Statistical Modeling of the Distribution Characteristics of High-Speed Railway Disruptions

Ping Huang [a,b,c,1]

[a] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu Sichuan 610031, China;

[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu Sichuan 610031, China;

[c] Railway Research Centre, University of Waterloo, Waterloo N2L3G1, Canada

[1] E-mail: huangping129@my.swjtu.edu.cn, +86 18200298902

Abstract

Studies on the spatiotemporal distribution and duration characteristics of railway disruptions are very significant for the advanced prediction of disruption and development of real-time dispatch strategies. In this study, historical disruption records of some Chinese High-Speed Railways (HSRs) lines from 2014–2016 were used to investigate the distribution characteristics of railway disruptions. The spatiotemporal probability distribution of four railway lines were calculated and their hotspots (coordinates with high probabilities) and coldspots (coordinates with low probabilities) were revealed using heatmaps. Furthermore, all the disruptions were classified into seven clusters based on their causes, and statistical analysis was carried out on each cluster. In addition, three right-skewed distribution models, namely Log-normal, Weibull, and Gamma distributions, were used to fit the duration of each cluster to uncover its duration regularities. Finally, goodness-of-fit test was performed on the models using the Kolmogorov-Smirnov method, indicating that the duration of each classified disruption can be estimated using a Log-normal distribution function. The obtained spatiotemporal probabilities and duration time distribution models thus can be further applied into estimating the occurrence and duration of railway disruption in real-time dispatching to help dispatchers make advanced decisions.

Keywords

High-speed-railway; disruption; spatial-temporal distribution; duration; Log-normal distribution

1 Introduction

The disruptions encountered in railway systems are caused by humans, equipment, and the environment, which can lead to considerable losses for managers and travelers. For example, the statistics from a Dutch railway network show that infrastructure-related disruptions occur approximately 22 times per day and each disruption lasts for an average of 1.7 h (Jespersen-Groth et al, 2009). Furthermore, the Austrian Federal Railways suffer huge financial losses of more than EUR 100 million every year due to flooding (Kellermann, Schönberger and Thieken, 2016). Meanwhile, the average departure punctuality in China at various origin stations was 98.8% in 2016. However, the average punctuality at the final destination stations was less than 90% due to various disruptions during operation, although delays smaller than 5 min are considered punctual (Lessan et al, 2018). Hence, train dispatchers are faced with the challenge of reducing the influence of disruptions by developing effective strategies in advance. In other words, the dispatchers can make effective decisions during or before disruptions for efficient timetable re-scheduling if they

can predict when and where the disruptions would occur and how long the disruptions would last. Therefore, studies on the rules and distribution characteristics of railway disruptions are significant for the real-time dispatch of trains.

However, there are several challenges in the accurate prediction of the occurrence of train disruption and duration which are as follows: 1) the disruption is unexpected; and 2) the maintenance duration is highly dependent on the experience and skill of the maintenance staff. Functional models are not sufficient to explain the complex relationship between the disruptions and their potential influence factors. However, skilled dispatchers usually predict the disruption duration empirically, which tends to cause ineffective dispatching when disruptions happen. However, data-mining approaches have recently gained more attention because they can efficiently model train operations and can support robust timetables and real-time dispatching (Wallander and Mäkitalo, 2012). Historical disruption records are considered as interactive consequences of all potential influence factors such that the disruption rules can be determined from the historical performances rather than influence factors. Thus, advanced data-mining techniques, as well as big data, enable us to address these problems using data analysis.

This paper aims to discover the spatiotemporal distribution and duration characteristics of railway disruptions based on data obtained from Guangzhou Railway Group in China. Thus, the spatiotemporal probability distribution of disruptions on four railway lines (Wuhan-Guangzhou HSR line, Shanghai-Shenzhen HSR line, Guangzhou-Shenzhen HSR line, and Guangzhou-Shenzhen intercity line) were analyzed. The disruptions were then classified into seven categories based on their source, and statistical analyses were conducted on the duration of each category. Furthermore, three right-skewed distribution models were used to fit the duration of disruption of each category. The histograms indicated that the duration has a right-skewed and heavy-tailed distribution. Finally, Kolmogorov-Smirnov method was used to perform the goodness-of-fit test in order to select the optimal models for each category.

2 Literature review

Generally, railway disruptions can be caused by exogenous factors, such as natural disasters, and bad weather conditions and endogenous factors, such as operation interference resulted from equipment failure, man-made faults, railway construction, temporary speed limitations, defective braking systems, signal and interlocking failures, and excessive passenger demand (Olsson and Haugland, 2004; Hartrumpf et al, 2009; Higgins, Kozan and Ferreira 1995). Many methods and models have been suggested to manage these disruptions. Traditionally, train operation simulated systems such as LUKS (Janecek and Weymann, 2010), RailSys (Wiklund, 2003), and OpenTrack (Nash and Huerlimann, 2004) have been used by railway researchers and managers worldwide. However, the disruption or delay parameters in these systems mainly depend on hypothetical and theoretical models. (Corman, D'Ariano and Hansen 2014) examined the resisting disturbance abilities of normal traffic and robust timetables using a simulation method. (Huisman and Boucherie, 2001) established a delay propagation model considering the routes occupation relations to predict the knock-on delays, under the condition that train delays follow an Exponential distribution.

Data-driven approaches are also widely used in railway disruption/delay management. These approaches aim to discover the delay and disruption patterns from historical train operation data or disruption records. (Murali et al, 2010) introduced a delay regression-based estimation technique that models delay as a function of train mix and network

topology. (Kecman and Goverde, 2015) developed separate predictive models for the estimation of running and dwell times by collecting data on the respective process types from a training set. (Lessan et al, 2018) 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. It was shown that the distributional form of primary delays, and the affected number of trains could be well-approximated by Log-normal distribution and linear regression models (Wen et al, 2017). A *q*-exponential function is used to demonstrate the distribution of train delays on the British railway network (Takimoto, 2000). 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 (Hartrumpf et al, 2009). Regression trees were trained using Hong Kong subway incident data to estimate the affected delay trains in (Weng et al, 2015). However, the environment of HSR trains is more complex than subway systems. Copula Bayesian networks were developed to predict the duration of turnout faults (Zilko, Kurowicka and Goverde, 2016). A hybrid Bayesian network model is also established to predict arrival and departure delays for Wuhan-Guangzhou HSR (Lessan, Fu and Wen, 2018).

3 Data description

The data used in this study were obtained from the disruption records of Guangzhou Railway Group from 2014-2016, for Wuhan-Guangzhou, Shanghai-Shenzhen, and Guangzhou-Shenzhen HSR lines, as well as Guangzhou-Shenzhen intercity line, as shown in Figure 1. The trains have a maximum speed of 350 km/h when operated on Wuhan-Guangzhou and Guangzhou-Shenzhen HSR lines and 250 km/h when operated on Shanghai-Shenzhen HSR line. In addition, the trains have a maximum speed of 200 km/h when operated on Guangzhou-Shenzhen intercity line. Thus, 2,256 disruptions attributed to nine causes were recorded which are Automatic Train Protection (ATP) system faults, turnout faults, track faults, pantograph faults, rolling stock faults, catenary faults, signal system faults, foreign body invasions, and severe weather. Table 1 shows four cases of the disruptions in the database.

Table 1: Records of HSR disruptions

Line	Date	Train	Time	Duration(min)	Cause
Wuhan-Guangzhou HSR	2014.05.19	G275	19:10	19	Catenary faults
Wuhan-Guangzhou HSR	2014.05.20	G6313	14:30	63	Severe weather
Wuhan-Guangzhou HSR	2015.09.27	G1133	17:06	15	Severe weather
Shanghai-Shenzhen HSR	2015.10.24	G530	16:42	19	Pantograph faults

Figure 1: Sketch map of HSR lines in the jurisdiction of Guangzhou Railway Group.

4 Spatiotemporal probability distribution of disruptions

Railway disruptions are unexpected. However, they tend to appear as regularities that can be investigated from large-scale historical records due to the influence of external factors such as weather and climate, and internal factors such as the characteristics and coordination of equipment, and train interval. Figures 1–4 show the spatiotemporal probability distributions of HSR disruptions, where darker colors represent higher probabilities. Owing to the low probabilities and frequencies of disruptions, each HSR line was divided into several segments to improve the statistical effects. For example, Wuhan-Guangzhou HSR line which has 17 stations was divided into four segments from south-north, such as GZS-SG, SG-HYE, HYE-CSS, and CSS-WH. Figures 1–4 indicate that different segments have different probabilities in the time domain. The peak hours occurred between 12:00 and 20:00. However, GZS-SG segment has the highest probabilities for Wuhan-Guangzhou HSR line, while SZN-SW and CS-ZA segments have higher probabilities for Shanghai-Shenzhen HSR line. Similarly, GZS-HM has the highest probabilities for Guangzhou-Shenzhen HSR line, while GZ-DG and ZMT-SZ have higher probabilities for Guangzhou-Shenzhen intercity line. The spatiotemporal characteristics of disruptions indicate that the probabilities of the disruptions depend on the number of train operations in time domain. However, its influence factors are complex in space domain owing to weak regularities. The probabilities in the space domain tend to be influenced by the status of the equipment, skill

and experience of dispatchers, weather, and climate. However, these factors are different for different locations.

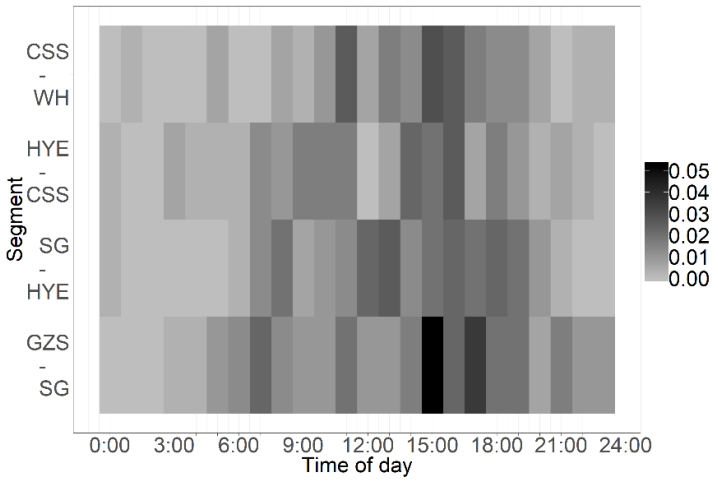

Figure 2: Spatial-temporal distribution of Wuhan-Guangzhou HSR disruptions.

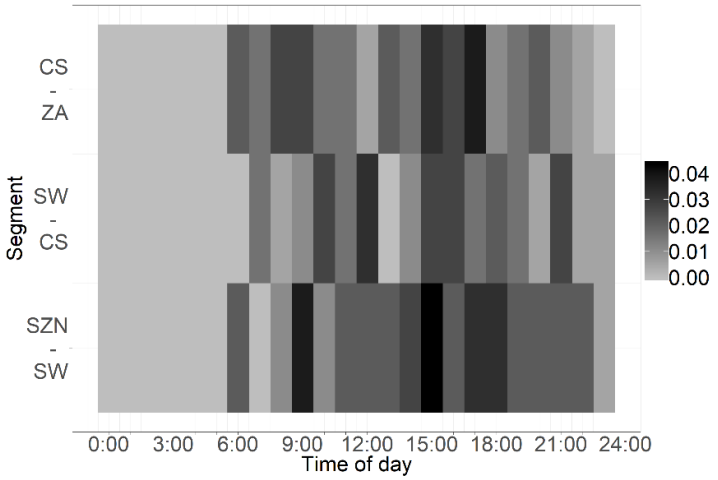

Figure 3: Spatial-temporal distribution of Shanghai-Shenzhen HSR disruptions.

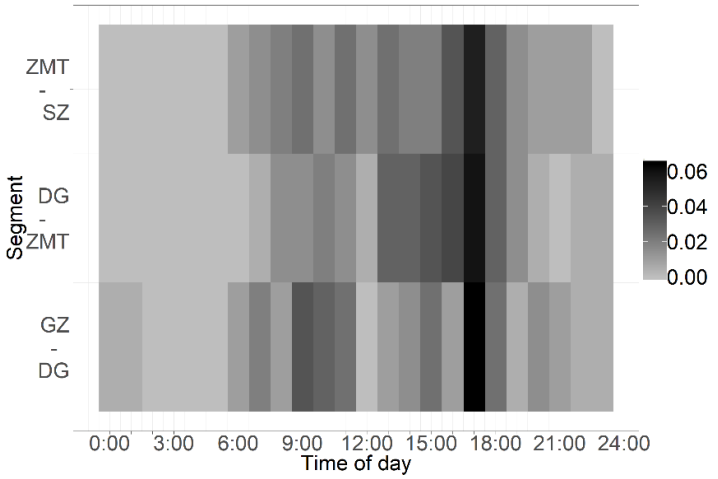

Figure 4: Spatial-temporal distribution of Guangzhou-Shenzhen Intercity Railway disruptions.

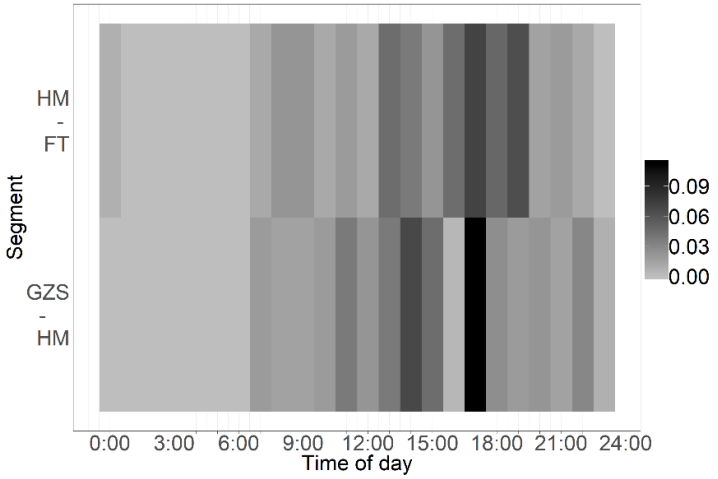

Figure 5: Spatial-temporal distribution of Guangzhou-Shenzhen HSR disruptions.

5 Investigation of disruption duration characteristics

The spatiotemporal distribution probabilities can help dispatchers predict the occurrence of disruptions. However, in practice, it is also necessary to know the duration of disruptions to better understand the characteristics of disruptions, and estimate their influences on train operation, as the durations of disruptions can have different influence on railway systems. Therefore, in this section, we examine the rules of disruption durations using statistical

method.

5.1 Statistics analyses

Based on the coordinated relationship between each equipment in HSR systems, pantograph faults and catenary faults can be regarded as a single category called power supply faults as they have the same effect on HSR systems. Likewise, track faults and turnout faults can be regarded as a single category called turnout-track faults. Thus, the disruptions were classified into seven clusters, namely ATP faults (ATPFs), rolling stock faults (RSFs), turnout-track faults (TTFs), power supply faults (PSFs), signal faults (SFs), severe weather (SW), and foreign body intrusions (FBIs). Statistical analyses were conducted to examine the differences in duration between each category, as shown in Table 2. The results show that the mean values of TTF and SW durations are higher than other values and are longer than 40 min, which indicates that these two categories have stronger influence on the HSR system. In addition, the variances of these two categories are larger than the other values, indicating that a larger uncertainty exists. Meanwhile, the mean and variance of ATPF duration have the least values, indicating that ATPF has the least influence on the HSR system. Its duration has a more centralized distribution.

Table 2: Statistics on duration time of disruptions with different causes(min).

Cause	Min	Mean	Max	Variance	Sample size
RSF	13	31.69	506	1148.22	472
ATPF	10	20.68	154	327.16	328
TTF	8	42.71	579	3185.21	149
PSF	9	33.21	295	1340.07	543
SW	4	41.97	286	2612.29	263
FBI	11	34.35	372	1336.38	289
SF	6	30.19	376	977.90	212
Total	4	27.39	577	1568.39	2256

5.2 Distributional models for disruption duration

The duration of disruption is the difference between its starting and ending time. Figure 6 shows a real disturbance in YDW-SG section on W-G HSR line. This figure defines the disruption length, which is from the time when the station/section is blocked to the time when the first train is allowed to pass. Longer durations can lead to stronger influence on the HSR system, causing more damage and significant losses to railway managers and travelers. Hence, the duration distribution models of the disruptions were investigated to discover the characteristics of disruptions so that dispatchers can predict and control the disruptions effectively. The database just recorded the disruptions whose length are longer than 4 minutes, because the delays longer than 4 minutes are labelled as delayed trains by the China Railway corporation. In addition, samples with durations longer than 120 min were regarded as outliers because they had extremely low frequencies. In Figure 7, the histograms show the duration distribution of each category and all samples, which indicate that both each cluster and all samples have a long-tailed and right-skewed distribution. To quantitatively examine their duration, three right-skewed probability models were selected to fit the data:

1) Log-normal distribution.
If the logarithm of a random variable follows a normal distribution, the random variable also follows a Log-normal distribution. The probability density of a Log-normal model is

$$f\left(x;\mu,\sigma\right)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(\ln x-\mu\right)^2}{2\sigma^2}\right) \tag{1}$$

where x is a random variable, σ is the standard deviation, and μ is mean.

2) Weibull distribution.

$$f\left(x;\lambda,k\right)=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left(-\left(x/\lambda\right)^k\right) & x\geq 0\\ 0 & x<0\end{cases} \tag{2}$$

where x is a random variable, λ>0 is the scale parameter, and k>0 is the shape parameter.

3) Gamma distribution.

$$f\left(x;\alpha,\beta\right)=\frac{1}{\beta^{\alpha}\Gamma\left(\alpha\right)}x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)\qquad x>0 \tag{3}$$

$$\Gamma\left(x\right)=\int_0^{\infty}t^{x-1}\exp\left(-t\right)dt \tag{4}$$

where x is a random variable, α is the shape parameter, and β is the scale parameter.

The models above were used to fit the duration of the disruptions as shown in Figure 7. Meanwhile, the fitted parameters of each category are shown in Table 3.

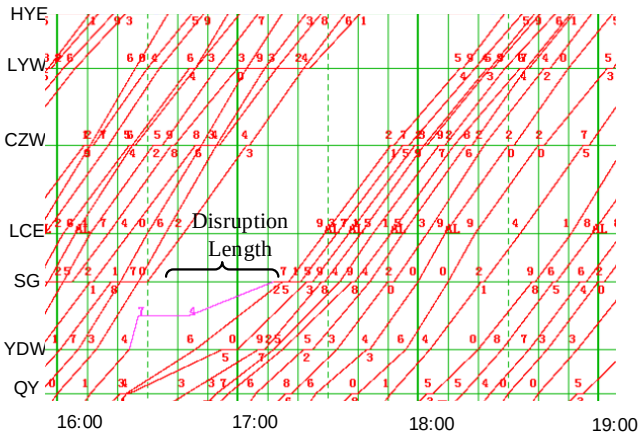

Figure 6: A real disturbance happened on W-G HSR line shown in time-space diagram (horizontal axis is time, and vertical axis is space).

Table 3: Fitted parameters of each category.

Cause	Log-normal		Weibull		Gamma	
	μ	σ	k	λ	α	β
RSF	3.100	0.772	1.416	32.418	1.954	0.066
ATPF	2.760	0.685	1.480	22.186	2.328	0.117
TTF	3.301	0.707	1.553	38.491	2.272	0.066
PSF	3.082	0.727	1.407	31.290	2.091	0.074
SW	3.091	0.780	1.304	32.774	1.761	0.058
FBI	3.138	0.733	1.434	33.230	2.067	0.069
SF	2.768	0.845	1.323	23.836	1.739	0.079
Total	3.029	0.766	1.376	30.211	1.933	0.070

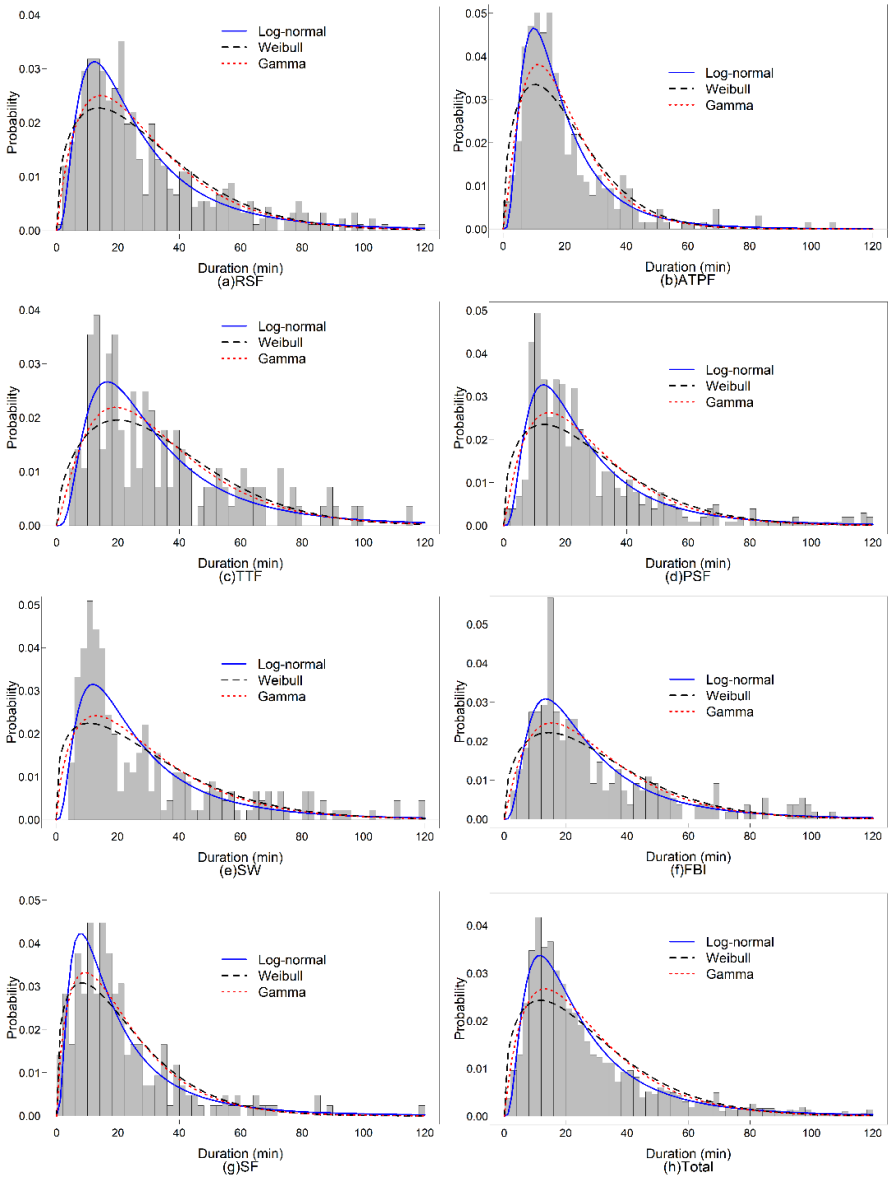

Figure 7: Fitting results of duration time of each disruption category.

5.3 Goodness-of-fit testing

To select the model that has the best performance for each category, a Kolmogorov-Smirnov (K-S) (Massey Jr, 1951) method was used to test the goodness-of-fit of the models. K-S method tests if one random variable follows a theoretical distribution, or if two random variables have the same distribution. Its null hypothesis is as follows:

H0：a random variable follows a theoretical distribution, or two random variables have the same distribution.

Its test statistic (*T*) is the largest difference between the cumulative distribution function (CDF) of the data and the theoretical distribution, as described by (5). However, some random numbers, which follow an uniform distribution were added to the data in order to satisfy the continuity requirement of K-S because the historical train operation data were recorded in the minute timescale

$$T = max\left|F'(x) - F(x)\right| \tag{5}$$

where $F'(x)$ is the CDF of the observed data, which consists of the duration of each category, $F(x)$ is the CDF of the theoretical distribution models, which consists of three alternative distribution models. A significance level of $\alpha = 0.05$ was chosen for the test. As *T* becomes smaller, the sample distribution tends to follow the theoretical distribution. The K-S test results of all the models are summarized in Table 4.

The results indicate that Log-normal models fitted using RSF, ATPF, FBI, all samples, as well as all alternative models fitted using TTF, and SF samples, passed K-S test. However, the Log-normal models had the least *T* value. Meanwhile, no model based on PSF and SW samples passed K-S test. However, Log-normal model had the least *T* value, and the p-values were very close to α. Therefore, the CDF of Log-normal model had the smallest distance with that of PSF, and SW, and Log-normal model thus was chosen as the distribution model of all HSR disruption clusters. The parameters of each category are shown in Table 5. The fitted probability models can be used to estimate the duration of any disruption, once its causes are ascertained.

Table 4: KS testing result of each cluster.

Cause	Log-normal		Weibull		Gamma	
	T	p-value	T	p-value	T	p-value
RSF	0.028	0.863	0.078	0.007	0.067	0.031
ATPF	0.041	0.635	0.099	0.003	0.086	0.016
TTF	0.052	0.848	0.072	0.450	0.073	0.436
PSF	0.066	0.021	0.109	0.000	0.083	0.001
SW	0.094	0.034	0.119	0.003	0.129	0.000
FBI	0.037	0.843	0.093	0.017	0.081	0.052
SF	0.065	0.324	0.083	0.104	0.071	0.224
Total	0.031	0.729	0.077	0.001	0.073	0.016

Note: underline fonts mean passing K-S test

Table 5: Fitted Log-normal distribution parameters for each category.

Cause	Model	μ	σ	Cause	Model	μ	σ
RSF	Log-normal	3.100	0.772	SW	Log-normal	3.091	0.780
ATPF	Log-normal	2.760	0.685	FBI	Log-normal	3.138	0.733
TTF	Log-normal	3.301	0.707	SF	Log-normal	2.768	0.845
PSF	Log-normal	3.082	0.727	Total	Log-normal	3.029	0.766

6 Conclusion

In this paper, we investigated the spatiotemporal distribution and duration distribution characteristics of railway disruptions based on the historical disruption records of four HSR lines in China. The conclusions made are as follows:

1) The probabilities of railway disruptions are spatiotemporally different.
2) Railway disruptions can be classified into seven categories based on their causes and influence on the HSR system.
3) The statistical analyses of each category revealed that the average duration of TTF and SW is the highest and longer than 40 min, whereas ATPF has the least value.
4) The duration of each category can be well fitted using Log-normal distribution model.

The results can assist dispatchers in understanding the distribution characteristics of disruptions, thereby improving the quality of their decisions. In particular, they can obtain the real-time and future probabilities of disruption at any coordinates of the timetable to enable them develop strategies that can prevent the disruptions. Furthermore, they can estimate the duration of disruptions using fitted Log-normal distribution models in order to make better decisions. The probability models can also improve train operations and disruption management in simulated systems as they are more accurate than hypothetical models. Hypothetical models introduce certain gaps into the simulations and usually overestimate or ignore some situations and constraints of train operations, which are needed by dispatchers in rescheduling the timetable.

Acknowledgments

This work was supported by the China Scholarship Council [grant number 201707000038]; National Nature Science Foundation of China [grant numbers 71871188, 61503311]; Science & Technology Department of Sichuan Province [grant number 2018JY0567]; and the Doctoral Innovation Fund Program of Southwest Jiaotong University [grant number D-CX201827]. We are grateful for the useful contributions made by our project partners, and we would like to thank the China Railway Guangzhou Group Co., Ltd for the data support.

References

Corman, Francesco, Andrea D'Ariano, and Ingo A Hansen. 2014. "Evaluating disturbance robustness of railway schedules." *Journal of Intelligent Transportation Systems* 18 (1):106-120.

Hartrumpf, M., T. Claus, M. Erb, and J. M. Albes. 2009. "Surgeon performance index: tool for assessment of individual surgical quality in total quality management."

European Journal Of Cardio-Thoracic Surgery 35 (5):751-758. doi: 10.1016/j.ejcts.2008.12.006.

Higgins, Andrew, E Kozan, and L Ferreira. 1995. "Modelling delay risks associated with train schedules." *Transportation Planning and Technology* 19 (2):89-108.

Huisman, Tijs, and Richard J. Boucherie. 2001. "Running times on railway sections with heterogeneous train traffic." *Transportation Research Part B: Methodological* 35 (3):271-292.

Janecek, David, and Frédéric Weymann. 2010. "LUKS-Analysis of lines and junctions." Proceedings of the 12th World Conference on Transport Research (WCTR).

Jespersen-Groth, Julie, Daniel Potthoff, Jens Clausen, Dennis Huisman, Leo Kroon, Gábor Maróti, and Morten Nyhave Nielsen. 2009. "Disruption management in passenger railway transportation." In *Robust and online large-scale optimization*, 399-421. Springer.

Kecman, Pavle, and Rob MP Goverde. 2015. "Predictive modelling of running and dwell times in railway traffic." *Public Transport* 7 (3):295-319.

Kellermann, Patric, Christine Schönberger, and Annegret H Thieken. 2016. "Large-scale application of the flood damage model RAilway Infrastructure Loss (RAIL)." *Natural Hazards and Earth System Sciences* 16 (11):2357-2371.

Lessan, Javad, Liping Fu, and Chao Wen. 2018. "A hybrid Bayesian network model for predicting delays in train operations." *Computers & Industrial Engineering*. In press, DOI:10.1016/j.cie.2018.03.017.

Lessan, Javad, Liping Fu, Chao Wen, Ping Huang, and Chaozhe Jiang. 2018. "Stochastic Model of Train Running Time and Arrival Delay: A Case Study of Wuhan–Guangzhou High-Speed Rail." *Transportation Research Record*. DOI:10.1177/0361198118780830.

Massey Jr, Frank J. 1951. "The Kolmogorov-Smirnov test for goodness of fit." *Journal of the American statistical Association* 46 (253):68-78.

Murali, Pavankumar, Maged Dessouky, Fernando Ordóñez, and Kurt Palmer. 2010. "A delay estimation technique for single and double-track railroads." *Transportation Research Part E: Logistics and Transportation Review* 46 (4):483-495. DOI: 10.1016/j.tre.2009.04.016.

Nash, Andrew, and Daniel Huerlimann. 2004. "Railroad simulation using OpenTrack." *WIT Transactions on The Built Environment* 74.

Olsson, Nils OE, and Hans Haugland. 2004. "Influencing factors on train punctuality—results from some Norwegian studies." *Transport policy* 11 (4):387-397.

Takimoto, T. 2000. "Development of efficient operational control using object representation." *Computers In Railways Vii* 7:837-841.

Wallander, Jouni, and Miika Mäkitalo. 2012. "Data mining in rail transport delay chain analysis." *International Journal of Shipping and Transport Logistics* 4 (3):269-285.

Wen, Chao, Zhongcan Li, Javad Lessan, Liping Fu, Ping Huang, and Chaozhe Jiang. 2017. "Statistical investigation on train primary delay based on real records: evidence from Wuhan–Guangzhou HSR." *International Journal of Rail Transportation* 5 (3):170-189. DOI:10.1080/23248378.2017.1307144.

Weng, Jinxian, Yang Zheng, Xiaobo Qu, and Xuedong Yan. 2015. "Development of a maximum likelihood regression tree-based model for predicting subway incident delay." *Transportation Research Part C: Emerging Technologies* 57:30-41.

Wiklund, Mats. 2003. "SERIOUS BREAKDOWNS IN THE TRACK INFRASTRUCTURE: CALCULATION OF EFFECTS ON RAIL TRAFFIC." *VTI MEDDELANDE* (959).

Zilko, Aurelius A, Dorota Kurowicka, and Rob MP Goverde. 2016. "Modeling railway disruption lengths with Copula Bayesian Networks." *Transportation Research Part C: Emerging Technologies* 68:350-368.