factors. Next, we applied five widely used probability density models and two kernel functions to fit the distributions of NAT and TDT. We then selected the best models for each cluster based on goodness-of-fit testing. Finally, the test data from the operation records from nine months were used to validate the generalization of the fitted models, which showed that these models could be applied to estimate the NAT and TDT of future disturbances.

2. Literature Review

Since the 1970s, there has been active research on disturbance management in train dispatching [5]. Most recently, the topic of the INFORMS 2018 Railroad Problem Solving Competition was “Predicting Near-Term Train Schedule Performance and Delay” using operational records. A large number of methods and algorithms have been proposed to improve rail operations, but due to unavailable or insufficient historical data, research was mainly based on simulation and mathematical delay propagation models.

Many approaches have been uncovered and proposed to manage railway disturbances. Exogenous factors, such as natural disasters, and bad weather conditions, and endogenous factors, such as operation interference resulting from equipment failure, man-made faults, railway construction, temporary speed limitations, defective braking systems, signal and interlocking failures, and excessive passenger demand can contribute, alone or together, to the primary delay [6–8]. Also, if the running and dwell times increase due to unexpected disturbances, it can result in knock-on delays and delays for other trains [9]. Serious disruptions such as switch or signal failures, if not managed effectively, can result in queueing of trains creating a chain of delayed trains. The experience from the Taiwan High-speed Rail shows that shortening the maintenance cycle can effectively alleviate the problem of train delay caused by signal failures [10]. Some studies have made contributions on statistical models of delay and the respective fitness models. The Weibull, Gamma, and Log-normal distributions have been adopted in several studies [11, 12]. It was shown that the distributional form of primary delays, and the affected number of trains could be well-approximated by classical methods such as Log-normal distribution and linear regression models [13]. A q-exponential function is used to demonstrate the distribution of train delays on the British railway network [14]. Using spatial and temporal resolution transport data from the UK road and rail networks and the intense storms of 28 June 2012 as a case study, a novel exploration of the impacts of an extreme event has been carried out in [7]. Given the HSR operation data, the maximum likelihood estimation method was used to determine the probability distribution of the different disturbances factors and the distributions of affected trains; however, the models of primary delay consequences have not been established in detail [15]. Probabilistic distribution functions of both train arrival and departure delays at the individual station were derived in general based on the data from Beijing-Shanghai HSR [16].

Data-driven research studies proposed for delay/disruption management mainly focused on using regression or distribution approaches to fit delay data. Van der Meer et al. mined data from peak hours, including rolling stock, and weather data and developed a predictive model involving the mining of track occupation data for delay estimations [9]. A data-mining approach was used for analyzing rail transport delay chains, with data from passenger train traffic on the Finnish rail network, but the data from the train running process was limited to one month [4]. Murali et al. reported a delay regression-based estimation technique that models delay as a function of train mix and network topology [17]. A statistical analysis of train delays in the Eindhoven Station in the Netherlands was used to explain systematic delay propagation based on the use of a robust linear regression model to uncover the correlations between arrival delays [18]. Recently, Kecman and Goverde developed separate predictive models for the estimation of running and dwell times by collecting data on the respective process types from a training set [19]. Javad et al. examined different distribution models for running times of individual sections in an HSR system and showed that the log-logistic probability density function is the best distributional form to approximate the empirical distribution of running times on the specified line [3]. A hybrid Bayesian network model is also established to predict arrival and departure delays for Wuhan-Guangzhou HSR [20].

A review of the literature reveals that only a few studies focus on the modeling of the NAT and TDT of disturbances, especially in HSR operations. It is crucial for HSR operating companies to predict and reduce disruptions in train operations and to operate as closely as possible to published timetables. It is therefore important that they identify the severity of displacements or interruptions to train services. This can help dispatchers reduce delay propagation and the possibility of aggravation through effective designing of timetables or real-time dispatching decisions. This study aims
Table 1: Train operation data format in database.

<table>
<thead>
<tr>
<th>Train</th>
<th>Station</th>
<th>Date</th>
<th>Actual arrival</th>
<th>Actual departure</th>
<th>Scheduled arrival</th>
<th>Scheduled departure</th>
<th>Occupied track</th>
</tr>
</thead>
<tbody>
<tr>
<td>G280</td>
<td>GZN</td>
<td>2015/3/24</td>
<td>7:02</td>
<td>7:02</td>
<td>7:01</td>
<td>7:01</td>
<td>II</td>
</tr>
<tr>
<td>G548</td>
<td>SG</td>
<td>2015/8/27</td>
<td>17:26</td>
<td>17:30</td>
<td>17:26</td>
<td>17:29</td>
<td>4</td>
</tr>
</tbody>
</table>

* The passage tracks at the station are labeled with Roman characters, while the dwelling tracks are labeled with numbers.

3. Preliminaries

3.1. Data Description. The W-G HSR, which has a length of 1,096 km, is one of the busiest passenger railway lines in the country and joins with the Guangzhou-Shenzhen, Hengyang-Liuzhou, and Shanghai-Kunming HSRS at the GZS, HYE, and CSS stations, respectively. Trains operating on this line are all equipped with the Chinese Train Control System (CTCS), which allows a maximum speed of 350 km/h, and an Automatic Train Supervision system, which records the movements of all trains. The W-G HSR line, where only high-speed trains run on, is totally separated from conventional lines, but it connects with Shanghai-Kunming HSR line at CSS station, Hengyang-Liuzhou HSR line at HYE station, and Guangzhou-Shenzhen and Guiyang-Guangzhou HSR lines at GZN station. At these stations, train termination, turn-over, and cross-line are allowed. We studied the movement of trains in the northbound direction on this line, that is, from GZS to CSS, as shown in Figure 1.

The data collected include 57,796 trains in the GZS-HYE section and 64,547 HSR trains in the HYE-CSS section. The data contain operational records covering the period from March 24, 2015 to November 10, 2016, comprising scheduled/actual arrival/departure records for each train at each station, train numbers, dates, and information on occupied tracks. All the data with a time format was recorded in full minutes due to the accuracy of the system, as shown in Table 1.

In order to better understand the distributional pattern of the influence of disturbances, we first analyzed the train delay regularity by visualizing the arrival and departure delay distribution at HSW station. The histograms in Figure 2 clearly show that both arrival and departure delays follow a heavy-tailed distribution, from which we can infer that the influence of disturbance (NAT and TDT), like other problems in complex systems, has a right-skewed and heavy-tailed distribution.

3.2. Problem Description. In both practice and research, train delays and disturbances are always classified according to their causes. However, this method seems to have a drawback, as some disturbances with different causes sometimes have the same influencing mechanism on train operation. For example, track failure in sections and power supply fault in sections are different causes, but they have the same effect on train operation, because in both conditions, trains have to wait for the availability of the section. In other words, from the perspective of railway management, it is significant to classify disturbances according to their impacts on train operation. Besides, other cases like signal fault and turnout fault, speed limitation for bad weather and speed limitation for construction, they both have the same effect on train operation. In this research, we intend to classify the disturbances focusing on some key factors that influence their effects on train operation.

4. Clustering Model for Disturbances

4.1. Influencing Factors. In order to meet passenger demand, train services tend to vary across different periods and
segments even on the same HSR line. Train interval is a key factor in delay propagation, as a disturbance tends to cause more severe effects in smaller interval periods and segments, and smaller effects in larger interval periods and segments. Figures 3(a) and 3(b) reveal the NAT and TDT of disturbances at GZN and HSW. It clearly shows that the NAT and TDT of W-G HSR differ significantly across time periods and segments. In addition, the disturbance length will directly influence their consequences; the longer the disturbance, the more delayed trains and total delays it will cause. Based on the analysis, the following features for clustering models and indexes to measure the influence of disturbances were ascertained:

(i) Train interval ($I$): the average scheduled train interval when a disturbance occurs (minute);

(ii) Occurrence time ($T$): the time when a disturbance starts (in the railway operation, it can be indicated by the scheduled arrival time of the first delayed train in its affected trains group);

(iii) Disturbances length ($L$): the time span from starting time to ending time of the disturbance (it can be indicated by the delay times of the first train in the affected train group (minute)).

(iv) NAT and TDT indexes are used to measure their influence.

Figure 4, which shows a real disturbance in YDW-SG section on W-G HSR line, demonstrates the calculating methods of the selected features. If seven trains were delayed due to this disturbance (the first seven trains were delayed but the eighth train was on schedule when they arrived at
Table 2: Examples of disturbances and their influences for modeling.

<table>
<thead>
<tr>
<th>Train</th>
<th>Date</th>
<th>Station</th>
<th>Occurrence time</th>
<th>Length (min)</th>
<th>Train interval (min)</th>
<th>NAT</th>
<th>TDT (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1406</td>
<td>2015/9/2</td>
<td>LYW</td>
<td>19:55</td>
<td>9</td>
<td>6.5</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>G6020</td>
<td>2015/9/6</td>
<td>ZZW</td>
<td>17:16</td>
<td>17</td>
<td>7.5</td>
<td>8</td>
<td>93</td>
</tr>
<tr>
<td>G102</td>
<td>2016/6/24</td>
<td>HSW</td>
<td>9:07</td>
<td>21</td>
<td>10.0</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>G1032</td>
<td>2016/9/14</td>
<td>GZN</td>
<td>13:26</td>
<td>46</td>
<td>7.9</td>
<td>17</td>
<td>183</td>
</tr>
<tr>
<td>G6486</td>
<td>2016/10/29</td>
<td>LCE</td>
<td>10:33</td>
<td>30</td>
<td>8.1</td>
<td>31</td>
<td>173</td>
</tr>
</tbody>
</table>

Input: dataset \( D = \{x_1, x_2, \ldots, x_N\} \), and initial centers \( K \); 
Output: categories of every sample \( C = \{C_1, C_2, \ldots, C_K\} \)
Initialization: Randomly classify samples as \( K \) categories, and calculate the initial mean vector \( (\mu_k) \); 
Repeating the following step until stopping condition is satisfied: 
(i) Calculating the sign \( (\lambda_i) \) of sample \( x_i \), \( \lambda_i = \arg(\min |x_i - \mu_k|) \) 
(ii) Classifying \( x_i \) into its nearest cluster, \( C_{\lambda_i} = C_{\lambda_i} \cup \{x_i\} \) 
(iii) Updating the mean vector, \( \bar{\mu}_k = (1/|C_k|) \sum_{x_i \in C_k} x_i \) 
(iv) Stopping condition: mean vectors do not change, thus: \( \mu_k = \bar{\mu}_k \)

Figure 4: A real disturbance happened on W-G HSR line shown in time-space diagram (horizontal axis is time, and vertical axis is space).

SG station), we thus define these seven trains as affected train group. In this case, NAT value is thus seven, TDT value is the sum of arrival delays of these seven trains at SG station, and the interval \( I \) is the average interval of these seven trains at SG station. The disturbance length \( L \) which is defined as the difference of its occurrence time and the time when this section is available in this research was calculated by the delay times of the first train in the affected train group.

Based on the indexes, data on 6006 disturbances and their consequences on train operations were extracted from the raw data; five cases are shown in Table 2. In order to validate the proposed model, the extracted data were split into a training dataset, which included 3154 disturbances in the preceding 12 months, and a validation dataset, which contained 2852 disturbances over the following 9 months. We extracted only those disturbances with a disturbance longer than 4 minutes according to the standard set by the Chinese Railway Company. Shorter disturbances, which can be assimilated by the time supplements distributed in timetables, tend not to cause delay propagation and are therefore eliminated from the dataset.

4.2. K-Means Clustering. A K-Means cluster is a typical and popular algorithm that has strong robustness on high-dimensional and multicollinear datasets in unsupervised learning. For the given dataset \( D = \{x_1, x_2, \ldots, x_N\} \), assuming that the clustering centers \( C = \{C_1, C_2, \ldots, C_K\} \) are initialized, the object of K-Means is to minimize the mean squared error (MSE):

\[
\min_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - u_k\|^2
\]

where \( u_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \) is the mean vector of \( C_k \).

Equation (1) indicates the nearness between samples in a cluster and their mean vector.

The core principle of the K-Means cluster is to choose \( K \) points in space as centers and assign the samples to their nearest cluster. By iteratively updating their centers, the objective is to minimize until the stopping condition is satisfied. The K-Means clustering algorithm can be concluded as shown in Algorithm 1 and more details are shown in literature [21].

4.3. Model Performance. The performance of clustering models is commonly evaluated from two perspectives: (1) the tightness of the samples in the cluster and (2) the distances between clusters. The widely used evaluation indexes are distance and covariance [22, 23], both of which require smaller values among samples in the same cluster and larger values among samples in different clusters. To systematically
evaluate the clustering models, we simultaneously chose distance- and covariance-based indexes, namely, the Silhouette Coefficient (SC) and Calinski-Harabasz Score (CHS) as shown in (2) and (3), respectively.

$$SC(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$ (2)

where \(a(i)\) is the average distance between \(x_i\) and other samples in a cluster; \(b(i)\) is the average distance between \(x_i\) and samples in other clusters; notifying: the range of SC is \([-1, 1]\).

$$CHS(K) = \frac{tr(B_K)}{(K - 1)} / \frac{tr(W_K)}{(m - K)}$$ (3)

where \(m\) is the number of samples, \(K\) is the number of clusters, and \(B_K\) is the covariance matrix among samples in the same cluster, while \(W_K\) is the covariance matrix among samples in different clusters, and \(tr(g)\) is the trace of the matrix. According to (2) and (3), better clustering results require larger SC and CHS.

To obtain reasonable clustering results, we also applied other clustering algorithms including the BIRTH, Spectral, and Agglomerative clustering models and investigated the number of clusters (K) of each model from two to twenty in order to categorize the disturbances. After standardization of the input data, clustering was performed, and the clustering result of each model is shown in Figure 5. We can first choose K-Means model as the best clustering model from the candidates, as it almost had the best performances on all different number of clusters on both SC and CHS metrics. An exception was when K equals two, which resulted in the highest SC index but the lowest CHS index of BIRCH model. Considering that clustering models should be evaluated from both distances- and covariance-based indexes to obtain systematical results, we thus chose K-Means model, which has both high SC and CHS metrics, as the clustering model for railway disturbances. Then, when we chose the K value, we found that three, four, and five clusters were all possible for K-Means algorithm. We, therefore, standardized the SC and CHS values of K-Means model between zero and one, and finally chose four as the best number of clusters for it resulted in the highest sum of standardized SC and CHS. Therefore, a K-Means algorithm that has four clusters was selected as the clustering model for HSR disturbances.

Finally, HSR disturbances were classified into four categories using the K-Means clustering algorithm, as shown in Figure 6. According to the distribution of each cluster, we can define them as follows:

(i) Cluster A: disturbances occurred between 7:00 am and 23:30 pm; train intervals range from 3 to 13 minutes; lengths range from 5 to 14 minutes.

(ii) Cluster B: disturbances occurred between 7:30 am and 23:30 pm; train intervals range from 5 to 25 minutes; lengths range from 14 to 31 minutes.

(iii) Cluster C: disturbances occurred between 7:30 am and 23:30 pm; train intervals range from 4 to 30 minutes; lengths are longer than 31 minutes.

(iv) Cluster D: disturbances occurred between 10:30 am and 23:30 pm; train intervals range from 13 to 30 minutes; lengths range from 5 to 22 minutes.

The statistics of NAT and TDT of each cluster are shown in Table 3.

### 5. Estimating Models of NAT and TDT

#### 5.1. Candidate Models

In order to reveal the rules of disturbance influences, we first investigated the histogram
of NAT and TDT, as shown in Figures 7 and 8, which indicates that both NAT and TDT appear to have right-skewed distributions. Since the locations and shapes of the histograms are very different, we fitted the data using five common right-skewed distribution models and two kernel functions as the candidate models. The distribution models including the Log-normal, Weibull, Gamma, Exponential, and Logistic and kernel functions, including Gaussian and Epanechnikov kernels were employed to fit the data from parametric and nonparametric perspectives. The probability distribution models and kernel functions of these models are as follows [24].

(i) Log-Normal Distribution

\[
f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}
\]  

where \( \sigma \) is the shape parameter and \( \mu \) is the location parameter.

(ii) Weibull Distribution

\[
f(x; \lambda, k) = \begin{cases} \frac{k}{\lambda} \left( \frac{x}{\lambda} \right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}
\]  

where \( \lambda > 0 \) is the scale parameter and \( k > 0 \) is the shape parameter.

(iii) Gamma Distribution

\[
f(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}
\]

\[
\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt
\]  

where \( \alpha \) is the shape parameter and \( \beta \) is the scale parameter.

(iv) Exponential Distribution

\[
f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x \leq 0 \end{cases}
\]  

where \( \lambda \) is the shape parameter.

(v) Logistic Distribution

\[
f(x; u, s) = \frac{1}{s \left(e^{(x-u)/2s} + e^{-(x-u)/2s}\right)^2}
\]

Where \( u \) is the location parameter and \( s \) is the scale parameter.

(vi) Gaussian Kernel

\[
k(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \mu^2}
\]
The parameters were estimated using maximum likelihood and are shown in Table 4, and the fitting results of the distribution models are shown in Figures 7 and 8. These figures clearly show that all the candidate probability density models mimic the shape of the nonparametric estimation using Gaussian and Epanechnikov kernels, which enables us to choose the best model from the parametric candidates as the estimating model of disturbance influences.

(vii) Epanechnikov Kernel

\[ k(\mu) = \begin{cases} 
\frac{3}{4} (1 - u^2) & \text{if } |u| \leq 1 \\
0 & \text{else} 
\end{cases} \]  

The parameters were estimated using maximum likelihood and are shown in Table 4, and the fitting results of the distribution models are shown in Figures 7 and 8. These figures clearly show that all the candidate probability density models mimic the shape of the nonparametric estimation using Gaussian and Epanechnikov kernels, which enables us to choose the best model from the parametric candidates as the estimating model of disturbance influences.

Table 4: Fitted parameters of NAT and TDT.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Log-normal (\mu)</th>
<th>Log-normal (\sigma)</th>
<th>Weibull (k)</th>
<th>Weibull (\lambda)</th>
<th>Gamma (\alpha)</th>
<th>Gamma (\beta)</th>
<th>Exponential (\lambda)</th>
<th>Exponential (u)</th>
<th>Logistic (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1.754</td>
<td>0.884</td>
<td>1.176</td>
<td>9.003</td>
<td>1.455</td>
<td>0.172</td>
<td>0.118</td>
<td>7.016</td>
<td>3.912</td>
</tr>
<tr>
<td>B</td>
<td>2.175</td>
<td>0.932</td>
<td>1.308</td>
<td>13.616</td>
<td>1.557</td>
<td>0.124</td>
<td>0.079</td>
<td>11.229</td>
<td>5.380</td>
</tr>
<tr>
<td>C</td>
<td>2.155</td>
<td>0.874</td>
<td>1.406</td>
<td>12.943</td>
<td>1.760</td>
<td>0.149</td>
<td>0.085</td>
<td>10.771</td>
<td>4.650</td>
</tr>
<tr>
<td>D</td>
<td>1.506</td>
<td>0.781</td>
<td>1.288</td>
<td>6.688</td>
<td>1.773</td>
<td>0.289</td>
<td>0.163</td>
<td>5.120</td>
<td>2.557</td>
</tr>
<tr>
<td>TDT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>3.233</td>
<td>0.861</td>
<td>1.146</td>
<td>38.660</td>
<td>1.439</td>
<td>0.039</td>
<td>0.027</td>
<td>29.683</td>
<td>17.195</td>
</tr>
<tr>
<td>B</td>
<td>4.386</td>
<td>0.728</td>
<td>1.616</td>
<td>113.677</td>
<td>2.273</td>
<td>0.022</td>
<td>0.010</td>
<td>95.280</td>
<td>37.595</td>
</tr>
<tr>
<td>C</td>
<td>4.769</td>
<td>0.567</td>
<td>2.086</td>
<td>154.544</td>
<td>3.560</td>
<td>0.026</td>
<td>0.007</td>
<td>131.530</td>
<td>41.077</td>
</tr>
<tr>
<td>D</td>
<td>3.188</td>
<td>0.764</td>
<td>1.334</td>
<td>35.677</td>
<td>1.859</td>
<td>0.057</td>
<td>0.031</td>
<td>27.779</td>
<td>13.421</td>
</tr>
</tbody>
</table>
5.2. Goodness-of-Fit Test. In this section, we evaluated the goodness-of-fit of the distribution models using the Kolmogorov-Smirnov (K-S) method [25] and selected the optimal distribution model for each category.

K-S testing is proposed to test whether a group of data follows a theoretical distribution model; its null hypothesis is that the dataset follows a theoretical distribution. Its testing statistics are the largest difference between the cumulative distribution function of the data and the theoretical distribution, as shown in (12). Because the number of trains is the integer, and the historical train operation data were recorded in the minimum unit of one minute, we inserted some random numbers that follow uniform distribution in order to meet the continuity requirement of the K-S method.

\[
D = \max \left| F'(x) - F(x) \right| \tag{12}
\]

where \( F'(x) \) is the cumulative distribution function of the samples, which are the NAT and TDT; \( F(x) \) is the cumulative distribution function of the theoretical distribution models, which are the five alternative distribution models. We chose the significance level \( \alpha = 0.05 \); when the sample size is large enough (over 50), the critical value of \( D \) should be

\[
D_{0.05} = \frac{1.36}{\sqrt{n}} \tag{13}
\]

where \( n \) is the sample size.

According to the rules of the K-S test, if \( D < D_{0.05} \), the null hypothesis is accepted and the samples are considered as the following theoretical distribution. The smaller \( D \) is, the closer the sample is to the theoretical distribution. The K-S testing results of all the models are shown in Table 5.

Finally, we chose the model that passes the K-S test and has the smallest \( D \) as the distribution model of each disturbance cluster. The parameters of distribution models for NAT and TDT of each cluster are shown in Table 6.
Table 5: Goodness-of-fit testing of NAT and TDT using K-S method.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Log-normal</th>
<th>Weibull</th>
<th>Gamma</th>
<th>Exponential</th>
<th>Logistic</th>
<th>$D_{0.05}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.025*</td>
<td>Yes</td>
<td>0.055</td>
<td>No</td>
<td>0.037</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>0.077</td>
<td>No</td>
<td>0.046</td>
<td>Yes</td>
<td>0.044*</td>
<td>Yes</td>
</tr>
<tr>
<td>C</td>
<td>0.088</td>
<td>No</td>
<td>0.045</td>
<td>Yes</td>
<td>0.038*</td>
<td>Yes</td>
</tr>
<tr>
<td>D</td>
<td>0.042*</td>
<td>Yes</td>
<td>0.066</td>
<td>Yes</td>
<td>0.043</td>
<td>Yes</td>
</tr>
<tr>
<td>TDT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.026*</td>
<td>Yes</td>
<td>0.082</td>
<td>No</td>
<td>0.058</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>0.057</td>
<td>Yes</td>
<td>0.038</td>
<td>Yes</td>
<td>0.037*</td>
<td>Yes</td>
</tr>
<tr>
<td>C</td>
<td>0.059</td>
<td>Yes</td>
<td>0.060</td>
<td>Yes</td>
<td>0.049*</td>
<td>Yes</td>
</tr>
<tr>
<td>D</td>
<td>0.037*</td>
<td>Yes</td>
<td>0.064</td>
<td>Yes</td>
<td>0.041</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Superscript ** indicates the best models.

Table 6: Parameters of the best distribution models.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Model</th>
<th>Parameters</th>
<th>Model</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAT</td>
<td>Log-normal ($\mu, \sigma$)</td>
<td>(1.754, 0.884)</td>
<td>Log-normal ($\mu, \sigma$)</td>
<td>(3.233, 0.861)</td>
</tr>
<tr>
<td></td>
<td>Gamma ($\alpha, \beta$)</td>
<td>(1.557, 0.124)</td>
<td>Gamma ($\alpha, \beta$)</td>
<td>(2.273, 0.022)</td>
</tr>
<tr>
<td></td>
<td>Gamma ($\alpha, \beta$)</td>
<td>(1.760, 0.149)</td>
<td>Gamma ($\alpha, \beta$)</td>
<td>(3.560, 0.026)</td>
</tr>
<tr>
<td>TDT</td>
<td>Log-normal ($\mu, \sigma$)</td>
<td>(1.506, 0.781)</td>
<td>Log-normal ($\mu, \sigma$)</td>
<td>(3.188, 0.764)</td>
</tr>
</tbody>
</table>

Table 7: Statistics and the testing result of NAT and TDT using K-S method.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Samples size</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
<th>SD</th>
<th>$D$ <em>value</em></th>
<th>Pass?</th>
<th>$D_{0.05}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1757</td>
<td>1</td>
<td>8.06</td>
<td>40</td>
<td>7.51</td>
<td>0.007</td>
<td>Yes</td>
<td>0.032</td>
</tr>
<tr>
<td>B</td>
<td>357</td>
<td>1</td>
<td>12.61</td>
<td>40</td>
<td>9.58</td>
<td>0.055</td>
<td>Yes</td>
<td>0.072</td>
</tr>
<tr>
<td>C</td>
<td>195</td>
<td>1</td>
<td>12.28</td>
<td>39</td>
<td>8.86</td>
<td>0.056</td>
<td>Yes</td>
<td>0.097</td>
</tr>
<tr>
<td>D</td>
<td>273</td>
<td>1</td>
<td>6.31</td>
<td>24</td>
<td>3.11</td>
<td>0.069</td>
<td>Yes</td>
<td>0.082</td>
</tr>
<tr>
<td>TDT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1757</td>
<td>5</td>
<td>35.13</td>
<td>238</td>
<td>34.73</td>
<td>0.023</td>
<td>Yes</td>
<td>0.032</td>
</tr>
<tr>
<td>B</td>
<td>357</td>
<td>15</td>
<td>99.18</td>
<td>294</td>
<td>64.13</td>
<td>0.027</td>
<td>Yes</td>
<td>0.072</td>
</tr>
<tr>
<td>C</td>
<td>195</td>
<td>32</td>
<td>135.90</td>
<td>299</td>
<td>66.85</td>
<td>0.046</td>
<td>Yes</td>
<td>0.097</td>
</tr>
<tr>
<td>D</td>
<td>273</td>
<td>5</td>
<td>32.61</td>
<td>162</td>
<td>26.29</td>
<td>0.067</td>
<td>Yes</td>
<td>0.082</td>
</tr>
</tbody>
</table>

5.3. Generalization Test. In order to investigate the generalization of the fitted distribution models, we used the disturbances in the following nine months of W-G HSR line to validate the models. We first fed the disturbances into the proposed K-Means clustering algorithm and obtained the clustering labels of each sample. The clusters that have the same labels as the training dataset are used to validate the fitted models. The fitting results are shown in Figures 9 and 10, and the descriptive statistics and their K-S testing results are shown in Table 7.

The testing results indicate that all the fitted models pass the goodness-of-fit and generalization testing. The fitted models can therefore accurately reveal the distributive disciplines of NAT and TDT and are of great significance to real-time train dispatching.

6. Conclusion

This paper presents a data-mining approach, which is composed of the K-Means clustering model and probability density fitting techniques to estimate NAT and TDT resulting from HSR operation disturbances, considering timetable and disturbance characteristics. The proposed clustering algorithm, which fully takes the disturbance characteristics and timetable structure into consideration, compensates the shortage of existing classification method of railway disturbance. In real-time operation, once a disturbance happens, its features are therefore known (including train interval, length and occurrence time of disturbance), which can be fed into the trained K-Means clustering algorithm, and its classifying label can be obtained. Then, according to the label,
Figure 9: Fitting results of NAT on test data.

Figure 10: Fitting results of TDT on testing data.
the specific distribution model can be chosen to calculate the probabilities of different outcomes. The probability models being fitted using real operation data can help dispatchers improve their decision-making qualities. With the probability distribution models, the dispatchers can obtain the real-time and future probabilities of NAT and TDT of any train operation disturbances. The influencing patterns of disturbances arising from the train operation data can be used to improve the rescheduling and adjusting abilities in disturbance situations and help dispatchers improve their decision-making qualities, by providing dispatchers with accurate estimation of NAT and TDT. Also, these probability models can improve the train operation and disturbance management in simulation systems, as they are more accurate than those hypothetical models that bring certain gaps between simulations and practice and usually over assume and ignore some situations and constraints of train operations.

The established distribution models are general models for W-G and X-S HSR lines. However, train services and infrastructure could both vary at different stations, leading to differences in distribution model parameters. Therefore, our future work will focus on establishing models for each station on the HSR network.

Data Availability

The data used to support the findings of this study are restricted by the China Railway Corporation. Data are available from Ping Huang, huangping129@my.swjtu.edu.cn, only for researchers who use them to review or better understand this paper.

Conflicts of Interest

No potential conflicts of interest were reported by the authors.

Acknowledgments

This work was supported by the National Nature Science Foundation of China [Grant no. 71871188]; the Science & Technology Department of Sichuan Province [Grant no. 2018JY0567]; and Doctoral Innovation Fund Program of Southwest Jiaotong University. We are grateful for the contributions made by our project partners.

References


