Rail. Eng. Science (2020) 28(4):424–434
https://doi.org/10.1007/s40534-020-00225-8

Delay recovery model for high-speed trains with compressed train dwell time and running time

Yafei Hou[1,2] · Chao Wen[1,2] · Ping Huang[1,2] · Liping Fu[3] · Chaozhe Jiang[1,2]

Received: 9 May 2020 / Revised: 28 October 2020 / Accepted: 3 November 2020 / Published online: 24 November 2020

Abstract Modeling the application of train operation adjustment actions to recover from delays is of great importance to supporting the decision-making of dispatchers. In this study, the effects of two train operation adjustment actions on train delay recovery were explored using train operation records from scheduled and actual train timetables. First, the modeling data were sorted to extract the possible influencing factors under two typical train operation adjustment actions, namely the compression of the train dwell time at stations and the compression of the train running time in sections. Stepwise regression methods were then employed to determine the importance of the influencing factors corresponding to the train delay recovery time, namely the delay time, the scheduled supplement time, the running interval, the occurrence time, and the place where the delay occurred, under the two train operation adjustment actions. Finally, the gradient-boosted regression tree (GBRT) algorithm was applied to construct a delay recovery model to predict the delay recovery effects of the train operation adjustment actions. A comparison of the prediction results of the GBRT model with those of a random forest model confirmed the better performance of the GBRT prediction model.

✉ Chao Wen
wenchao@swjtu.edu.cn

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

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

[3] Intelligent Transport Systems Center, Wuhan University of Technology, Wuhan 430070, China

Keywords High-speed train · Delay recovery · Train operation adjustment actions · Gradient-boosted regression tree

1 Introduction

High-speed trains encounter many random disturbances during operation that can cause train delays. The anti-interference ability of the train timetable and the ability to recover from delays after being affected by disturbances are among the most critical concerns that affect the service quality of high-speed railways. Delay recovery refers to the reduction of the delay time, and effective train delay recovery is a top priority in the daily work of dispatchers. When a train is delayed, the dispatcher should select an appropriate train operation adjustment action to coordinate with the operation situation and transport needs; this decision is made on the basis of the dispatcher's own experience and, of course, the dispatching rules. There are three widely used train adjustment actions, namely compressing the train dwell time at stations, compressing the train running time in sections, and changing the travel interval. Different train recovery effects will occur under different transportation situations and train operation adjustment actions. It is often difficult to quantify and model the dispatcher's decision-making process as the effects of different train operation adjustment actions.

Real-world train operation records, e.g., actual train timetables, records of delay causes, and equipment utilization statuses, have been well preserved. These records include a wealth of useful information, such as the trains'

arrival and departure times, delays, and the relationships between trains. The establishment of models that can quantify the delay recovery effects under different train operation adjustment actions will be of considerable significance to support the development of intelligent dispatching for high-speed rail systems.

Based on the operation performance of a high-speed train, the decision-making process and rules used by dispatchers are determined in this study based on existing data. Then, taking the delay recovery time as the evaluation criterion for the effects of train operation adjustment actions, the gradient-boosted regression tree (GBRT) machine learning model is used to establish delay recovery models under different operation adjustment actions. First, the characteristic factors that affect delay recovery under different dispatching actions are extracted from actual train operation data, and stepwise regression is then used to select the characteristic parameters from the multivariate variables based on their degrees of influence. Finally, an algorithm is used to establish models that consider these influencing factors in the delay recovery process when employing the compression of the train dwell time at stations and the compression of the train running time in sections, and the relationships between the factors and their degrees of impact on delay recovery are analyzed.

Models are constructed to reflect the mechanisms by which delay is recovered using different train operation actions, the most common of which are changing the running time or dwell time. Trains that arrive before the scheduled time can usually be easily addressed, and there are very few cases in which dispatchers need to adopt actions to recover the delays. Commonly, trains with positive arrival delays will not depart from the station before the scheduled departure time.

The presented delay recovery model can better learn the delay recovery effect of each scheduling action under different train delay conditions, operation diagram structural parameters, and redundant time layouts. The model can elaborate the delay recovery effects of different train operation adjustment actions according to the parameters, and can provide decision-making assistance to the dispatcher; i.e., the dispatcher can make decisions according to the real-time prediction of relevant parameters. Knowing the possible effect of a train operation adjustment action can subsequently realize auxiliary scheduling decisions.

2 Literature review

The delay of a high-speed train not only leads to the change of the established train schedule and the unreasonable allocation of resources, but also causes the entire transportation system to fluctuate and reduces the quality of transportation service due to the transmission characteristics of train delay. Train delays are caused by a variety of unpredictable interference factors. According to the different types of interference, the causes of delay can be categorized as either internal or external factors. According to statistical analysis, equipment failure, vehicle failure, pantograph signaling network failure, and line failure are the main factors that lead to delay. Errors in dispatchers' scheduling commands and the reception and dispatching of train operation information by the station staff may lead to train conflicts and delays due to resource competition. Moreover, the large passenger flow in the station during peak periods will extend the established boarding and landing times and interfere with the normal operation process, ultimately resulting in the late departure of the train from the station.

Traditional mathematical model-driven methods are widely used by scholars to study supplement time scheduling and train operation conflict resolution to realize train delay recovery. These methods mainly include queuing theory, optimal scheduling algorithms, and optimization theory.

Dispatchers often set supplement times by extending train running times and increasing dwell times to achieve train delay recovery. Train timetable elasticity is considered an important criterion for the description and measurement of delay absorption and recovery capabilities, but an increase in the supplement time will result in wasted capacity [1]. Simulation methods are considered effective for the establishment of joint delay recovery models [2, 3]. Krasemann [4] used a depth-first greedy algorithm to assist the integer adjustment model for train operation adjustment planning and considered the supplement time of the operation timetable as a tool to eliminate the effect of random interference on train operation. Yuan et al. [5] established a stochastic theoretical model of delay propagation at stations that took into account the recovery parameters for delayed trains at stations and in sections. Stochastic analysis models, integer programming models, decision tree methods, and depth-first greedy algorithms have commonly been used to study the problems of delay recovery and supplement time scheduling [4–7]. Dollevoet et al. [8] proposed a delay management method based on the micrograph model of surrogate graphs, and iteratively solved a microscopic train scheduling optimization problem to minimize passenger operation delays and passenger transfer conflicts. Cheng [9] established a train operation network graph model and proposed train scheduling actions considering conflict resolution measures, such as the priority of the arrival train, local optimization, and the critical path method.

Benefiting from the development of railway informatization and big data technology in recent years, research on

dispatch modeling based on railway operation performance has gradually increased [10]. Commercial simulation software, such as OpenTrack and RailSys, has also been developed and widely used [11, 12]. Many studies have also addressed the issues of the comparison and selection of train operation adjustment actions and delay recovery.

By implementing different scheduling actions, the dispatcher can change the interval running time, dwell time, and train interval to achieve delay recovery. Interval running time estimation is a problem to which researchers have devoted considerable attention. It requires the consideration of factors such as random interference and the setting of supplement time during the entire operation of the train. Regarding the statistical and stochastic models for the establishment of train delay propagation models, Yamamura et al. [13] first used heat maps to visualize train operation data, and then performed static and dynamic analyses on the delay at each station; finally, methods by which to reduce train delays were proposed. Naohiko [14] discussed delay recovery actions for trains in the Tokyo Circle. Liebchen et al. [15] introduced the concept of recovery robustness and proposed an optimization action for train delay recovery based on actual train performance. Kecman and Goverde [16] developed a model by collecting running times and dwell times from a training set and investigated a machine model for accurate prediction of train event times based on a timed event graph with dynamic arc weights after a train was delayed [17]. It is also worth noting that Khadilkar [18] studied the delay distribution probability, and the analysis of historical data revealed that the average delay recovery rate of the Indian Railway was 0.13 min/km. This value was then used as the delay recovery ability in the constructed delay recovery model; however, it is difficult to reflect the delay recovery ability of a train in each section and station using the average value, which will affect the prediction ability of the model. Regarding domestic research, Guo et al. [19] regarded train operation as a series of discrete events based on the Beijing–Shanghai high-speed railway, and the operating performances of five railway stations were used to establish a linear regression model for delay prediction. Based on the train performance of the Wuhan–Guangzhou high-speed railway, a chain process of delay propagation [20], multiple linear regression of the initial delay recovery, and a random forest model [21] have been investigated, and multiple types of machine learning models for the prediction of delay recovery have been compared [22].

The dispatcher can adjust the train operation, recover delays, and improve the robustness of the train timetable by utilizing the resource of supplement time among sections, stations, and trains. Therefore, the advantages and disadvantages of a scheduling action are reflected in the use of supplement time to recover train delays. The layout and utilization of the supplement time vary with different stations, trains, sections, and time periods. Based on statistical methods, Yuan and Hansen [23] determined that with an increase in the supplement time between train lines, the absorption capacity for joint delays will increase exponentially, the supplement time available for dispatching may be greater, and the delay recovery effect will be improved. According to the guideline "UIC CODE 451-1 OR" issued by the International Railway Union, the scheduled supplement time is determined using the average travel distance or formation time of the train, and the supplement time scheduling scheme is calculated in minutes per kilometer (min/km) or as a percentage (%). However, statistical schemes do not distinguish between trains, stations, sections, etc., and the resulting scheme is thus not targeted. In recent years, scholars in different countries have studied the problem of delay recovery [24–26], and, based on random delays in arrivals and departures, the method of rearranging redundant time to improve the utilization of the supplement time has been gradually recognized [27, 28].

As described above, experts and scholars have carried out numerous related studies and achieved fruitful results in theories of railway traffic control; besides, more researchers have dedicated efforts to the application of big data in the modeling of railway operations, and more train operation data become available for train delay modeling. However, related theories and methods for high-speed railways, especially those for dispatching in China's high-speed railways, are still insufficient; there is a lack of research on the application of machine learning methods to train delays, delay recovery, and scheduling decision mechanisms; and there are also deficiencies in the regular extraction of scheduling decisions made by dispatchers. With the rapid development of machine learning methods, it is possible to apply these methods to establish data-driven models of train operation adjustment decision-making and delay recovery, but the research in this area is rarely reported.

3 Problem statement and data description

3.1 Problem statement

The process of a train changing from a delayed state to an on-time state, or of the reduction in the degree of a delay, is called delay recovery. The process of increasing the delay of a train is called a delay increase, while delay stasis occurs when the delay of a train remains unchanged during operation. As shown in Fig. 1, train i changes from the punctual state S_i^0 to a delayed state S_i^1. At this time, the

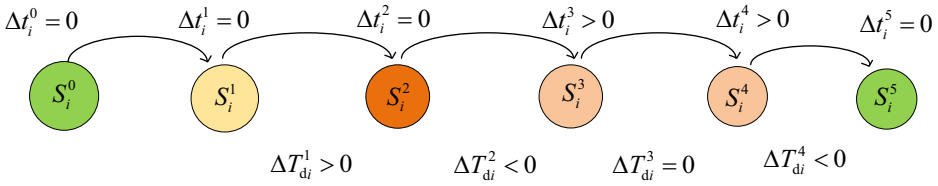

Fig. 1 Changes in the delay state of a train

degree of delay is small, after which it increases to reach state S_i^2. During the process of transitioning from S_i^2 to S_i^3, the degree of delay is reduced; i.e., the delay is recovered. During the process of transitioning from S_i^3 to S_i^4, the degree of delay is unchanged; i.e., the train remains delayed. Finally, after a further delay recovery, train i changes from a delayed state S_i^4 to an on-time state S_i^5. In this analysis, S $(S_i^0, S_i^1, \ldots, S_i^n)$ represents the state of a train. In addition,

$$\begin{cases} \Delta t_i^n = T_{ai}^n - T_{si}^n \\ \Delta T_{di}^n = \Delta t_i^{n+1} - \Delta t_i^n \end{cases},$$

where T_{ai}^n represents the actual arrival time in state S_i^n, T_{si}^n represents the scheduled arrival time in state S_i^n, Δt_i^n represents the delay time of train i in state S_i^n, ΔT_{di}^n represents the change in the delay time of delayed train i corresponding to the change from state S_i^n to S_i^{n+1}, and n represents the number of the state.

The dispatcher can adopt different train operation adjustment actions, which will affect the process by which train i changes from state S_i^n to state S_i^{n+1}. It is necessary to establish delay recovery models for different train operation adjustment actions to determine their effects on delay recovery. These models can predict the possible effects of the implementation of a certain train operation adjustment action under particular train operating parameters, thereby assisting the dispatcher in estimating the role of the selected dispatching action and improving the efficiency and accuracy of dispatching decisions.

3.2 Data description

This study was based on the operation of the Wuhan–Guangzhou high-speed railway from July 1, 2015 to June 30, 2016 from Guangzhou North Station to Chibi North Station; a total of 40,888 train running data points were utilized. The data were obtained from the scheduled and actual train operation timetables and included both the scheduled and actual train arrival and departure times. Via statistical analysis and data modeling, the delay recovery mechanisms for different scheduling actions were elucidated.

The data format is shown in Table 1, where the data can be divided into two types of recovery actions, namely the compression of the downtime and compression of the recovery intervals, based on the different adjustment actions employed. The respective data sets for these adjustment actions were extracted, yielding 26,877 and 61,963 data points, respectively. This indicates that, during actual train operation, the compression of the train dwell time at stations to achieve delay recovery occurs far less often (about 1/3) than the compression of the train running time in sections.

The six variables related to delay recovery are recorded as the delay recovery time (T_r), the delay time (T_d), the scheduled supplement time (T_s), the running interval (T_i), the occurrence time (T), and the place where the delay occurred (L). To investigate the correlation between variables, the time data variable T was converted into numerical data; for example, the time "2015/11/3 8:44" was converted to the value 8.73. In particular, a higher number indicates that the station or section occurs further toward the end of the line.

The Pearson correlation coefficient was used to analyze the correlations between variables when the compression of the train dwell time at stations was employed, as summarized in Table 2.

As can be seen from Table 2, in the recovery data for the compression of the train dwell time at stations, there is a strong positive correlation between T_r and T_s, and a weak positive correlation between T_r and T_d; in addition, there are negative correlations between T_r and each T_i, T and L.

The Pearson correlation coefficients for the six variables in the delay recovery data for the adjustment action of the compression of the train running time in sections were also analyzed, and the results are summarized in Table 3.

In the recovery data reported in Table 3, there are positive correlations between T_r and each T_d, T_s, T, and L; T_r has strong positive correlations with T_s and L, while it has weak positive correlations with T_d and T. Additionally, T_r has a negative correlation with T_i.

A significance test was conducted to verify that the sample data accurately represent the overall data characteristics. As presented in Table 4, the significance test results indicate that the correlation coefficients between T_r

Table 1 Data format

Train no.	Recovery time (min)	Delay time (min)	Scheduled supplement time (min)	Running interval (min)	Time	Place
G542	2	5	4	8	2016/6/18 11:16	CZW
G1104	3	3	5	3	2015/10/1 9:45	HYE
G280	1	1	5	6	2015/11/3 8:44	QY-YDW
G6110	1	5	9	7	2016/5/28 16:57	CZW-LYW
G432	2	2	14	41	2015/12/12 22:19	CSS-MLE

Table 2 Correlation coefficients between the delay recovery variables under the action of compression of the train dwell time at stations

–	T_r	T_d	T_s	T_i	T	L
T_r	1	0.1171	0.5855	− 0.1867	− 0.0187	− 0.1946
T_d	0.1171	1	0.0604	0.1066	0.0192	0.1243
T_s	0.5855	0.0604	1	− 0.1691	0.0549	− 0.1698
T_i	− 0.1867	0.1066	− 0.1691	1	0.0450	0.3478
T	− 0.0187	0.0192	0.0549	0.0450	1	0.2920
L	− 0.1946	0.1243	− 0.1698	0.3478	0.2920	1

Table 3 Correlation coefficients between the delay recovery variables under the action of compression of the train running time in sections

–	T_r	T_d	T_s	T_i	T	L
T_r	1	0.1355	0.3129	− 0.0237	0.1214	0.4149
T_d	0.1355	1	0.0826	0.0302	0.0918	0.1401
T_s	0.3129	0.0826	1	0.0531	0.2491	0.3641
T_i	− 0.0237	0.0302	0.0531	1	0.2056	− 0.0109
T	0.1214	0.0918	0.2491	0.2056	1	0.2313
L	0.4149	0.1401	0.3641	− 0.0109	0.2313	1

Table 4 Significance test results of the correlation coefficients under the action of compression of the train running time in sections

–	T_r	T_d	T_s	T_i	T	L
T_r	0.00	0.00	0.00	0.00	0.00	0.00
T_d	0.00	0.00	0.00	0.00	0.00	0.00
T_s	0.00	0.00	0.00	0.00	0.00	0.00
T_i	0.00	0.00	0.00	0.00	0.00	0.01
T	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.01	0.00	0.00

and each T_d, T_s, T_i, T, and L are all significant and nonzero; therefore, the original hypotheses of a correlation degree of zero can, respectively, be rejected.

4 Multiple variable feature selection based on stepwise regression

The correlation analysis revealed the nature and strength of the correlations between the six variables (T_r, T_d, T_s, T_i, T, and L) in the delay recovery data related to the compression of the train dwell time at stations and the compression of the train running time in sections. To further quantify the influencing factors and establish a mathematical model of delay recovery, it is necessary to filter and optimize these variables (with the exception of T_r), remove uncorrelated and redundant feature variables, and select the truly related simplified feature variables by differentiation. This process can not only reduce the running time of the model, but can also improve the accuracy of the model. There are many methods that can be used for feature selection, including the direct method, single-variable feature selection, and multivariable feature selection. Because there are many influencing factors of delay recovery, a stepwise regression algorithm was applied for multivariable feature selection to select the feature variables that affect delay recovery.

Multivariate feature selection refers to the combination and simultaneous optimization of multiple variables. Stepwise regression is a multivariate regression method for variable selection [29]. In this method, the independent variables that have the most significant impacts on the dependent variables are introduced sequentially, and the significances of the original independent variables in the model are tested to eliminate any nonsignificant variables;

the resulting model includes all the significant independent variable features of the dependent variables.

The feature selection of five variables (T_d, T_s, T_i, T, and L) related to T_r in the delay recovery data of the operation actions of the compression of the train dwell time at stations and the compression of the train running time in sections was carried out. First, the five variables T_d, T_s, T_i, T, and L were, respectively, used to establish regression models for T_r. The F-statistics were then calculated, and the variable with the most significant F-statistic (i.e., the largest F-statistic and $p < 0.01$) was selected and input into the regression model. Then, the most important variable of the remaining four variables was selected and input into the regression model, and the significance (p value) of the two variables after entering them into the model was tested. If p was not less than 0.01, the second step was repeated to judge all variables. After iterative calculation, the characteristic variables that affect delay recovery were finally determined. The statistical analysis results of the regression models are summarized in Table 5.

Taking the variable selection for the compression of the dwell time as an example, among the five F-statistics determined in Step 1, the F-statistic corresponding to T_s was the largest, and $p < 0.01$. Therefore, T_s was selected to be entered into the regression model. Then, in Step 2, among the remaining four variables (T_d, T_i, T, and L), the variable L was selected for inclusion in the regression model because it had the highest F-statistic. In this way, the variables were obtained sequentially by the 5 steps.

It must be noted that, in the regression model obtained in Step 5, the p-values of T_d, T_s, L, T_i, and T were, respectively, 2.12×10^{-14}, $< 2.2 \times 10^{-16}$, 8.01×10^{-8}, 1.18×10^{-7}, and 0.0777. Among them, the p-value of T was greater than 0.01, so the variable T was not used in the model, and the variable selection process returned to Step 4. Finally, for the delay recovery data of both the compression of the train dwell time at stations and the compression of the train running time in sections, it was ultimately determined that the characteristic variables that affect delay recovery are $T_r \sim T_s + L + T_d + T_i$.

5 Delay recovery model based on a gradient-boosted regression tree

The GBRT algorithm is a machine learning regression algorithm based on the gradient boosting method. This algorithm is an integrated model by use of the gradient descent algorithm to obtain a lifting tree, the goal of which is to minimize the mean squared error [30–32]. In the GBRT algorithm, subsequent decision trees are formed by learning the conclusions and residuals of the previous decision tree. By adjusting the weight of the weak learner, the accuracy of the regression prediction is gradually improved. Advantages of the GBRT algorithm include a high model accuracy, the ability to handle nonlinear data, and strong robustness to outliers.

During the process of carrying out the GBRT algorithm, the loss function, number of iterations, learning rate, resampling ratio, and depth of the decision tree are the most important model parameters. Based on experience, the value of the learning rate shrinkage is generally

Table 5 Statistical analysis results of the regression models

	Regression model	Compression of the train dwell time		Compression of the train running time	
		F-statistics	Significance (p value)	F-statistics	Significance (p value)
STEP 1	$T_r \sim T_d$	51.39	9.104×10^{-13}	1135	$< 2.2 \times 10^{-16}$
	$T_r \sim T_s$	1928	$< 2.2 \times 10^{-16}$	6587	$< 2.2 \times 10^{-16}$
	$T_r \sim T_i$	133.5	$< 2.2 \times 10^{-16}$	34.13	5.188×10^{-9}
	$T_r \sim T$	1.29	0.256	908.8	$< 2.2 \times 10^{-16}$
	$T_r \sim L$	145.4	$< 2.2 \times 10^{-16}$	1263	$< 2.2 \times 10^{-16}$
STEP 2	$T_r \sim T_s + T_d$	992.6	$< 2.2 \times 10^{-16}$	6586	$< 2.2 \times 10^{-16}$
	$T_r \sim T_s + T_i$	997.9	$< 2.2 \times 10^{-16}$	7701	$< 2.2 \times 10^{-16}$
	$T_r \sim T_s + T$	974.7	$< 2.2 \times 10^{-16}$	6330	$< 2.2 \times 10^{-16}$
	$T_r \sim T_s + L$	1004	$< 2.2 \times 10^{-16}$	6344	$< 2.2 \times 10^{-16}$
STEP 3	$T_r \sim T_s + L + T_d$	696.6	$< 2.2 \times 10^{-16}$	5300	$< 2.2 \times 10^{-16}$
	$T_r \sim T_s + L + T_i$	680.1	$< 2.2 \times 10^{-16}$	5162	$< 2.2 \times 10^{-16}$
	$T_r \sim T_s + L + T$	670.7	$< 2.2 \times 10^{-16}$	5125	$< 2.2 \times 10^{-16}$
STEP 4	$T_r \sim T_s + L + T_d + T_i$	533	$< 2.2 \times 10^{-16}$	4000	$< 2.2 \times 10^{-16}$
	$T_r \sim T_s + L + T_d + T$	523.2	$< 2.2 \times 10^{-16}$	3978	$< 2.2 \times 10^{-16}$
STEP 5	$T_r \sim T_s + L + T_d + T_i + T$	427.3	$< 2.2 \times 10^{-16}$	3200	$< 2.2 \times 10^{-16}$

between 0.001 and 0.01, and the resampling ratio is preferably 50%. In this study, during the process of establishing the regression models, the squared error is $\sum_{i=1}^{n}(\overline{y}_i - y_i)^2$, where indexes over some training set of the size of actual values of the output variable $\overline{y}_i$ is the predicted value, and y_i is the observed value.

The delay recovery data under the train operation adjustment actions of the compression of the train dwell time at stations and the compression of the train running time in sections were sorted according to time. The first 75% of the data was used as the training set, and the remaining 25% of the data was used as the test set. Based on the feature variables selected in Sect. 3, a regression model of T_d, T_s, T_i, and L against T_r was established using the GBRT algorithm.

5.1 Preferences

Before a model is established, the parameter values need to be determined. In this work, the tenfold cross-validation method was used to calculate the number of iterations, and the sum of the squared error loss was used as the evaluation criterion [33]. The optimal model iterations of compressed train dwell time and running time are determined as shown in Fig. 2. For these models, the sum of the squared error loss became stable when the numbers of iterations were, respectively, 2622 and 7389, as determined by the tenfold cross-validation method. Therefore, the optimal numbers of iterations for the delay recovery models in cases of compressing the train dwell time at stations and compressing the train running time in sections were determined to be 2622 and 7389, respectively.

Next, the depths of decision trees were determined, and the root-mean-square errors of the cross-validation were compared at different decision tree depths. It was found that the root-mean-square error was the least when the decision tree depth was 5, as presented in Fig. 3.

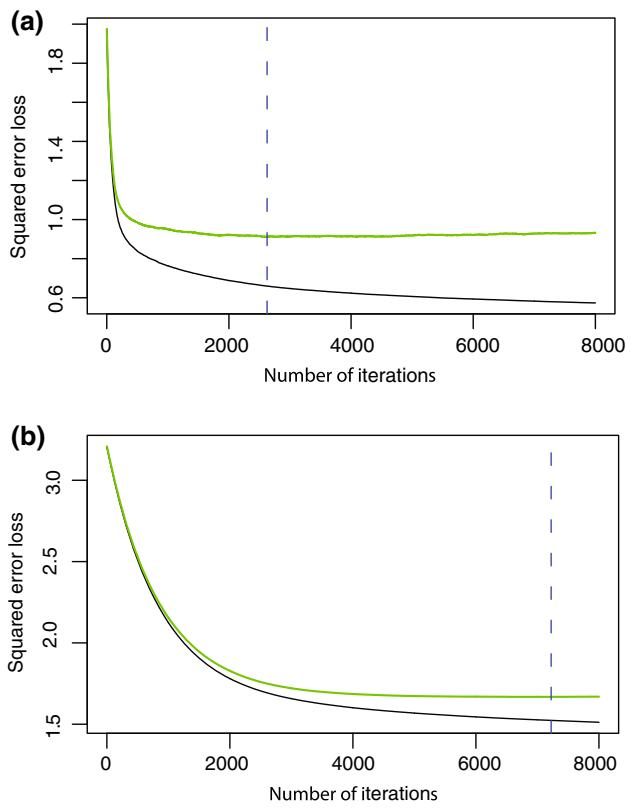

Fig. 2 Analysis of the optimal number of model iterations: **a** compression of the train dwell time and **b** compression of the train running time. The black curves and green curves represent the sum of the squared error loss and the squared error loss, respectively, and the vertical blue dotted lines represent the optimal numbers of iterations of the respective regression models

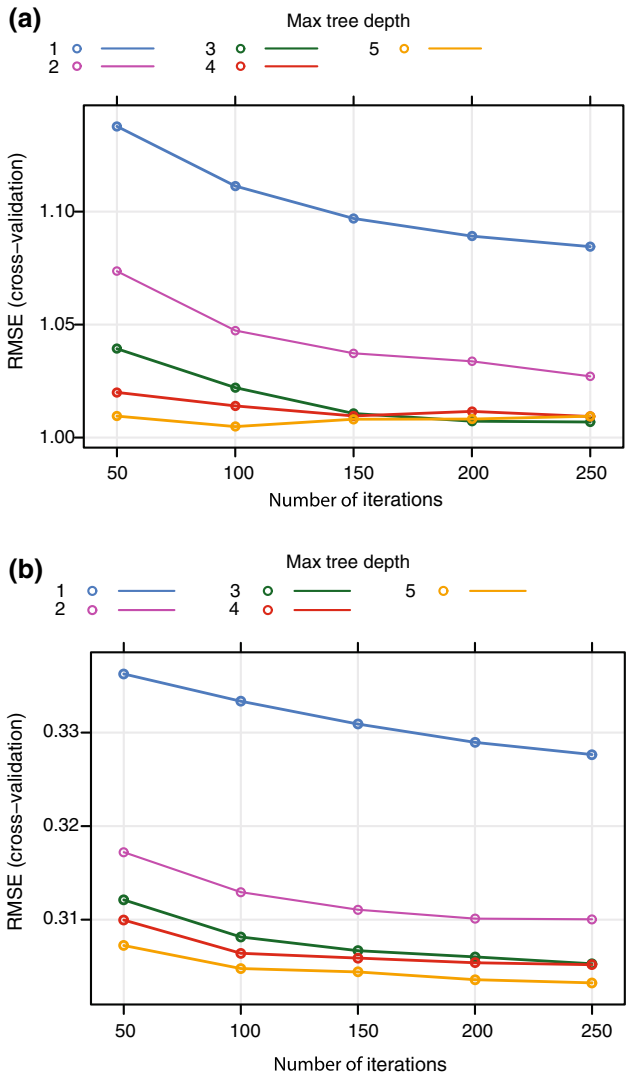

Fig. 3 Analysis of the decision tree depth: **a** compression of the train dwell time and **b** compression of the train running time

5.2 Modeling

After the model parameters were determined, the training data set was separately used for training and modeling. Due to the large amount of delay recovery data for the train intervals, the learning rate shrinkage was set as 0.05 to reduce the model running time while ensuring the accuracy of the model. In the delay recovery model of the compression of the train dwell time at stations, the learning rate shrinkage was set to 0.01.

The sum of the squared residuals was then calculated for the two training models; the results of the model with compression of the train dwell time at stations and compression of the train running time in sections were 0.926 and 0.290, respectively. It is normal for the values to differ so greatly because the two models are independent of each other. As delay recovery occurs more frequently and more supplement time can be used during train operation in sections than when the trains stop at stations, the value of the sum of squared error of the delay recovery model with compression of the train running time was much larger than that of the model with compression of the train dwell time.

The importance of each variable in the two regression models was calculated, and the results are presented in Fig. 4. It was found that, for the delay recovery models with both the compression of the train dwell time at stations and the compression of the train running time in sections, the supplement time has the highest importance, the delay time has the second-highest importance, and the interval between trains has the lowest importance.

According to this result, to recover train delays, regardless of which action is adopted, increasing the scheduled supplement time is the most effective approach, while changing the headway between trains has little effect.

Finally, the training models were used to predict the test data set, and the results are presented in Fig. 5.

The curves for the true and predicted values exhibited in Fig. 5 have high degrees of coincidence and small deviations; thus, the prediction models are considered to have performed well.

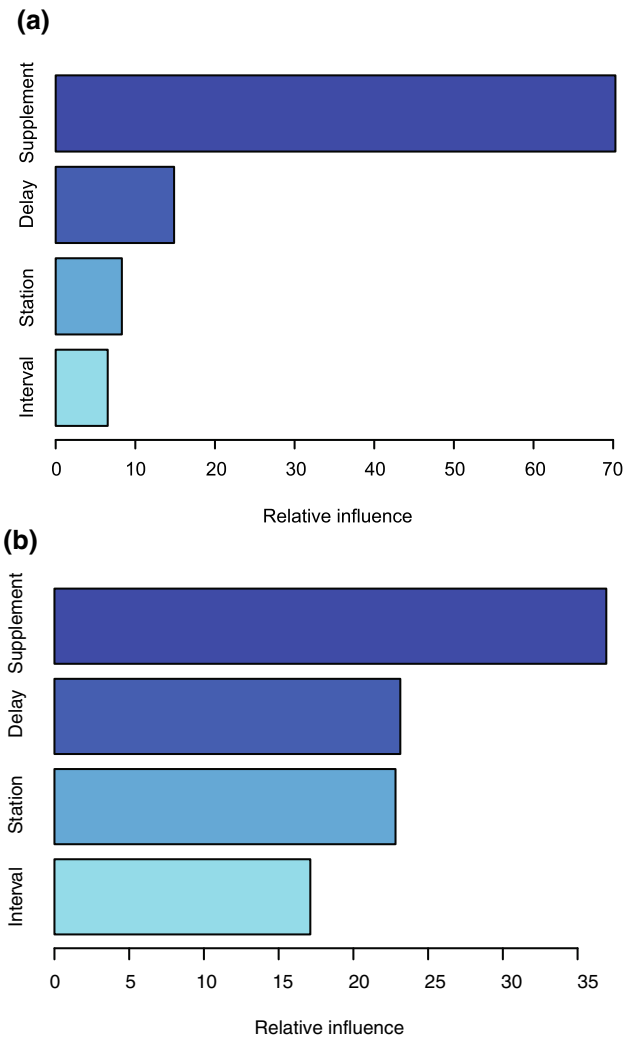

Fig. 4 Relative importance of variables: **a** compression of the train dwell time and **b** compression of the train running time

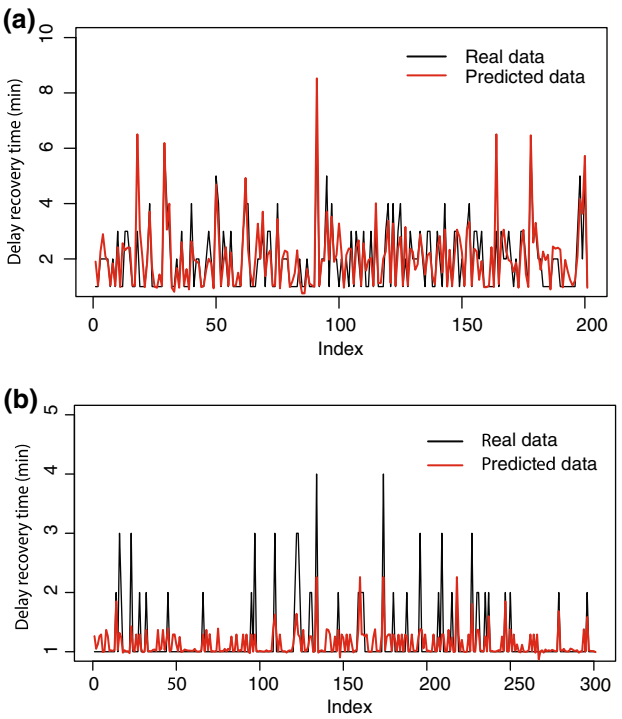

Fig. 5 Comparison of the true and predicted values for the test set: **a** compression of the train dwell time and **b** compression of the train running time

5.3 Model analysis

To further explore whether GBRT is the optimal regression model for delay recovery, a random forest (RF) model was used. The RF is a widely used and well-behaved integrated model for data prediction and has been confirmed to perform better than multiple linear regression and support vector machines when modeling delay recovery [22]. Generally, the distribution of residuals for the true and predicted values is used to analyze the effectiveness of the prediction. In the distribution figure, if most of the area under the curve is concentrated near zero, it means that the prediction effect is good; otherwise, it means that the prediction effect is not ideal. The residual distributions in Fig. 6 reveal that the areas under the residual curves of both the GBTR model and RF model are concentrated near zero, and the GBRT model has smaller residuals than the RF model, especially when the residuals are less than 2.5 for the action of compression of the train dwell time. This means that the GBRT model has achieved a better prediction accuracy than the RF model to a certain extent. In addition to its superior prediction accuracy, the advantages of the GBRT method also lie in the speed of parallel operation and its ability to process large amounts of data [30].

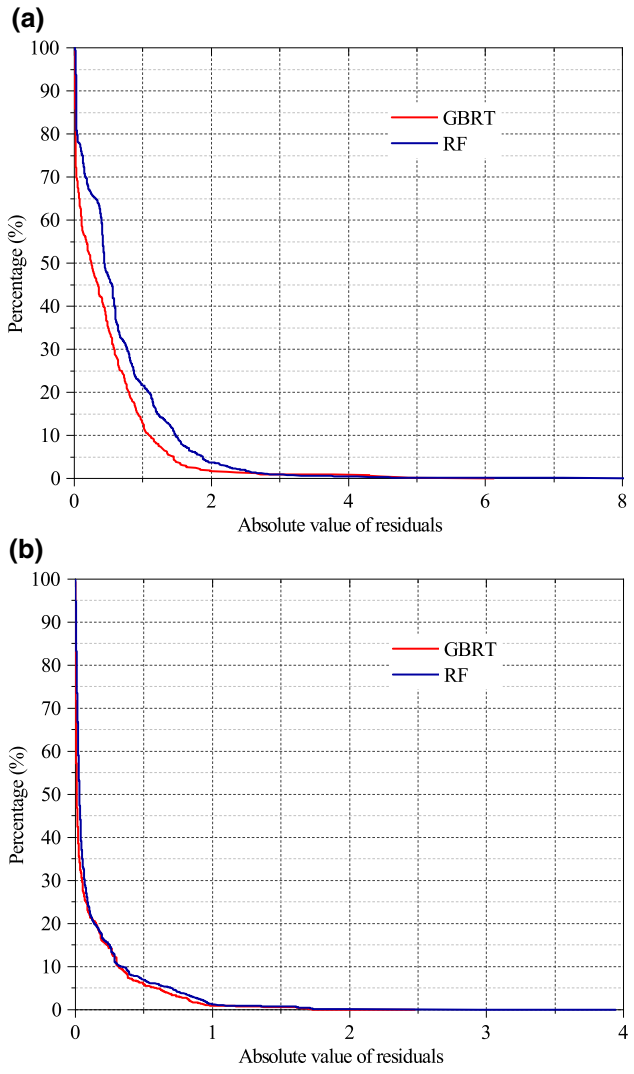

Fig. 6 Distribution of residuals: **a** compression of the train dwell time and **b** compression of the train running time

6 Conclusions

Modeling research on the effects of delay recovery for high-speed railways under different train operation adjustment actions can provide decision-making assistance for dispatchers in the selection of reasonable scheduling actions; it also can provide theoretical support for the development of intelligent scheduling for high-speed railways.

In this study, the variables in train operation data that may affect the delay recovery under different dispatching actions were first determined to be the delay recovery time (T_r), the delay time (T_d), the scheduled supplement time (T_s), the running interval (T_i), the occurrence time (T) and the place where the delay occurred (L). The Pearson correlation coefficients for the six variables were then calculated, and it was found that correlations existed between the six variables. To establish delay recovery models under different train operation adjustment actions and elucidate the delay recovery mechanism, the stepwise regression method was used to select multiple variables among the five identified influencing variables (T_d, T_s, T_i, T, and L). To demonstrate the superiority of the GBRT algorithm, an RF model was selected for comparison. From the results, the following conclusions can be drawn.

1. In the delay recovery models under different train operation adjustment actions, four explanatory variables, namely the scheduled supplement time, the delay time, the occurrence place, and the running interval, were found to have the same order of importance in cases of compressing the train dwell time at stations and compressing the train running time in sections, i.e., $T_s > T_d > L > T_i$.
2. The sums of the squared residuals of the GBRT regression prediction models for both the compression of the train dwell times and the compression of the train running time were both very small, indicating that the prediction performances of the models were good. In other words, the GBRT prediction model can effectively explain the delay recovery mechanism.

However, as delay recovery occurs more frequently and more supplement time can be used during train operation in sections than when trains stop at stations, the value of the sum of squared errors of the compressed train running time model was much larger than that of the compressed train dwell time model.

3. By comparing the prediction abilities of the GBRT prediction model and RF model, the GBRT prediction model was found to have higher prediction accuracy, and is characterized by a fast speed of parallel operation and the ability to process large amounts of data.

In the future, how to utilize the proposed model in the automation of dispatching knowledge should be studied in depth to assist dispatchers to make decisions easily and precisely and support the development of intelligent train dispatching systems. A goal of future research is to establish a database based on these models, which should include knowledge of delay recovery that accounts for different delay recovery actions to indicate which actions may be adopted according to the situations of the delayed trains. For example, once there is a delay, and once the related variables of the train can be retrieved, the possible delay recovery actions can be generated automatically to facilitate the decision-making of dispatchers.

Acknowledgements This work was supported by the National Nature Science Foundation of China (Nos. 71871188 and U1834209), the Science and Technology Department of Sichuan Province (No. 2018JY0567). We are grateful for the contributions made by our project partners. We also thanks Dr. Javad Lessan for his meaningful suggestions for the revision of the manuscript.

References

1. Mattsson L-G (2007) Railway capacity and train delay relationships. Springer, Berlin
2. Keiji K, Naohiko H, Shigeru M (2015) Simulation analysis of train operation to recover knock-on delay under high-frequency intervals. Case Stud Transp Policy 3(1):92–98
3. Kariyazaki K, Hibino N, Morichi S (2013) Simulation model for estimating train operation to recover knock-on delay earlier. Asian Transp Stud 2:284–294
4. Krasemann JT (2012) Design of an effective algorithm for fast response to the re-scheduling of railway traffic during disturbances. Transp Res Part C: Emerg Technol 20(1):62–78
5. Yuan J, Hansen IA (2007) Optimizing capacity utilization of stations by estimating knock-on train delays. Transp Res Part B Methodol 41(2):202–217
6. Cadarso L, Marín Á, Maróti G (2013) Recovery of disruptions in rapid transit networks. Transp Res Part E: Logist Transp Rev 53:15–33
7. D'Ariano A, Pranzo M, Hansen IA (2007) Conflict resolution and train speed coordination for solving real-time timetable perturbations. IEEE Trans Intell Transp Syst 8(2):208–222
8. Dollevoet T, Huisman D, Kroon L, Schmidt M, Schöbel A (2014) Delay management including capacities of stations. Transp Sci 49(2):185–203
9. Cheng Y (1998) Hybrid simulation for resolving resource conflicts in train traffic rescheduling. Comput Ind 35(3):233–246
10. Roberts C, Easton JM, Kumar AVS, Kohli S (2017) Innovative applications of big data in the railway industry. IGI Global, Pennsylvania
11. De Fabris S, Longo G, Medeossi G (2010) Automated analysis of train event recorder data to improve micro-simulation models. In: Timetable planning and information quality, pp 125–134 10.2495/978-1-84564-500-7/1.
12. Bendfeldt J, Mohr U, Muller L (2000) RailSys: a system to plan future railway needs. WIT Trans Built Environ 50:249–255
13. Yamamura A, Koresawa M, Adachi S, Tomii N (2014) Taking effective delay reduction measures and using delay elements as indices for Tokyo's metropolitan railways. WIT Trans Built Environ 135:3–15
14. Naohiko H, Osamu N, Shigeru M, Hitoshi I, Norio T (2017) Recovery measure of disruption in train operation in Tokyo Metropolitan Area. Transp Res Procedia 25:4374–4384
15. Liebchen C, Lübbecke M, Möhring R, Stiller S (2009) The concept of recoverable robustness, linear programming recovery, and railway applications. In: Robust and online large-scale optimization. Springer, Berlin, pp 1–27
16. Kecman P, Goverde RMP (2015) Online data-driven adaptive prediction of train event times. IEEE Trans Intell Transp Syst 16(1):465–474
17. Kecman P, Goverde RM (2015b) Predictive modelling of running and dwell times in railway traffic. Public Transp 7(3):295–319
18. Khadilkar H (2016) Data-enabled stochastic modelling for evaluating schedule robustness of railway networks. Transp Sci 51(4):1161–1176
19. Guo J, Meng L, Kecman P, Corman F (2015) Modeling delay relations based on mining historical train monitoring data: a Chinese railway case. In: Proceedings of the 6th international seminar on railway operations modeling and analysis (RailTokyo 2015), March 23–26, Chiba Institute of Technology, Tokyo, Japan
20. Huang P, Wen C, Yang Y, Jiang C, Chen Y, Li J (2017) Delay propagation mechanism of high-speed railway. In: 96th TRB annual meeting. Transportation Research Board, Washington DC
21. Wen C, Lessan J, Fu L, Huang P, Jiang C (2017) Data-driven models for predicting delay recovery in high-speed rail. In: 4th international conference on transportation information and safety (ICTIS), IEEE, pp 144–151
22. Jiang C, Huang P, Lessan J, Fu L, Wen C (2019) Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression. Can J Civ Eng 46(5):353–363

23. Yuan J, Hansen IA (2008) Closed form expressions of optimal buffer times between scheduled trains at railway bottlenecks. In: 11th international IEEE conference on intelligent transportation systems, IEEE, pp 675–680
24. Safapour E, Kermanshachi S, Alfasi B, Akhavian R (2019) Identification of schedule performance indicators and delay recovery strategies for low-cost housing projects. Sustainability 11(21):6005
25. Alkhonaini M, El-Sayed H (2018) Minimizing delay recovery in migrating data between physical server and cloud computing using reed-solomon code. In: 20th international conference on high performance computing and communications (HPCC), Exeter, United Kingdom, pp 718–724
26. Yasufumi O, Yoshiki M, Norio T (2018) Analysis of delay recovery operation of railway drivers using decision trees. IEEJ Trans Ind Appl 138(11):877–883 **(in Japanese)**
27. Palmqvist C-W, Olsson NOE, Winslott-Hiselius L (2017) An empirical study of timetable strategies and their effects on punctuality. In: 7th international conference on railway operations modeling and analysis, Lille, France, 5–7 April, 2017
28. Goverde R, Hansen I (2000) TNV-prepare: analysis of Dutch railway operations based on train detection data. Comput Railw 7:779–788
29. Pope P, Webster J (1972) The use of an F-statistic in stepwise regression procedures. Technometrics 14(2):327–340
30. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77(4):802–813
31. Zhang H, Yang Q, Shao J, Wang G (2019) Dynamic streamflow simulation via online gradient-boosted regression tree. J Hydrol Eng 24(10):04019041
32. Ivatt PD, Evans MJ (2020) Improving the prediction of an atmospheric chemistry transport model using gradient-boosted regression trees. Atmos Chem Phys 20(13):8063–8082
33. Fushiki T (2011) Estimation of prediction error by using K-fold cross-validation. Stat Comput 21(2):137–146