Mining Train Delay Propagation Pattern from Train Operation Records in a High-Speed System

Ping Huang [a,b,c,1], Chao Wen [a,b,c] Zhongcan Li [a,b]
[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

This study aims to investigate delays, delay increases, and delay recovery characteristics, by using statistical methods to clarify delay propagation patterns according to historical records of the Wuhan-Guangzhou high-speed railway (HSR) in China in 2014 and 2015. Specifically, we examined arrival and departure delay duration distributions and used heatmaps to demonstrate the spatiotemporal frequency distribution of delays, delay increases, and delay recovery, and the heatmaps clearly show hot spots (coordinates with high frequencies) in a timetable. Then, we separated delays as discrete intervals according to their severity, and analyzed the delay increasing frequency and the delay increasing severity within each interval, so as to clarify the relationships of delay increasing probability and delay increasing severity with delay extents. Next, we investigated the observed delay recoveries and prescheduled buffer times at (in) station (section), which demonstrate the recovery ability of each station and section. Finally, to understand the key influencing factor of delay propagation, we analyzed the relationship between capacity utilization and delays, delay increases, and delay recoveries, by examining their Pearson correlation coefficients. These indicate that delay frequencies and delay increasing frequencies with Pearson correlation coefficients as high as 0.9 are highly dependent on capacity utilization. The uncovered delay propagation patterns can enrich dispatchers' experience, and improve their decision-making ability during real-time dispatching in HSR.

Keywords

High-speed railway, train delays, delay increases, delay recoveries, capacity utilization

1 Introduction

Train operations are subject to various disturbances, such as severe weather, power outages, and facility failure, and all of these can result in train delays and lead to considerable losses for both railway operators and travellers (Khadilkar, 2016). For instance, the statistics from a Dutch railway network show that infrastructure-related disruptions occur approximately 22 times per day, and each disruption on average can last 1.7 hours (Jespersen-Groth et al. 2009). The Austrian Federal Railways had to cope with financial losses of more than EUR 100 million every year due to flooding (Kellermann, Schönberger and Thieken, 2016). In China, the average departure punctuality in origination stations was as high as 98.8% in 2016, but because of various disturbances during their operations, the average punctuality at the final destination stations was less than 90%, though delays smaller than 5 minutes are considered punctual (Lessan et al, 2018). For the train dispatchers, the key steps to reduce loss are not only managing the unexpected delays, but also making decisions in advance. In other words, if the dispatchers can know the delay probabilities at different times and locations and how the delay would evolve and propagate, they can make decisions before train delays that can result in more efficient timetable re-scheduling. Therefore, examining patterns of delay propagation is of great significance for improving railway delay management and real-time decision-making abilities.

However, accurate train delay and propagation pattern recognition presents challenges, mainly because: 1) disturbances are totally unexpected; 2) delay propagation is spatiotemporal, and its influencing factors are complex; and 3) prescheduled time supplements and buffer times cannot be fully utilized (delay recoveries are stochastic). In practice, some skilled dispatchers usually predict delays and delay propagation empirically, leading to different decision-making standards even for the same dispatcher. Data-mining 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 (Wallander and Mäkitalo, 2012). Historical train operation records can be regarded as the interactive consequences of all potential influencing factors, and this supports us exploring the rules of delays and propagation from their historical performances, rather than fitting functional expressions. Hence, mining the delay and propagation patterns from historical operation records can provide more accurate and comprehensive results for railway operation managers.

This study aims to recognize train delay and delay propagation patterns from train operation data of the Wuhan-Guangzhou (W-G) HSR in China. To this end, we first analyzed the duration and spatiotemporal distribution of train delays. Next, we split train delays as discrete intervals with a width of 5 minutes according their length, and investigated the delay increasing probabilities and severity on different delay extents. We also examined the delay recovery abilities and prescheduled buffer times at(in) each station(section). Finally, in order to understand the influences of capacity utilization on delays, delay increases, and delay recoveries, we investigated their relationships by calculating Pearson correlation coefficients.

2 literature review

Generally, train delays are caused by 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 (Olsson and Haugland, 2004; Hartrumpf et al, 2009; Higgins, Kozan and Ferreira 1995). In addition, if the running and dwell times increase due to unexpected disturbances, it can result in knock-on delays and delays for other trains (Milinković et al, 2013). Serious disruptions such as switch or signal failures, if not managed effectively, can result in queuing of trains, creating a chain of delayed trains. The experience from the Taiwan HSR shows that shortening the maintenance cycle can effectively alleviate the problem of train delay caused by signal failures (Hasan, 2011). 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 (Yuan, Goverde and Hansen, 2002; Higgins and Kozan, 1998). 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 (Wen et al, 2017). A qexponential 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). Given the HSR operation data, the maximum likelihood estimation method was used to determine the probability distribution of the different disturbance factors and the distributions of affected trains. However, the models of primary delay consequences have not been established in detail (Xu, Corman and Peng, 2016). 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 (Liang et al, 2009).

Data-driven research studies proposed for delay management mainly focused on using regression or distribution approaches to fit delay data. (Milinković et al, 2013) 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. 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 (Wallander and Mäkitalo, 2012). (Murali et al, 2010) reported a delay regression-based estimation technique that models delay as a function of train mix and network topology. 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 (Goverde, 2005). Recently, (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. A hybrid Bayesian network model is also established to predict arrival and departure delays for the Wuhan-Guangzhou HSR (Lessan, Fu and Wen, 2018).

3 Data description

The data used in this study are the real-world train operation records of the Wuhan-Guangzhou HSR in China. This line connects to the Guangzhou-Shenzhen HSR line at GZS station, to the Hengyang-Liuzhou HSR line at HYE station, and to the Shanghai-Kunming HSR line at CSS station, respectively. All the trains operating on this line are equipped with the Chinese Train Control System (CTCS), which allows a maximum speed of 350 km/h, and the Automatic Train Supervision system that records the movements of all trains. We considered data from trains in the northbound direction comprising the segment from Guangzhou South (GZS) to Changsha South (CSS), as shown in Figure 1. The collected data contain 57,796 HSR trains in the GZS-HYE section and 64,547 HSR trains in the HYE-CSS segment, comprising information about train operations from March 24, 2015 to November 10, 2016. The scheduled/actual arrival/departure records of each train and station, the number of trains, dates, occupied tracks, and section lengths were collected to construct a database with data recorded every minute. Figure 2 shows the accumulative HSR trains of each station during each hour, clearly indicating the differences of train services along with space and time axes. Therefore, in the following sections, we will investigate the spatiotemporal differences of delay and delay propagation characteristics.

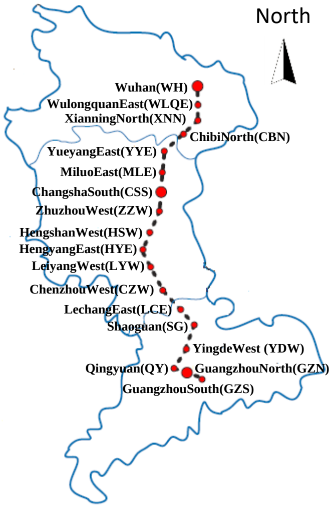

Figure 1: Map of Wuhan–Guangzhou high-speed railway line

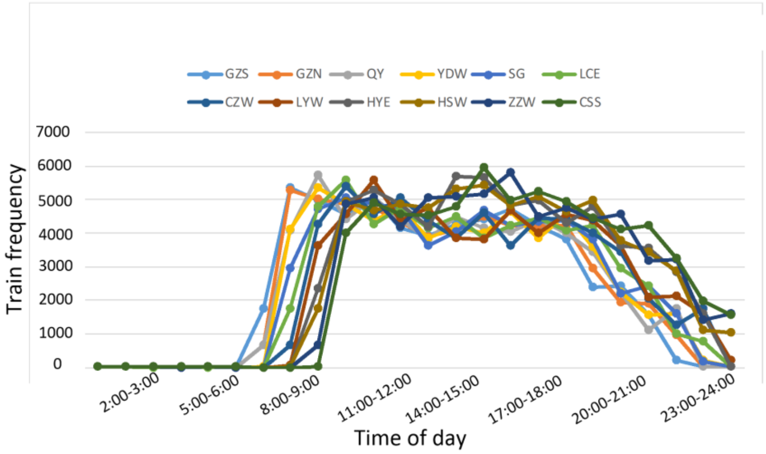

Figure 2: Cumulative HSR trains per hour

4 Delay Characteristics Investigation

Longer delays can have stronger influences on railway systems and can propagate farther, whereas shorter delays have smaller influences on railway systems, or can even be assimilated at the moment of occurrence. To understand their characteristics, we first visualized the duration distribution of arrival and departure delays. The histograms in Figure 3 clearly show that both arrival and departure delays follow a right-skewed and heavy-tailed distribution, which indicates that the longer the delays, the lower the frequencies. Also, train delays can propagate along time and space axes, which can result in different delay frequencies in a timetable. To understand the spatiotemporal distribution pattern of train delays, we analyzed the frequency distribution of delays. We separated delays as longer than 4 minutes and as longer than 30 minutes, to better understand the spatiotemporal distribution characteristics of different delay severities. The length of 4 minutes was chosen because it is the criteria set by the Chinese Railway Company to label trains as delayed, and the length of 30 minutes was chosen to understand the spatiotemporal distribution of longer delays. Figure 4 and Figure 5 clearly show that both 4 minute and 30-minutes-or-longer delay frequencies at the original station are extremely low, but they became much higher with the operation of trains. In addition, along the time axis, their frequencies are low during off-peak hours, and high during peak hours. In short, the hot spots of both 4 minute and 30-minutes-or-longer delays appear during 14:00 to 20:00, in the LCE-CSS section.

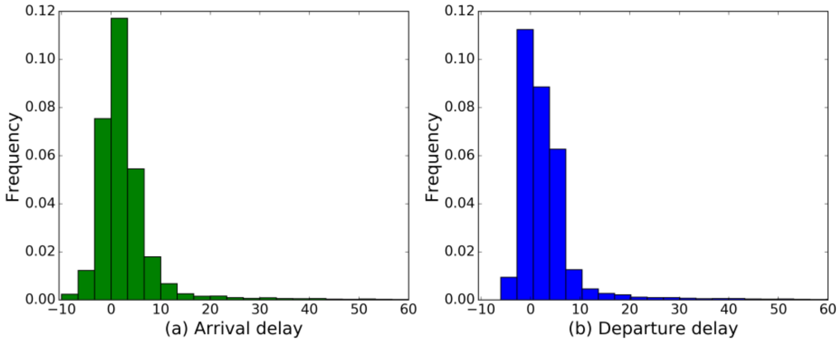

Figure 3: Delay length (min) distribution

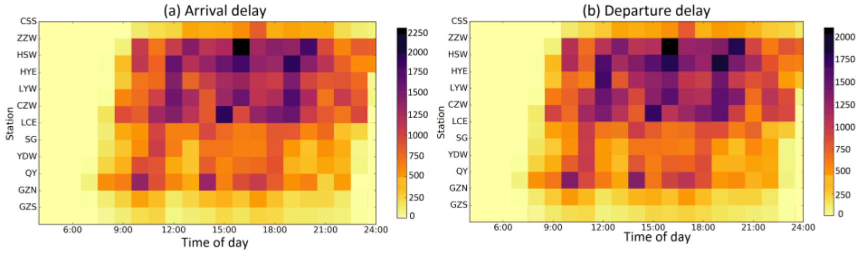

Figure 4: Spatiotemporal distribution of delays

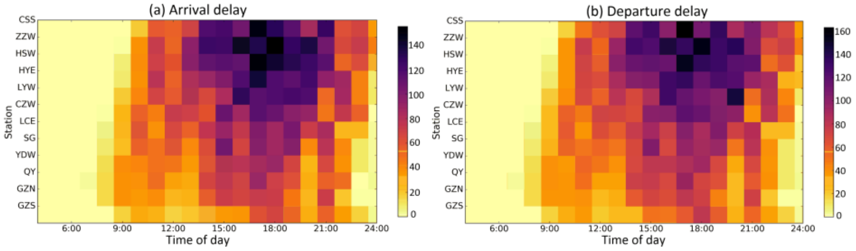

Figure 5: Spatiotemporal distribution of delays longer than 30 minutes

5 Delay propagation pattern investigation

5.1 Delay increasing characteristics

Delays can increase due to secondary disturbances. In order to understand the delay increasing pattern in the timetable, we first made statistics about the delay increasing (growth more than 4 minutes) frequency, at station and in section. In this process, a train whose departure delay was 4 minutes longer than its previous arrival delay was labeled as delay increase at station, and an arrival delay that was 4 minutes longer than its previous departure delay was labeled as delay increase in section. Figure 6 clearly shows that the delay increasing frequency at station is high at junction stations (CSS, HYE, and GZS). This conclusion is understandable, as the junction stations have more tasks (such as trains turning-over, crossing-line, and terminating) than other stations, which makes the equipment utilization more frequent, leading to the higher disturbance probability. However, without evident task volume differences, probabilities in sections do not appear with any explicable regularity, as they are mainly related to equipment status, climate, weather, and the experience and skills of dispatchers.

Then, we conducted sensitive analyses of delay increasing frequency and delay increasing severity on different delay severities. We transferred train delays as discrete intervals with a width of 5 minutes, and separated delays into the intervals that they fall in. Likewise, the sensitive analyses were also conducted on delay increases at station and in section, as shown in Figure 7 and Figure 8.

Figure 7 shows that both delay increasing frequencies at station and in section rise with the growth of delay extent, meaning that longer delays are more likely to encounter secondary disturbances. An exception happens on the early-arrival-trains interval (the first interval), where the delay increasing probability at station is abnormally high, but that is not in the case in section. This can be explained by the dispatching principle that trains are only allowed to arrive early, but cannot depart early due to the passenger boarding requirements. Therefore, early arrival trains tend to be given more dwelling times to depart on schedule. Figure 8 shows the sensitivity of delay increasing severity, where, at stations, it is shorter with the growth of delay length, but, in sections, it keeps stable with the growth of delay length. Also, an exception happens on the interval of early arrival trains at station (the first interval). This result was caused by the recovery of early arrival trains, as their early arrival times (the smallest is -10 minutes) are not as long as delays (can reach 190 minutes), thus limiting their increasing extent.

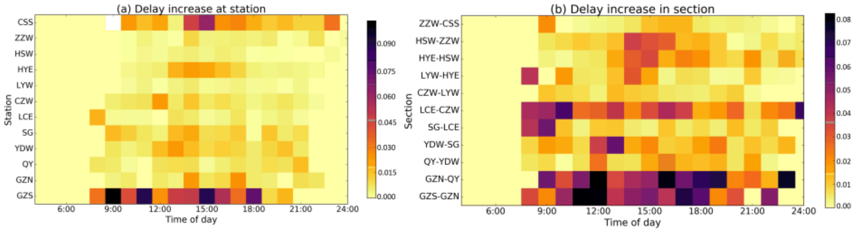

Figure 6: Spatiotemporal frequency distribution of delay increase

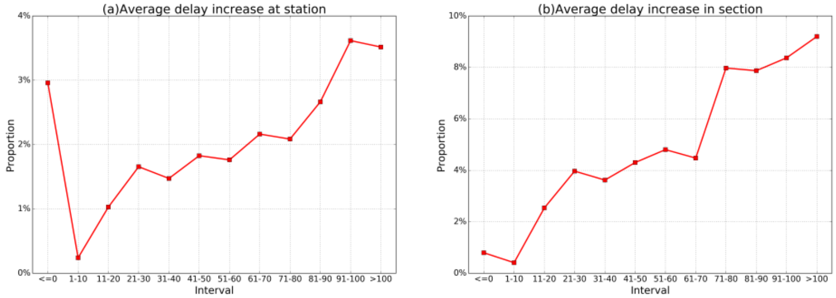

Figure 7: Sensitive analysis of delay increasing frequency on delays severity

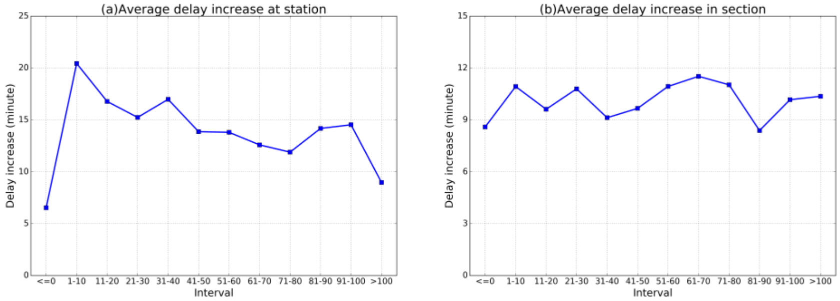

Figure 8: Sensitive analysis of delay increasing severity on delays severity

5.2 Delay recovery characteristics

Delays can be recovered using buffer times prescheduled in sections and stations. To understand the delay recovery characteristics of each station and section, we conducted statistics analyses about the spatiotemporal probability distribution of delay recoveries, which is the proportion of the trains with delay recoveries to all delayed trains, as shown in Figure 9. Like the spatiotemporal distributions of delay increases in sections, delay recoveries do not have centralized hot spots, but their probabilities in section are much higher than those at station. Besides, delay recovery probabilities at(in) one station(section) are evidently different from others, coordinating the empirical conclusion that delay recoveries are dominantly influenced by the buffer times distribution in the timetable. In practice, the buffer time allocation methods differ from timetables/railway lines, such as allocating according to section length and travel times or according to the specific recovery requirements of sections (Huang et al, 2018). Therefore, we investigated the observed delay recoveries and pre-scheduled buffer times of each station and section as shown in Figure 10 and Figure 11, respectively. Figure 10 denotes the comparisons of scheduled running (dwell) times and practical running (dwell) times at(in) each station(section). Comparisons of the bar pairs indicate that the practical running and dwelling times are smaller than the scheduled running and dwelling times, implying that buffer times were somewhat effective in reducing delays at(in) station(section). However, different stations and sections appear to have different recovery values, as the left-hand bars were ranked from small to large (from top to bottom), but the right-hand bars were opposed to this rule, and the recovery volumes in sections are lower than those at stations. We thus calculated the available (prescheduled) buffer times at each station using prescheduled running times and minimum running times, and those in sections using prescheduled dwell times and minimum dwell times, given by (1) and (2).

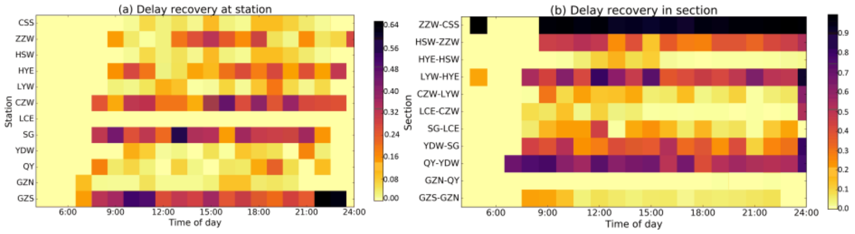

Figure 9: Spatiotemporal distribution of delay recovery probabilities

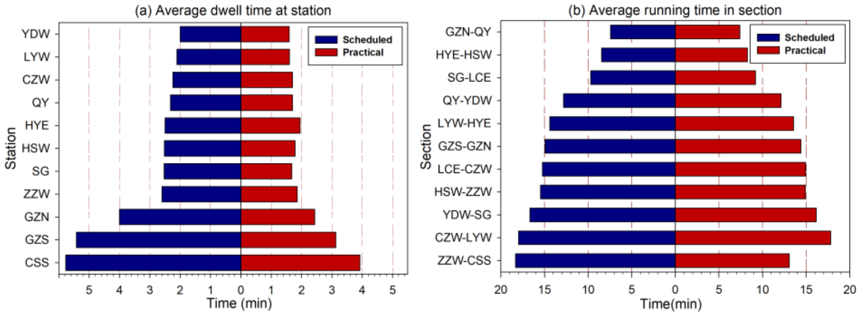

Figure 10: Comparison of scheduled and practical dwell (running) times

$$BT_{\text{station}} = T_{\text{station}} - T_{\min} \, , \qquad (1)$$

$$BT_{\text{section}} = T_{\text{section}} - {}^{L}\!\!\diagup\!_{S_{\max}} \, . \qquad (2)$$

where BT_{station} and BT_{section} are buffer times at(in) station(section); T_{station} and T_{section} are prescheduled dwell times and running times at(in) station(section); $T_{\min}$ is the minimum dwell times of trains at station, where, at junction stations, it is 2 minutes, and at other stations, it is 1 minute; L is the distance of every adjacent station; and S_{max} is the maximum speed of HSR trains, i.e., 310 km/h during the time span in the collected data, according to the technique documents from China Railway Company. The available buffer times distributions are shown in Figure 11, where buffer times value is extremely high at the GZS station and in the ZZW-CSS section, which can explain the high probabilities at GZS stations and in the ZZW-CSS section in Figure 9, and the large recovery ability of ZZW-CSS in Figure 10.

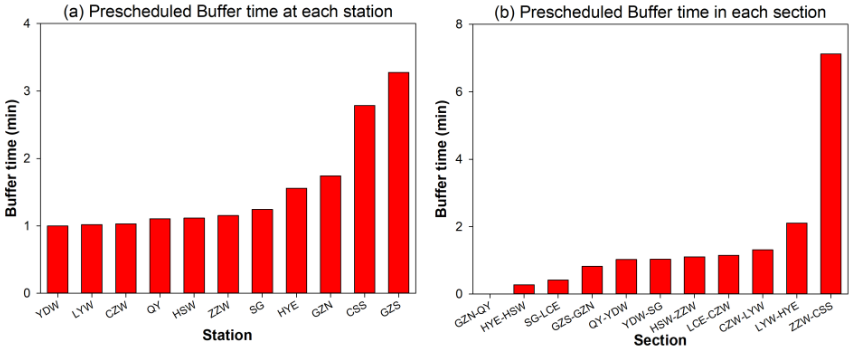

Figure 11: Prescheduled buffer times at(in) each station(section)

5.3 Correlation analyses with capacity utilization

Figure 2, Figure 4, Figure 5, and Figure 6 indicate that both train delays and delay increasing have high frequencies during peak hours. To quantitatively estimate their relationships with capacity utilization, we calculated the Pearson correlation coefficients (φ , see (4)) of delays and delay increases with the number of trains per hour (N), given by (3).

$$N = {}^{N_{total}}\!/_{d} \; . \tag{3}$$

where N_{total} is the total train services of each hour shown in Figure 2, and d is the number of days the data included.

$$\varphi_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(Y)}\sqrt{E(Y^2) - E^2(X)}} \; . \tag{4}$$

In the above equation, X and Y are two variables, and $E(X)$ and $E(Y)$ are the expectations of X and Y, respectively.

Table 1 clearly shows that delays (including arrival and departure delays) and delay increases (including delay increases at station and in section) have strong relationships with the number of trains per hour whose Pearson correlation coefficients can reach as high as 0.9, but the Pearson correlation coefficients between delay recoveries and the number of trains per hour are as low as 0.413 and 0.598 at station and in section, respectively. Pearson correlation coefficients further clarify the quantitative relationships between the number of trains per hour with delays, delay increases, and delay recoveries. Specifically, the scatter-lines in Figure 12 and Figure 13 show the matching effects of delays and delay increases with the number of trains per hour. Considering these figures and the Pearson correlation coefficients, the linear relationship between the probabilities of delays and delay increases with the number of trains per hour is high.

Table 1: Pearson correlation coefficients of delays, delay increases, and delay recoveries and capacity utilization

Arrival delays	Departure delays	Delay increase at station	Delay increase in section	Delay recovery at station	Delay recovery in section
0.936	0.933	0.915	0.945	0.413	0.598

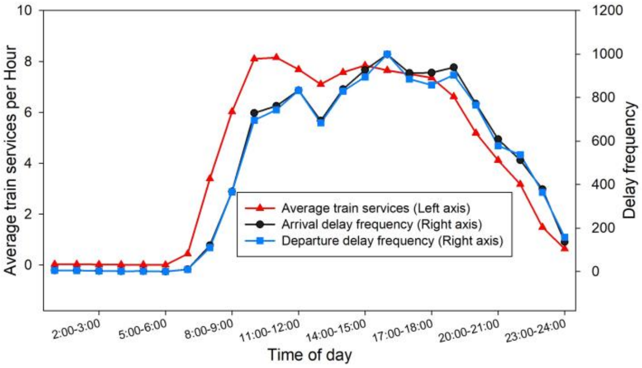

Figure 12: Correlation of delay and capacity utilization

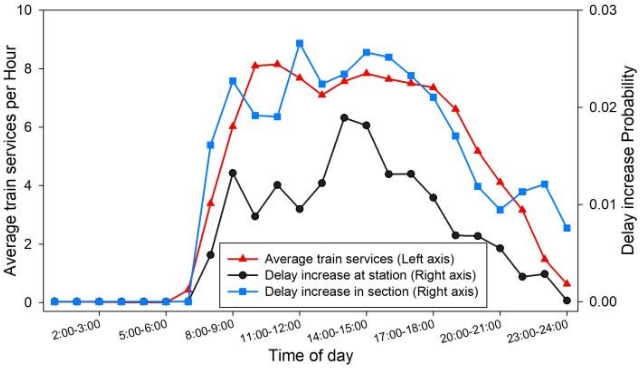

Figure 13: Correlation of delay increase and capacity utilization

6 conclusion

The paper presents how to recognize train delays and delay propagation patterns from historical train operation records. The following conclusions were obtained.

1) Train delay frequencies and delay increasing probabilities are spatiotemporally different, and are highly dependent on capacity utilization; the more the capacity utilization, the higher probabilities of delays and delay increases (with Pearson correlation coefficients over 0.9).

2) For both arrival and departure delays, the longer the delays, the higher the delay increasing probabilities.

3) Longer arrival delays could result in shorter delay increases, whereas departure delays do not influence the delay increasing extent.

4) The delay recoveries, which are mainly influenced by prescheduled buffer times, have higher probabilities but lower volumes in section as compared against those at stations.

The spatiotemporal probabilities and analyses on delay increase and recovery can help dispatchers improve their decision-making qualities. Explicitly, with the spatiotemporal distributions, the dispatchers can obtain the real-time and future probabilities of delays, delay increases, and delay recoveries. With the sensitivity analyses between delays and delay increase, the dispatchers can acquire their increase probabilities and severities under any delay length; with the relationship analyses between delay recoveries and total buffer times, the dispatchers can have a better understanding of the recovery abilities of each station and section. Additionally, the spatiotemporal probabilities can also be applied to train operation simulation systems to optimize disturbance setting and timetable rescheduling programs, as they are more practical than hypothetical models that bring certain gaps between simulations and practice, and usually over assume and ignore some situations and constraints of train operations.

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

Goverde, Rob MP. 2005. "Punctuality of railway operations and timetable stability analysis." Delft.

Hartrumpf, Martin, Thomas Claus, Michael Erb, and Johannes 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.

Hasan, Nazmul. 2011. "Direct Fixation Fastener (DFF) Spacing and Stiffness Design." 2011 Joint Rail Conference, *American Society of Mechanical Engineers*.

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

Higgins, Andrew, and Erhan Kozan. 1998. "Modeling train delays in urban networks." *Transportation Science* 32 (4):346-357.

Huang, Ping, Chao Wen, Qiyuan Peng, Javad Lessan, Liping Fu, and Chaozhe Jiang. 2018. "A data-driven time supplements allocation model for train operations on high-speed railways." International Journal of Rail Transportation:1-18.

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.

Khadilkar, Harshad. 2016. "Data-enabled stochastic modeling for evaluating schedule robustness of railway networks." *Transportation Science* 51 (4):1161-1176.

Lessan, Javad, Liping Fu, and Chao Wen. 2018. "A hybrid Bayesian network model for predicting delays in train operations." *Computers & Industrial Engineering*. 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

Liang, Zhang, Liu Jianhua, Wu Ruofei, and Gong Xiaobin. 2009. "Design of Performance Testing System for Train Air Conditioning." Energy and Environment Technology, 2009. ICEET'09. International Conference on IEEE 1: 85-89.

Milinković, Sanjin, Milan Marković, Slavko Vesković, Miloš Ivić, and Norbert Pavlović. 2013. "A fuzzy Petri net model to estimate train delays." *Simulation Modelling Practice and Theory* 33:144-157. DOI: 10.1016/j.simpat.2012.12.005.

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.

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." *WIT Transactions on The Built Environment* 50.

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):1-20.

Xu, Peijuan, Francesco Corman, and Qiyuan Peng. 2016. "Analyzing railway disruptions and their impact on delayed traffic in Chinese high-speed railway." *IFAC-PapersOnLine* 49 (3):84-89.

Yuan, J, RMP Goverde, and IA Hansen. 2002. "Propagation of train delays in stations." *WIT Transactions on The Built Environment* 61.