Proceedings of the Seventh International Conference on
Railway Technology: Research, Development and Maintenance
Edited by: J. Pombo
Civil-Comp Conferences, Volume 15, Paper 21.8
Civil-Comp Press, Edinburgh, United Kingdom, 2026
ISSN: 2753-3239, doi: 10.4203/ccc.15.21.8
©Civil-Comp Ltd, Edinburgh, UK, 2026

A Study on Railway Departure Delay Prediction Using a Retrieval-Augmented Time Series Diffusion Model

T. Fukuda, S. Takahashi and H. Nakamura

College of Science and Technology, Nihon University, Japan

Abstract

In high-frequency urban railway operations, departure delays at stations may propagate to subsequent trains and develop into chronic delays. Conventional delay prediction methods generally output a single predicted value, although actual delays are probabilistic phenomena affected by passenger flow, boarding and alighting time, and interactions among trains. This paper proposes a railway departure delay prediction method using a Retrieval-Augmented Time Series Diffusion Model (RATD). The proposed method retrieves past delay sequences similar to the observed historical sequence and uses them as reference information to guide the diffusion-based generation process. By generating multiple prediction sequences under the same input condition, the method represents departure delay prediction as a probability distribution. Using actual train operation records from a commuter railway line in the Tokyo metropolitan area, the generated distributions were compared with actual departure delays. The results indicate that the proposed method can qualitatively capture major delay levels and increasing, stable, and decreasing delay trends, suggesting its potential for uncertainty-aware train operation control support.

Keywords: railway departure delay prediction, retrieval-augmented time series diffusion model, probabilistic forecasting, generative models, train operation control, uncertainty quantification.

1 Introduction

Railway operation in the Tokyo metropolitan area of Japan is conducted at high frequency, with headways of two minutes, in order to accommodate enormous commuting demand. Therefore, train headways are short, and factors such as increased boarding and alighting time due to congestion cause delays. Delays occurring in some trains then propagate and expand successively to following trains, resulting in chronic delays. When a large delay occurs, adjustment is performed by extending station dwell times for the purpose of equalizing the intervals between trains. At this time, the decision regarding how many seconds the station dwell time of a train should be extended is made based on the experience of the station dispatcher. Therefore, no method has been established for determining how many seconds the dwell time should be extended. Accordingly, in Japan, where the working population is decreasing, clearly indicating the criteria for taking countermeasures is a very important issue for implementing countermeasures without relying on experience.

Therefore, the objective of this study is to provide a basis for deciding whether operational countermeasures should be implemented by predicting departure delay times at stations and identifying increasing or decreasing delay trends. In previous studies, it has been common to output a single delay value using regression models. In many studies, the accuracy of delay time prediction within an error of 5 seconds is approximately 60% [1]. This is considered to be because actual delays are probabilistic phenomena that occur through the combination of many factors, such as variations in boarding and alighting time, passenger flow, and the acceleration and deceleration of preceding and following trains. Considering these characteristics, a single predicted value cannot represent the range of possible delay times or their probability distribution, and is therefore not sufficient as decision support in train operation management. The inability to obtain prediction information that includes uncertainty is an issue because it makes it difficult to detect signs of delay propagation in advance and to make appropriate decisions on headway adjustment. In response to this issue, by using a generative model, it becomes possible to generate multiple delay times under the same conditions and perform probabilistic prediction, thereby enabling decision-making that takes possible outcomes into account. However, probabilistic generative models for time series data face limitations in information content and insufficient numbers of data, making learning and stable generation difficult.

In this paper, to address these issues, we propose a delay prediction method using a Retrieval-Augmented Time Series Diffusion Model (hereafter RATD) [2]. RATD is a model that aims to stabilize learning and generation by utilizing past similar time series as conditional information to address issues specific to diffusion models, such as limitations in information content and insufficient numbers of data in time series data. In this paper, focusing on the probabilistic generation characteristics of the model, we extend the framework for uncertainty evaluation in the delay prediction problem by performing generation multiple times from the same input conditions and analyzing the resulting prediction distribution. This makes it possible to predict the range of delay propagation that cannot be represented by a single predicted value.

2 Methods

In this paper, actual train operation records from commuter railway lines in the Tokyo metropolitan area of Japan are used. The target consists of inbound trains operated on weekday mornings from 7:00 to the 10 o'clock hour in fiscal years 2018 and 2019. The target section for evaluation consists of the six stations from Station A to Station F, and Station D is used as the target station for prediction. The train operation records include the actual arrival time and actual departure time of each train at each station, recorded in seconds. In this paper, the train operation records and the train timetable are compared, and the following seven types of features are calculated for each train and each station.

$$\mathrm{ArrivalDelay}_{j,s} = T_{j,s}^{\mathrm{arr,act}} - T_{j,s}^{\mathrm{arr,sch}} \quad (1)$$
$$\mathrm{DepartureDelay}_{j,s} = T_{j,s}^{\mathrm{dep,act}} - T_{j,s}^{\mathrm{dep,sch}} \quad (2)$$
$$\mathrm{RunningTime}_{j,s} = T_{j,s}^{\mathrm{arr,act}} - T_{j,s-1}^{\mathrm{dep,act}} \quad (3)$$
$$\mathrm{InterstationDelay}_{j,s} = \mathrm{ArrivalDelay}_{j,s} - \mathrm{DepartureDelay}_{j,s-1} \quad (4)$$
$$\mathrm{StoppingDelay}_{j,s} = \mathrm{DepartureDelay}_{j,s} - \mathrm{ArrivalDelay}_{j,s} \quad (5)$$
$$\mathrm{DwellTime}_{j,s} = T_{j,s}^{\mathrm{dep,act}} - T_{j,s}^{\mathrm{arr,act}} \quad (6)$$
$$\mathrm{ArrivalInterval}_{j,s} = T_{j,s}^{\mathrm{arr,act}} - T_{j-1,s}^{\mathrm{arr,act}} \quad (7)$$

Here, j represents the train number, s represents the station, j-1 represents the preceding train, and s-1 represents the previous station in the direction of travel. In addition, arr represents arrival, dep represents departure, sch represents the scheduled timetable, and act represents actual data. Each train is also assigned an arrival-order label indicating the order in which it arrived at Station D after 7:00, as discrete information representing the time of day.

The input sequence consists of eight consecutive trains and the six stations from Station A to Station F. Each train and each station are given the seven types of features described above. Among the eight trains, the first five trains are used as the historical sequence x_H, and the latter three trains are used as the prediction sequence x_P. The generative model generates the prediction sequence for the latter three trains, and in this paper, the departure delay at Station D among them is used as the main evaluation target.

The data from fiscal year 2018 are used to train the model, and the data from fiscal year 2019 are used for testing. In addition, the train operation records base referenced by the retrieval module is constructed from the data from fiscal year 2018 in both training and inference. This enables the prediction performance of the proposed method to be evaluated under the condition that the test data from fiscal year 2019 are not included in the retrieval database.

The model configuration of RATD is shown in Fig. 1. This model consists of a retrieval module, which extracts past sequences similar to the input historical sequence from the train operation recordsbase, and a generation module, which generates the prediction sequence using the extracted retrieved reference sequences as conditional information. RATD is a diffusion model that uses past similar time series as reference information in order to compensate for limitations in the number of training data and the lack of conditional information in the generation process for time series data. In this paper, this framework is applied to train delay prediction, and future delay sequences are generated probabilistically while referring to past delay trends similar to the historical sequence.

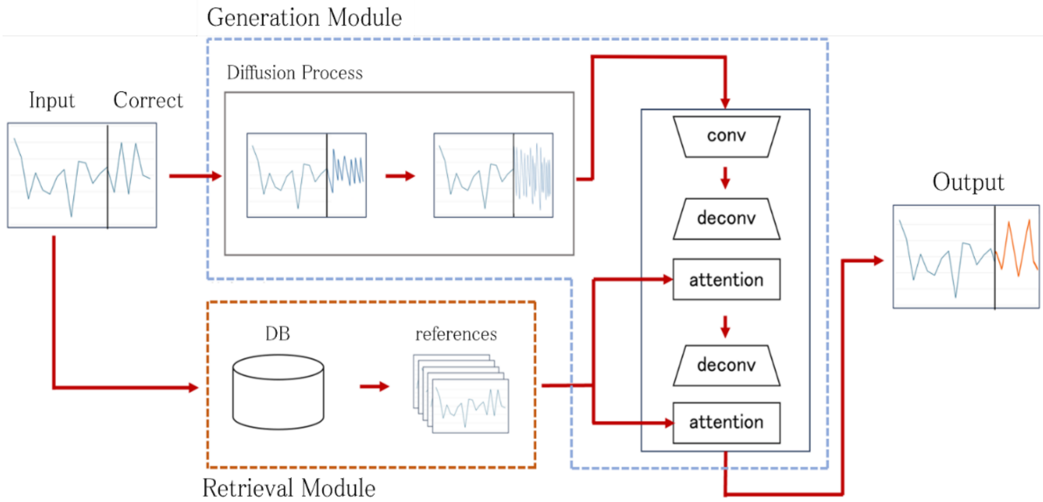


Fig. 1. Model configuration diagram of RATD

In the retrieval module, a database constructed from the train operation records for fiscal year 2018 is referenced, and sequences with high similarity to the input historical sequence are extracted. The database stores sequences in the same format as the input data structure described in the previous subsection. In this paper, in order to apply the retrieval process of RATD to railway operation data, a retrieval score that considers the similarity of feature vectors obtained by the encoder, as well as the magnitude of departure delay and the proximity of the time of day, is used. The retrieval score S_i is defined by Eq. (8) for sequence i in the database.

$$S_i = 0.830 s_{\mathrm{cos},i} + 0.098 s_{\mathrm{delay},i} + 0.072 s_{\mathrm{rank},i} \tag{8}$$

Here, $s_{\mathrm{cos},i}$ is the cosine similarity between the feature vectors of the historical sequence and the reference sequence, $s_{\mathrm{delay},i}$ is the similarity based on the difference in the total departure delay over five trains and six stations, and $s_{\mathrm{rank},i}$ is the similarity based on the difference in arrival-order labels. However, $s_{\mathrm{delay},i}$ and $s_{\mathrm{rank},i}$ are normalized indicators such that a larger value indicates higher similarity. The coefficient of each term is determined by Bayesian optimization using Optuna. The calculated S_i values are sorted in descending order, and the top five sequences are extracted as retrieved reference sequences.

A pre-trained encoder is used to calculate the cosine similarity. The encoder converts the historical sequence and the retrieved reference sequences in the database into 32-dimensional feature vectors. The train delays handled in this paper have spatial characteristics that propagate across both the train direction and the station direction. Therefore, a convolutional neural network is used as the encoder. Multiple convolutional layers extract increasing and decreasing delay trends that extend across stations and trains, and the resulting information is finally converted into low-dimensional feature vectors by fully connected layers. In addition, the encoder is pre-trained using an autoencoder structure, and by minimizing the reconstruction error of the input sequence, it represents the representative structures of delay sequences in the embedding space.

In the generation module, noise is gradually added to the prediction sequence, and the denoising process is learned using the historical sequence and the five retrieved reference sequences extracted by the retrieval module as conditions. In this paper, Denoising Diffusion Probabilistic Models (DDPM) [3] are used as the generative framework. The network responsible for denoising is composed of convolutional layers and transposed convolutional layers. The convolutional layers extract spatial trends in delay changes from a three-dimensional tensor composed of trains, stations, and features. Furthermore, an attention mechanism that integrates conditional features obtained from the historical sequence and the retrieved reference sequences is introduced into the transposed convolutional layers. This makes it possible to refer simultaneously to delay trend patterns such as increase, flatness, and decrease that are contained in multiple retrieved reference sequences, rather than depending excessively on a single nearest-neighbor sequence.

The loss function is defined as the mean squared error based on the difference between the noise estimated by the model and the noise actually added at each diffusion timestep. In addition, by applying weights according to the diffusion timestep, learning places emphasis on denoising performance at timesteps where the amount of noise is large. During inference, the prediction sequence part is initialized with random noise, and the reverse diffusion process using the trained model is applied sequentially to generate a prediction sequence that is consistent with the historical sequence and the retrieved reference sequences. By performing this process multiple times for the same historical sequence, a prediction distribution representing the range that delay time can take is obtained, rather than a single predicted value.

3 Results

In this paper, focusing on the fact that the proposed method outputs a prediction distribution rather than a single predicted value, we examined whether the generated delay distribution reflects the distributional characteristics of the actual data and the trends in delay progression. The evaluation target was the departure delay at Station D during the weekday morning time of day in fiscal year 2019. For each input sequence, prediction generation was performed 100 times, and the distribution of the generated prediction sequences was compared with the distribution of the same trains in the actual data. Since the actual data consist of one sequence for each input sequence, whereas the generated data consist of 100 sequences for each input sequence, the frequency counts of the histogram of the generated data were divided by 100 for normalization in the distribution comparison.

Figure 2 shows the distributions of actual data and generated data for departure delay at Station D. Both the actual data and the generated data have major frequency bands for each train, and the generated distribution generally corresponds to the delay levels frequently observed in the actual data. This result indicates the possibility that the proposed method can reflect delay levels that are likely to occur under the input conditions in the generation process by using past retrieved reference sequences as reference informatio.

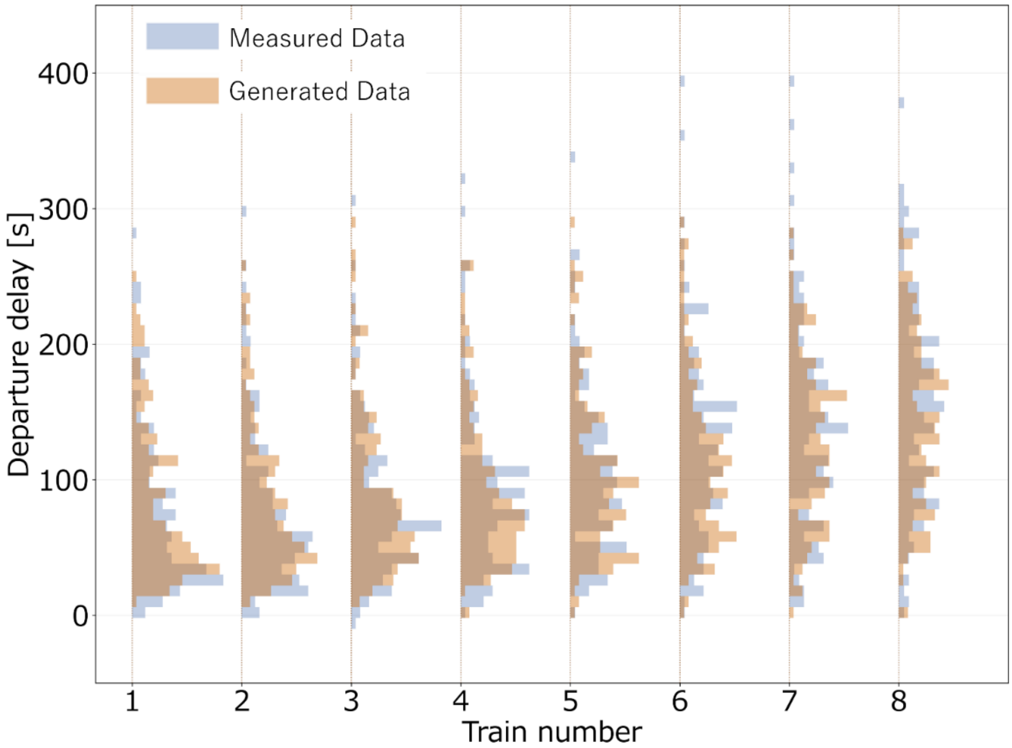
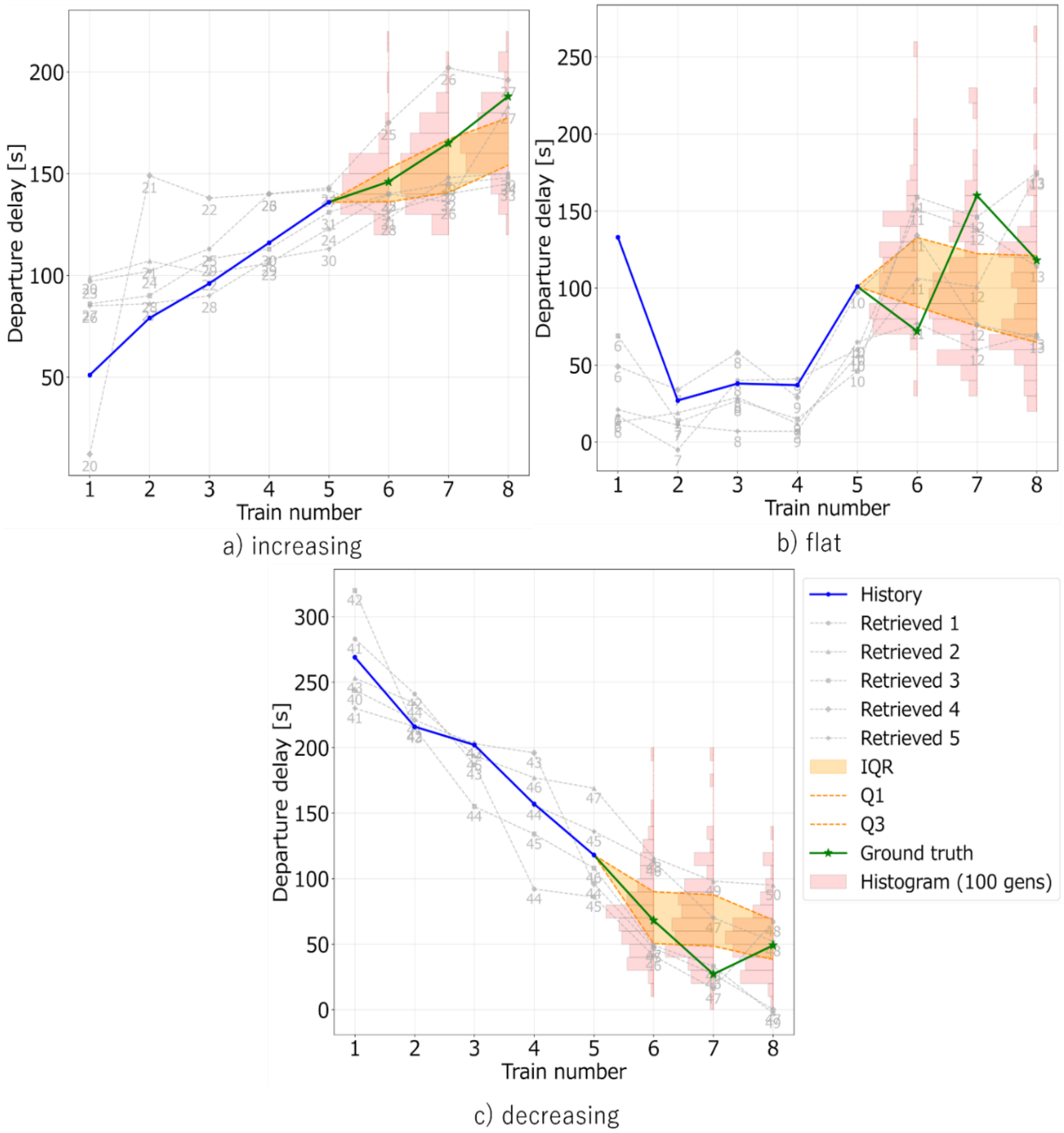


Fig. 2. Comparison of the distributions of generated delay times and actual delay times

Next, delay progression was divided into three patterns, namely increasing, flat, and decreasing, and we examined what kind of prediction distribution the proposed method generates for each trend. Figure 3 shows representative examples of prediction generation for each trend. In the figure, the blue line indicates the historical sequence, the green line indicates the ground-truth sequence, and the gray dotted lines indicate the five retrieved reference sequences extracted by the retrieval module. In addition, the pink histogram indicates the distribution of departure delay at Station D obtained from 100 generation results, and the orange range indicates the interquartile range of the generated distribution. In this paper, with application to train operation control support in mind, we focus not on matching the ground-truth value as a single value, but on whether the central range and trend of the generated distribution are consistent with the actual delay progression.



a) increasing

b) flat

c) decreasing

Fig. 3. Generated results by delay trend

In the increasing trend shown in Fig. 3a), the interquartile range of the generated distribution moves upward as the sequence proceeds to the prediction target trains. This is considered to be because the delay propagation trend included in the retrieved reference sequences was reflected in the generation process. In addition, the ground-truth sequence is also located within or near the interquartile range, and the proposed method represents future scenarios in which delays expand not as a single predicted value but as a distribution with a possible range of delays. Therefore, in the case of the increasing trend, it can be interpreted that the risk of increasing delay can be represented as an upward shift in the prediction distribution.

In the flat trend shown in Fig. 3b), the generated distribution does not deviate greatly toward a clear increasing or decreasing direction, but forms a range centered around a constant delay level. From this, the proposed method is considered to generally represent the tendency of delays to remain at a similar level without greatly expanding or converging. On the other hand, the ground-truth sequence includes local upward and downward fluctuations, and some ground-truth values are outside the interquartile range. In the flat trend, delays may increase or decrease within a short time due to variations in boarding and alighting time and temporary headway adjustments. Therefore, even when the trends of retrieved reference sequences are referenced, there may be cases in which local fluctuations cannot be represented sufficiently.

In the decreasing trend shown in Fig. 3c), the interquartile range of the generated distribution moves downward as the sequence proceeds to the prediction target trains. This is considered to be because the delay recovery trend included in the retrieved reference sequences was reflected in the generated results. The ground-truth sequence also shifts overall in the decreasing direction, and the proposed method represents the trend in which delays tend to recover as a generated distribution. However, at some points, the ground-truth value is located near the lower end of the interquartile range or outside the range, suggesting that the convergence speed of delays may not be represented sufficiently. This point suggests the need to include additional information, such as the selection of retrieved reference sequences for decreasing trends, passenger congestion rates, headways, and the presence or absence of train operation control, in the retrieval indicators.

From the above, the proposed method showed the possibility of reflecting delay progression trends contained in multiple retrieved reference sequences and generating multiple future scenarios from the same input conditions. In particular, the prediction distribution shifts upward for the increasing trend and downward for the decreasing trend, indicating that conditional information based on retrieved reference sequences is reflected in the shape of the generated distribution. On the other hand, there are cases in which local increases and decreases observed in the flat trend and the convergence speed in the decreasing trend cannot be sufficiently represented by the interquartile range of the generated distribution. Therefore, in the future, in addition to evaluation targeting multiple stations and time of days, it is necessary to conduct quantitative evaluation of distributional prediction performance using CRPS and prediction interval coverage probability.

4 Conclusions and Contributions

This paper focused on the issue that conventional single-value delay prediction methods cannot consider future uncertainty and have limitations in their use for operation support, and proposed a probabilistic prediction method using RATD. We confirmed that the generated prediction distribution shifts according to the composition of the multiple retrieved reference sequences and that the distribution is qualitatively consistent with the distributional shape of the actual data. From this result, the proposed method enables prediction that reflects variations in the trends inherent in retrieved reference sequences, and demonstrates a certain level of effectiveness as a probabilistic prediction method.

In the future, a remaining issue is to realize uncertainty representation in which the variance of the prediction distribution changes according to the diversity of trends inherent in retrieved reference sequences. Furthermore, improving the validity of the prediction distribution is expected by introducing weather and passenger congestion rates into the retrieval indicators.

References

[1] Takumi Fukuda, Shuniti Sato, Sei Takahashi, and Hideo Nakamura: “An Examination of Machine Learning Methods for Predicting Station Departure Delay Time Based on Actual Stopping Time Data of Trains at Stations”, IEEJ Transactions on Electronics, Information and Systems, Vol. 141, No. 1, pp. 1-7, 2021.

[2] Jingwei Liu,Ling Yang,Hongyan Li,Shenda Hong: “Retrieval-Augmented Diffusion Models for Time Series Forecasting”, 38th Conference on Neural Information Processing Systems (NeurIPS), 2024.

[3] Jonathan Ho,Ajay Jain,Pieter Abbeel: “Denoising Diffusion Probabilistic Models”, Advances in Neural Information Processing Systems, Vol. 33, pp.6840–6851, 2020.