Modelling the Influences of Primary Delays Based on High-speed Train Operation Records

ZHONGCAN LI [a], PING HUANG [a,b], CHAO WEN [a,b,1], YIXIONG TANG [a]

[a] School of Transportation and Logistics, Southwest Jiaotong University
Nr.111, North 1st Section of Second Ring Road, 610031, Chengdu, China
[b] High-speed Railway Research Center, University of Waterloo
Waterloo, N2L3G1, Canada
[1] E-mail: c9wen@uwaterloo.ca, Phone: 1-2269788096

Abstract
Primary delays (PDs) are the driving force of delay propagation. Hence, accurate predictions of the number of affected trains (NATs) and the total time of affected trains (TTATs) due to PDs can provide a theoretical background for the dispatch of trains in real time. Train operation data were obtained from Wuhan-Guangzhou High-Speed Railway (HSR) station from 2015 to 2016, and the NAT and TTAT influence factors were determined after analyzing the PD propagation mechanism. The NAT predictive model was established using eXtreme Gradient Boosting (XGBOOST) algorithm which was more efficient than other machine learning methods after comparison. Furthermore, the TTAT predictive model was established based on the NAT model using the support vector regression (SVR) algorithm. The results indicate that the XGBOOST algorithm has good performance on the NAT predictive model, whereas SVR is the best method for the TTAT model using Lessthan5 variable, which is the ratio of the difference between the sample size of actual and the predicted values in less than 5 min and the total sample size. In addition, 2018 data were used to evaluate the application of NAT and TTAT models over time. The results indicate that NAT and TTAT models have a good application over time.

Keywords:
High-speed railway, Primary delay, Number of affected trains, Total time of affected trains, Machine learning

1. INTRODUCTION

High-speed railway (HSR) transportation is becoming more popular than other modes of transportation worldwide owing to their high speed, safety, and density. In China, HSR trains have become one of the major means of transportation. High punctuality of these trains is an important factor considered by railway companies in attracting passengers (Yuan et al., 2002). However, they are influenced by bad weather, mechanical failure of the systems, and organization strategies during operation, which could lead to delays. These delays disrupt railway operation and transportation, increase travel time of passengers, and reduce the passenger travel experience, thereby making HSR trains less reliable.

Delays are categorized as primary delays (PDs) and secondary (knock-on) delays. PDs are the driving force of delay propagation. They occur when some uncertain events directly disrupt the train operations. However, secondary delays are attributed to the delay propagation caused by PDs. When a PD occurs, the operation adjustment mainly depends on the experience of train dispatchers. However, there are no scientific theories and methods that support the strategies used. Meanwhile, the number of affected trains (NATs) and the total time of affected trains (TTATs) due to a PD can be used to estimate the influence of PD and accurately determine the severity of the delay. Therefore, NAT and TTAT predictive models can assist the train dispatcher in estimating the train operation state, provide the theoretical basis for the rescheduling strategy, facilitate more scientific and reliable rescheduling decisions and adjustment based on the station work plan (Wen et al., 2018). Furthermore, NAT and TTAT predictive models are vital in the automatic operation of trains and the intelligent dispatch of HSRs.

The impact of the NAT and TTAT predictive models on PD propagation is determined in this study. The models were built based on the data obtained from Wuhan-Guangzhou HSR station (Guangzhou Railway Bureau, China) from March 2015 to November 2016, and evaluated using common machine learning classification and regression algorithms. The results indicated that eXtreme Gradient Boosting (XGBOOST) and support vector regression (SVR) algorithms had the best predictive results for NAT and TTAT models, respectively. Furthermore, the models were evaluated using 2018 data in order to test their effectiveness over time. The results show that the models have good predictive abilities and can be used for a long time.

This paper is structured as follows: Section 1 introduces the background and significance of the research. Section 2 reviews some studies conducted on delay propagation while Section 3 presents the problem to be solved and also describes the data used. The NAT and TTAT predictive models are established and tested in Section 4 while the conclusions are discussed in Section 5.

2. LITERATURE REVIEW

PDs may be caused by exogenous events such as irregularities in the natural environment or vehicle faults, accidents, facility failures, etc., in internal systems (Goverde, 2005). The severity of the delay is measured using a delay probability distribution model when the delay distribution corresponds to an exponential distribution and secondary delays are induced in different traffic scenarios (Huisman and Boucherie, 2001). (Meester and Muns, 2007) obtained the knock-on delay distribution from PD distributions using a phase-type distribution. However, (Goverde et al., 2013) found that Weibull distributions can be fitted on the PD distribution using empirical data. Meanwhile, (Wen et al., 2017) indicated that PD distributions could be well approximated by log-normal distributions while line regression models can be used to approximate NAT distributions. However, studies on predictive models of delay propagations are mostly based on mathematical optimization methods. (Huisman et al., 2002) and (Milinković et al., 2013) estimated train delays using Queuing and Petri net models, respectively. Meanwhile, (Hansen et al., 2010) proposed an online model for the prediction of running time and arrival time using timed event graphs. In addition, (Kecman and Goverde, 2015) proposed a timed event graph approach for the accurate prediction of train event times using dynamic arc weights model. Furthermore, (Goverde, 2007) established a delay propagation model using the max-plus algebra theory.

Data-driven studies are increasingly used in delay/disruption management. (Goverde,

2005) studied the systematic delay propagation in trains and employed a robust linear regression model to investigate the correlation among arrival delays using data obtained from Eindhoven Railway Station, Netherlands. Meanwhile, (Kecman et al., 2015) discussed the dynamics of train delays over time and space, and modeled the uncertainty of train delays based on a Markov stochastic process. (Şahin, 2017) also described the train operation process as a Markov chain and concluded that the train states at certain event timesteps could be determined by transition probability matrices. Furthermore, (Corman and Kecman, 2018) proposed an online Bayesian network to predict train delay over time using historical data in Sweden, while (Lessan et al., 2018) established a hybrid Bayesian network to estimate train arrival and departure delays based on real data in China. Artificial neural networks (ANNs) have been widely used to predict the delays in passenger trains (Chapuis, 2017; Pongnumkul et al., 2014; Yaghini et al., 2013). However, (Marković et al., 2015) indicated that SVR is more accurate for predicting train arrival delays in comparison with ANN algorithms based on Serbian Railways data. Meanwhile, (Tang et al., 2018) discovered the relationship between the causes of PD and the duration based on NAT and TTAT models using SVR. However, the NAT is unknown when a PD occurs that will lead to the model cannot predict online.

3. PROBLEM STATEMENT AND DATA DESCRIPTION

3.1 Problem statement

The headway between two trains in a station comprises the minimum interval time and the timetable supplement time. If a train is delayed before it arrives the station while the preceding train is not delayed, the delayed train is considered as a PD train. In other words, a delayed train is regarded as a PD train if a minimum threshold (e.g., 5 min in Wuhan-Guangzhou HSR station) exists between the arrival time (or scheduled arrival time) of the delayed train and the actual arrival time of the preceding train. The PD train greatly influences the motion of the subsequent train, thereby leading to the secondary delay. This process occurs for all successive trains. However, the PD train has less influence on the subsequent trains when the delay duration is less than 5 min such that the rescheduling of the trains is not necessary. Hence, only PD durations of more than 5 min are considered in this paper. Meanwhile, the delays are reduced by timetable supplement time until they are eliminated. Hence, there is a sequence in the PD influence where the number of PD and knock-on trains is classified by NAT and TTAT, which is the sum of the PD and knock-on delay time.

Figure 1 shows the process of PD propagation at two stations (Station A and Station B) in Wuhan-Guangzhou HSR station. The red and black lines are actual train lines and scheduled train lines, respectively. A minimum time interval exists between Train 1 and the preceding train, such that Train 1 is a PD train having a delay duration of t_1. Meanwhile, Train 2 is delayed as the interval between the actual arrival time of Train 1 and the scheduled arrival time of Train 2 is less than 5 min, thereby leading to a delay in Trains 3 and 4. The PD stops at Train 4 due to the supplement time t_{sup}^{i} such that Train 5 returns to normal operation. The delayed trains (Trains 1–4) form a PD sequence where NAT is 4 and TTAT is $\sum_{i=1}^{4} t_i$.

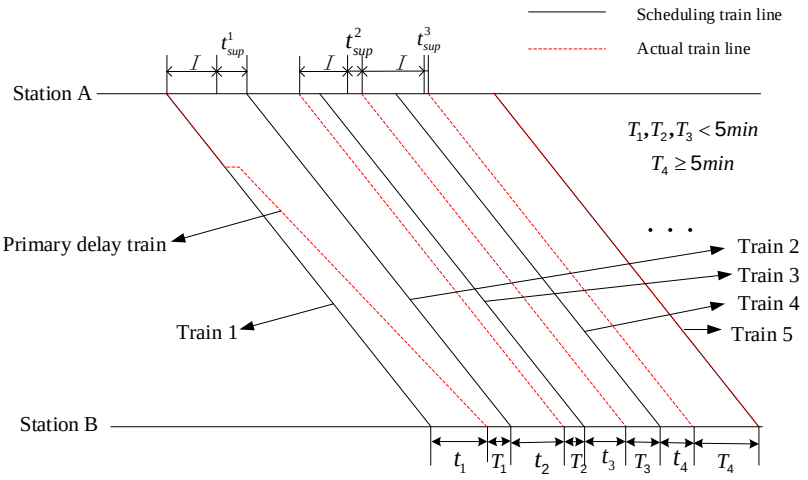

Figure 1: PD propagation process at two stations

The trains overtake one another if the actual arrival sequence is different from the scheduled arrival sequence. This sequence is due to rescheduling. However, the propagation process is complicated as many influence factors need to be considered. Hence, these sequences are not considered in this paper.

3.2 Data description

The data used in this study were obtained from the station operation records of Wuhan-Guangzhou HSR station (Guangzhou Railway Bureau, China) for Guangzhou North (GZN), Qingyuan (QY), Yingde West (YDW), Shaoguan (SG), Lechang East (LCE), Chenzhou West (CZW), Leiyang West (LYW), Hengyang East (HYE), Heangshan West (HSW), Zhuzhou West (ZZW), and Changsha South (CSS). Table 1 summarizes a portion of the data.

Table 1: Raw data from Guangzhou Station

Train NO	Date	Station	Scheduled arrive time	Scheduled departure time	Actual arrive time	Actual departure time
G280	2015/3/24	GuangzhouNorth	7:00:00	7:00:00	7:01:00	7:01:00
G636	2015/3/24	GuangzhouNorth	7:07:00	7:07:00	7:07:00	7:07:00
G1102	2015/3/24	GuangzhouNorth	7:13:00	7:13:00	7:14:00	7:14:00
G6102	2015/3/24	GuangzhouNorth	7:20:00	7:20:00	7:20:00	7:20:00

The primary influence predictive model was established by preprocessing the data in a series of steps summarized as follows:

- Step 1: Gather the data from the database and eliminate abnormal entries such as duplicate entries, errors and invalid entries.
- Step 2: Sort the data by actual arrival time in the station.
- Step 3: Select the PD train and obtain the train sequences which do not overtake based on PD influence.

- Step 4: Extract the features of the influence factors and calculate NAT and TTAT based on the PD influence sequences.

Thus, the feature sets of the influence factors of NAT and TTAT were obtained by analyzing the mechanism of the PD propagation. These influence factors are described as follows:

D: Primary delay duration of PD,

I: Scheduled interval between the PD train and the subsequent adjacent train,

B: 0-1 variable, which is 0 when the PD train does not stop at the station and 1 otherwise,

T: Period of a PD occurrence, and classify the period by hour

N: The number of affected trains if supplement times are fully utilized.

Table 2 summarizes a sample data after pre-processing:

Table 2: A sample of modeling data

D	***I***	***B***	***T***	***N***	**NAT**	**TTAT**
5	6	0	8:00-9:00	2	2	9
6	7	0	16:00-17:00	3	3	12
5	8	1	8:00-9:00	3	2	7
6	6	0	9:00-10:00	3	5	28
6	7	0	17:00-18:00	2	2	11

In this study, *D* presents the primary delay train delay duration; *I* record the scheduled headway between the PD train and the first train subsequently; *B* is a 0-1 variable, and it equals to 0 when the PD train does not stop at the station. Otherwise it equals to 1; Classify the period by hour and marked *T* as the period of PD occurs. *N* indicates the number of affected trains when the supplement times were fully utilized. All the factors above are obtained when PD occurs based on a real-time timetable. Hence, real-time rescheduling is possible if NAT and TTAT predictive models are investigated using these factors.

The predictive models were established using the data obtained from March 2015 to November 2016. Seventy percent of the data was used as the training data while 30% was used as the validation data for the model in order to prevent overfitting. Finally, the models were evaluated by using data obtained in 2018 as the test data.

4. PREDICTIVE MODEL OF NAT AND TTAT

4.1 The predictive model of NAT

Figure 2 shows the heatmap and 3D histogram of the intensity distribution of PD influence over time, which can assist train dispatchers in carrying out risk warnings. The PD duration and the period of PD occurrence for GZN station were plotted on the horizontal and vertical coordinates, respectively.

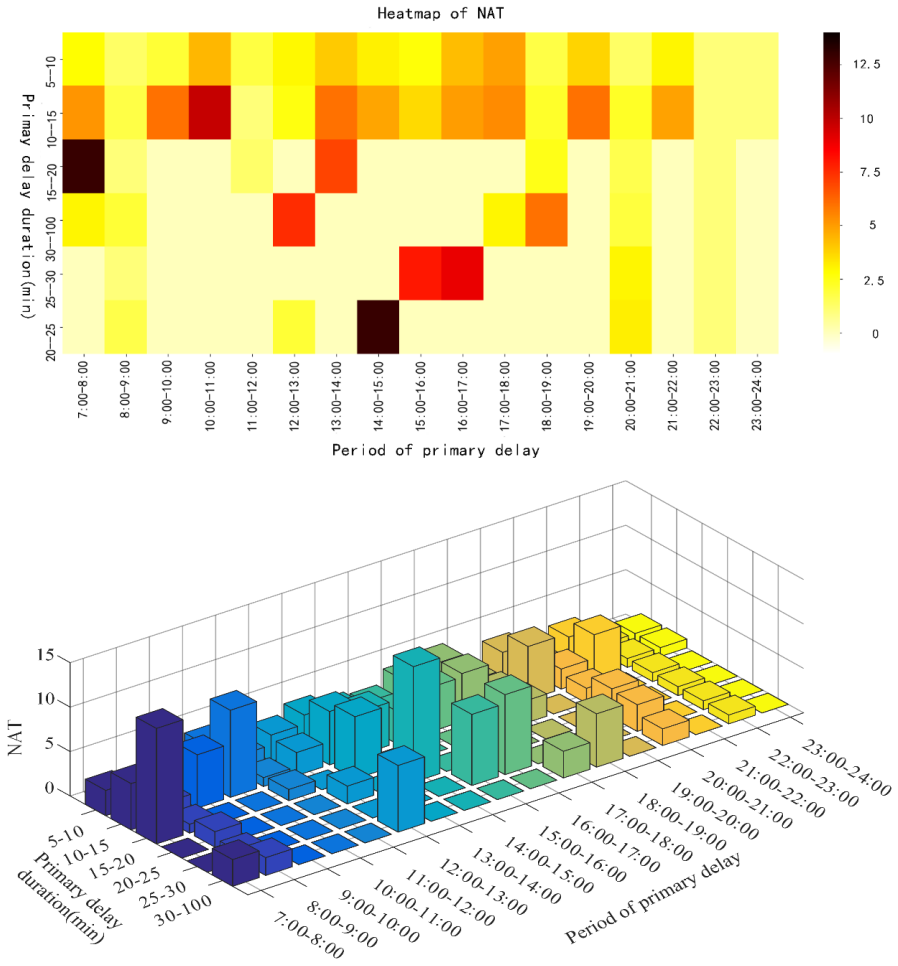

Figure 2: The NAT heatmap and 3D histogram of GZN station

The influence factors of NAT (G) are *D*, *B*, *I*, *T*, and *N*. NAT is a discrete random variable whose prediction is a classification problem. The output of the model is set to *S* while the feature set of the influence factors is the input such that the relationship between *S* and *G* is

$$S = \Phi(D, B, T, I, N) \tag{1}$$

where Φ is the classification algorithm. When NAT > 5, the sample size corresponding to each value is small, and the distribution is discrete. Thus, the NAT values that were greater than 5 were classified as 6 and more. Finally, NAT was divided into six categories (1 / 2 / 3 / 4 / 5 / 6 and more).

Meanwhile, XGBOOST was used as the classification algorithm. It is an improved algorithm based on gradient boosting decision tree which is highly efficient and flexible

and can be used for solving regression and classification problems. For a given dataset with *n* ensembles and *m* features, the result $\hat{y}_i$ is given by an ensemble represented by the model as follows:

$$D = \left\{ (x_i, y_i) : i = 1, 2, ...n, x_i \in R^m, y_i \in R \right\} \tag{2}$$

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{3}$$

$$F = \left\{ f(X) = w_{q(x)} \right\} \left(q : R^m \longrightarrow T, \quad w \in R^T \right) \tag{4}$$

where f_k is a regression tree (also known as CART), $f_k(x_i)$ represents the score given by the *k*-th tree to the *i*-th sample in the data, *q* represents the structure of each tree that maps an example to the corresponding leaf index, and *T* is the number of leaves in the tree. Each f_k corresponds to an independent tree structure *q* and leaf weight *w*.

Minimizing the regularized function to give the objective function：

$$\ell(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{5}$$

where *l* is the loss function and Ω is the penalty term to prevent overfitting and complexity of the model, given as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \tag{6}$$

where γ and λ control the penalty based on *T* and *w*, respectively.

Furthermore, an iterative method was used to minimize the objective function. The objective function which is minimized at *t*-th iteration is

$$\ell^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{7}$$

Using Taylor expansion, Eqn. (7) can be derived for loss reduction after the tree splits from the given node as

$$\ell_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{8}$$

where

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \tag{9}$$

where I is a subset of the available observations in the current node, and I_R and I_L are subsets of the available observations in the left and right node after the split, respectively. The best split can be found using Eqn. (8) at any given node, which is based on the regularization parameter (λ) and the loss function.

The detailed derivation is presented by (Chen and Guestrin, 2016).

To evaluate the predictive accuracy of the XGBOOST algorithm, other classification algorithms such as random forest (RF), support vector machine (SVM), Logistic Regression (LR) and K-nearest neighbor (K-NN) were used as the evaluation criteria. The optimal parameter value of each algorithm was calculated using hyperparametric search. Accuracy was then used as the standard measure to assess the predictive precision of the model, which is calculated as follows：

$$ACCURACY = \frac{N_c}{N_a}$$

where N_c : Sample size of correct classification, and

N_a : Total sample size.

The accuracy of each classification algorithm using validation data at different stations is shown in Table 3 and Figure 3.

Table 3: NAT predictive accuracy using different classification algorithms

	RF	**XGBOOST**	**SVM**	**LR**	**KNN**
GZN	0.7711	**0.7766***	0.7520	0.6676	0.7084
QY	0.7105	**0.8005***	0.6972	0.5642	0.7864
YDW	0.7200	**0.7200***	0.7200	0.6400	0.6933
SG	0.6453	**0.6816***	0.6065	0.5375	0.6271
LCE	0.7573	**0.7908***	0.7414	0.6837	0.7774
CZW	0.7239	**0.7692***	0.6916	0.6099	0.7658
LYW	0.7173	**0.7589***	0.6922	0.6182	0.7543
HYE	0.7544	**0.7424***	0.6393	0.5773	0.7246
HSW	0.7316	**0.7677***	0.6677	0.6098	0.7231
ZZW	0.6799	**0.7266***	0.6173	0.6072	0.7165
CSS	0.6805	**0.7427***	0.6473	0.6017	0.6390

* indicate the best predictive accuracy

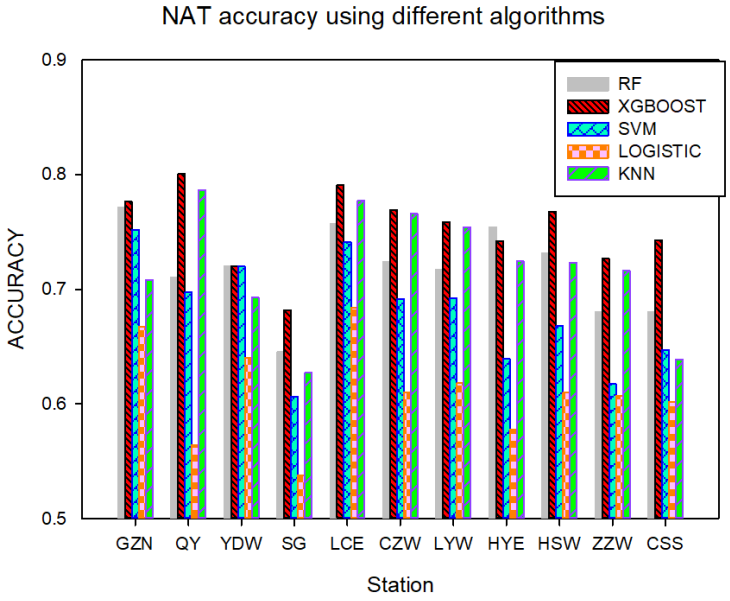

Figure 3: NAT predictive accuracy using different classification algorithms

The results show that (1) the XGBOOST algorithm has the highest accuracy at all stations in comparison with other algorithms; (2) the accuracy value of XGBOOST algorithm maintained high levels (up to 0.7) at all stations except at SG. This proves that the NAT predictive model based on the XGBOOST algorithm has good precision.

The timetable and infrastructure of the Wuhan-Guangzhou HSR station from 2015 to 2016 do not change significantly in comparison with 2018 data. Hence, the train operation data can be used as validation data to evaluate the precision of the model based on the data obtained from 2015 to 2016. Meanwhile, the data obtained from March to July 2018 were used as test data to evaluate the application of the model over time. The results of the predictive accuracy at different stations are shown in Figure 4.

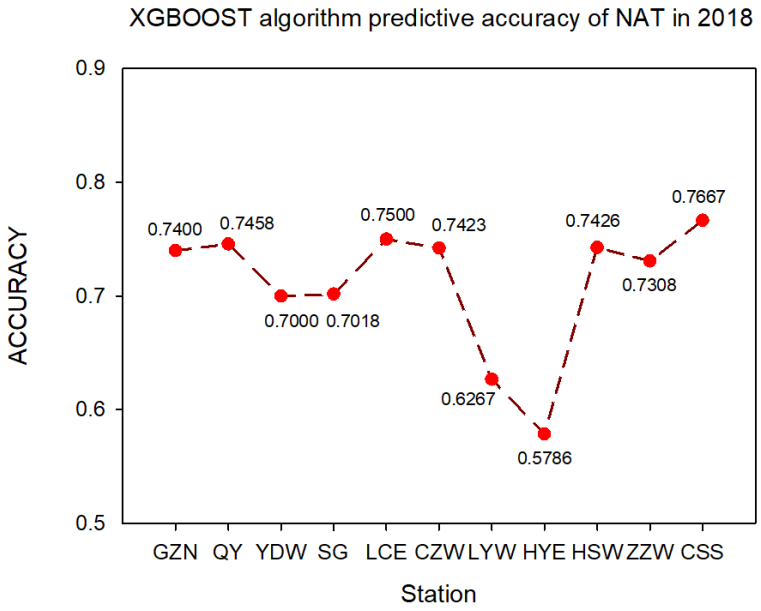

Figure 4: XGBOOST algorithm predictive accuracy of NAT in 2018

The model has a good precision and high accuracy (up to 0.7) in the stations at Wuhan-Guangzhou HSR station except for LYW and HYE. When this is combined with the accuracy values of the validation data, the results indicate that the model based on XGBOOST algorithm can accurately predict the number of affected trains by PD at Wuhan-Guangzhou HSR station.

4.2 The predictive model of TTAT

TTAT is another indicator that measures the severity of the PD influence. The overall scope of influence can be determined by combining TTAT and NAT results. The specific derivation process is described below:

Given a PD influence sequence, the TTAT and NAT are given as T_{td} and N_1, respectively, while the delay duration of i-th train is $T_{at}^{\ i}$. The discriminant relationship is obtained as follows:

IF $i=1$; THEN, the TTAT of the PD sequence is T_{td}, while NAT is N_1,

IF $1<i\le N_1$; THEN, the subsequent TTAT of the PD sequence is $T_{td}-\sum_{i=1}^{N_1} t_{at}^{\ i}$, while NAT is N_1-i.

The heatmap and 3D histogram of the TTAT for GZN station are shown in Figure 5.

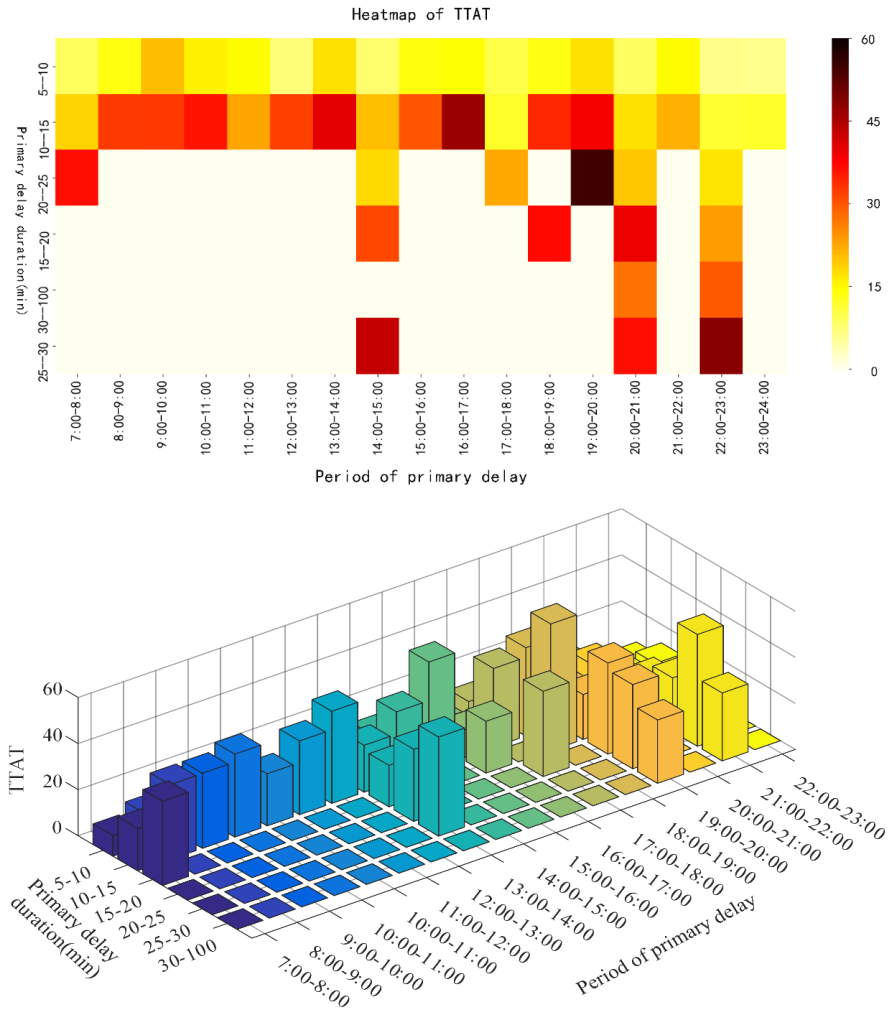

Figure 5: The TTAT heatmap and 3D histogram of GZN station

Because TTAT strongly depends on NAT, the predictive model is established based on the NAT model. Thus, the prediction is set as S' and Y for NAT and TTAT, respectively. Hence, TTAT predictive model is expressed as

$$Y = \varphi(D, B, T, I, N, S') \tag{10}$$

φ is a regression algorithm as TTAT is a continuous variable. The TTAT model was established using SVR, and compared with several algorithms such as RF, XGBOOST, Ridge regression (Ridge), and Lasso regression (LASSO)

Given a data set $D = \{(x_i, y_i) : i = 1, 2, ...n, x_i \in R^m, y_i \in R\}$, where x_i denotes the i

nput and y_i the output of the sample. The goal of SVR is to find a function $f(\mathbf{x})$ t hat has the most deviation (ε) from the actual and predicted values. $f(\mathbf{x})$ is defined as $f(\mathbf{x}) = w^{\mathrm{T}}\mathbf{x}+b$, where w is a hyperplane direction and b is an offset scalar.

The objective function is expressed as follows：

$$\begin{aligned} &\min_{w,b,\xi_i,\xi_i^*} = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right) \\ &\textbf{s.t.} \begin{cases} -\varepsilon - \xi_i^* \le f\left(\mathbf{x}_i\right) - y_i \le \varepsilon + \xi_i \\ \xi_i^*, \xi_i \ge 0, i = 1, 2, ..., m \end{cases} \end{aligned} \tag{11}$$

where *C* is a penalty factor which determines the trade-off between the flatness of *f* and the values to which deviations larger than ε are tolerated. The ε-insensitive loss function $|\xi|_\varepsilon$ is given as

$$|\xi|_\varepsilon := \begin{cases} 0, & if\ |\xi| \le \varepsilon; \\ |\xi| - \varepsilon, & otherwise. \end{cases} \tag{12}$$

Using Lagrange multipliers, Eqn. (11) can be expressed as

$$\begin{aligned} &L(w, b, \alpha, \alpha^*, \xi_i, \xi_i^*, u, u^*) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{m} u_i\xi_i - \sum_{i=1}^{m} u_i^*\xi_i^* \\ &+\sum_{i=1}^{m}\alpha_i\left(f(\mathbf{x}_i) - y_i - \varepsilon - \xi_i\right) + \sum_{i=1}^{m}\alpha_i^*\left(f(\mathbf{x}_i) - y_i - \varepsilon - \xi_i^*\right) \end{aligned} \tag{13}$$

The optimal solution can be obtained by solving Eqn. (13) to yield

$$\begin{aligned} &f(\mathbf{x}) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b \\ &b = y_i + \varepsilon - \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b \end{aligned} \tag{14}$$

The detailed derivation is presented by (Smola and Schölkopf, 2004). To evaluate the model, Lessthan5 variable was defined which is given as

$$Lessthan5 = \frac{N_d}{N_a}.$$

where N_d : The sample size of the absolute value of the difference between the actual and predicted values in less than 5 min.

N_a : Total sample size.

The optimal parameter value of each algorithm was calculated using hyperparametric search. The Lessthan5 value of each algorithm is shown in Table 4 and Figure 6.

Table 4: TTAT Lessthan5 value using different algorithms

	RF	XGBOOST	SVR	Ridge	LASSO
GZN	81.638	81.638	**85.311***	79.096	80.508
QY	77.526	77.526	**78.739***	76.395	77.850
YDW	70.000	70.000	**74.286***	70.000	71.429
SG	74.444	74.444	**76.173***	73.827	74.321
LCE	83.761	83.761	**84.444***	78.291	79.915
CZW	76.590	76.590	**77.009***	72.676	72.467
LYW	76.410	76.410	**77.098***	72.765	73.040
HYE	73.829	73.829	**74.582***	72.324	71.739
HSW	74.917	74.917	**75.116***	69.927	69.661
ZZW	76.362	76.362	**76.510***	72.680	71.355
CSS	80.090	80.090	**81.900***	78.281	78.281

* indicate the maximum Lessthan5 value in different regression algorithms

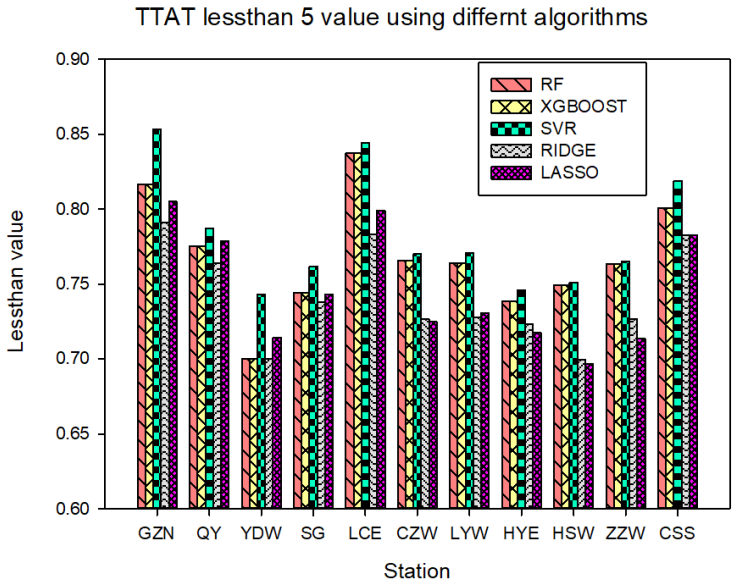

Figure 6: Lessthan5 value of TTAT using different algorithms

The results indicate that (1) the SVR algorithm has the highest Lessthan5 value at the stations in comparison with other algorithms. This proves that the SVR algorithm is the best algorithm for the TTAT predictive model. (2) The TTAT predictive accuracy of SVR algorithm at all stations was ~0.74, which proves that the SVR algorithm has good predictive accuracy.

Furthermore, 2018 data were used as the validation data to evaluate the application of TTAT model over time. The Lessthan5 values of the validation data for Wuhan-Guangzhou HSR stations are shown in Figure 7.

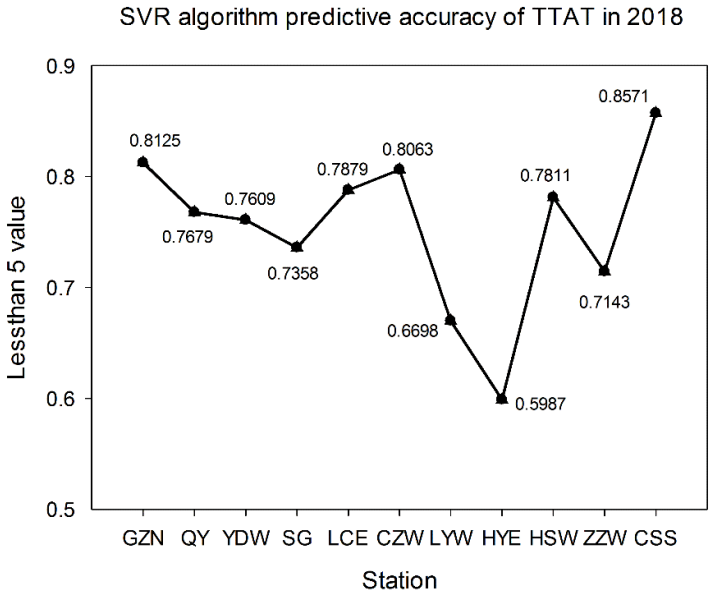

Figure 7: SVR algorithm predictive accuracy of TTAT in 2018

The TTAT predictive model has good predictive accuracy (~0.71) in most of the stations except LYW and HYE. The low precision of TTAT model for LYW and HYE is due to the low precision of NAT predictive model for these stations.

5. CONCLUSION

Prediction of the severity of PD influence in a station can assist the train dispatcher to develop rescheduling strategies and adjust the work plan of the station accordingly. The NAT and TTAT influence factors were determined by analyzing the mechanism of the PD propagation process. Moreover, the NAT and TAT predictive models were established and compared with several algorithms using the influence factors as model input. Data obtained from March 2015 to November 2016 were used to establish the models while the application of the models over time were evaluated using 2018 data. The main conclusions are as follows:

(1) NAT predictive model has a good predictive accuracy at Wuhan-Guangzhou HSR station based on the XGBOOST algorithm. When 2018 data were used as the test data, the results showed the NAT predictive model had a good application over time.

(2) NAT prediction results were used as the input values of the TTAT predictive model. The TTAT model was established using the SVR algorithm and compared with other regression algorithms. Furthermore, 2018 data were used as test data to test the application of TTAT model over time. The results indicate that the TTAT predictive model also has a good predictive accuracy over time.

(3) When a PD occurs, the influence scope can be obtained accurately using the NAT

and TTAT predictive models at each station. This provides a theoretical background needed by the dispatcher to develop rescheduling strategies and adjust the station work plan accordingly.

ACKNOWLEDGMENT

This work was supported by the National Nature Science Foundation of China [grant number 71871188]; the Science & Technology Department of Sichuan Province [grant number 2018JY0567]; We are grateful for the contributions made by our project partners.

REFERENCE

Chapuis, X., 2017. "Arrival Time Prediction Using Neural Networks", In: *7th International Conference on Railway Operations Modelling and Analysis. Lille (France): International Association of Railway Operations Research*, pp. 1500-1510.

Chen, T., Guestrin, C., 2016. Xgboost: "A scalable tree boosting system",In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 785-794.

Corman, F., Kecman, P., 2018. "Stochastic prediction of train delays in real-time using Bayesian networks". *Transportation Research Part C: Emerging Technologies* 95, 599-615.

Goverde R M P. Punctuality of railway operations and timetable stability analysis[D]. TU Delft, Delft University of Technology, 2005.

Goverde, R.M., 2007. "Railway timetable stability analysis using max-plus system theory". *Transportation Research Part B: Methodological* 41(2), 179-201.

Goverde, R.M., Corman, F., D'Ariano, A., 2013. "Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions". *Journal of Rail Transport Planning & Management* 3(3), 78-94.

Hansen, I.A., Goverde, R.M., van der Meer, D.J., 2010. "Online train delay recognition and running time prediction",In: *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, pp. 1783-1788.

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

Huisman, T., Boucherie, R.J., Van Dijk, N.M., 2002. "A solvable queueing network model for railway networks and its validation and applications for the Netherlands". *European Journal of Operational Research* 142(1), 30-51.

Kecman, P., Corman, F., Meng, L., 2015. "Train delay evolution as a stochastic process",In: *6th International Conference on Railway Operations Modelling and Analysis-RailTokyo2015*.

Kecman, P., Goverde, R.M., 2015. "Online data-driven adaptive prediction of train event times". *IEEE Transactions on Intelligent Transportation Systems* 16(1), 465-474.

Lessan, J., Fu, L., Wen, C., 2018. "A hybrid Bayesian network model for predicting delays in train operations". *Computers & Industrial Engineering*.

Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P., 2015. "Analyzing passenger train arrival delays with support vector regression". *Transportation Research Part C:*

Emerging Technologies 56, 251-262.
Meester, L.E., Muns, S., 2007. “Stochastic delay propagation in railway networks and phase-type distributions”. *Transportation Research Part B: Methodological* 41(2), 218-230.
Milinković, S., Marković, M., Vesković, S., Ivić, M., Pavlović, N., 2013. “A fuzzy Petri net model to estimate train delays”. *Simulation Modelling Practice and Theory* 33, 144-157.
Pongnumkul, S., Pechprasarn, T., Kunaseth, N., Chaipah, K., 2014. “Improving arrival time prediction of Thailand's passenger trains using historical travel times”, In: *Computer Science and Software Engineering (JCSSE), 2014 11th International Joint Conference on*. IEEE, pp. 307-312.
Şahin, İ., 2017. “Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances”. *Journal of Rail Transport Planning & Management* 7(3), 101-113.
Smola, A.J., Schölkopf, B., 2004. “A tutorial on support vector regression”. *Statistics and computing* 14(3), 199-222.
Tang, Y., Wen, C., Huang, P., Li, Z., Li, J., Yang, Y., 2018. Support Vector Regression Models for Influenced Time Prediction in High-Speed Rail System. In: *Transportation Research Board 97th Annual Meeting*, Washington DC, United States.
Wen, C., Li, Z., Lessan, J., Fu, L., Huang, P., Jiang, C., 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.
Wen, C., Li, Z., Lessan, J., Fu, L., Huang, P., Jiang, C., Muresan, M.I., 2018. Analysis of Causes and Effects of Primary Delays in a High-Speed Rail System. In: *Transportation Research Board 97th Annual Meeting*, Washington DC, United States.
Yaghini, M., Khoshraftar, M.M., Seyedabadi, M., 2013. “Railway passenger train delay prediction via neural network model”. *Journal of advanced transportation* 47(3), 355-368.
Yuan, J., Goverde, R., Hansen, I., 2002. “Propagation of train delays in stations”. WIT Transactions on The Built Environment 61.