ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20230522000656
Original scientific paper

Short-time Prediction of Urban Rail Transit Passenger Flow

Jing XUAN, Jiulin SONG, Jingya LIU, Qiuyan ZHANg, Gang XUE*

Abstract: Accurate prediction of short-term passenger flow in urban rail transit systems plays a crucial role in optimizing operations and enhancing passenger experience. This study presents a scientific approach to predict subway passenger flow by analyzing characteristic patterns, identifying key factors influencing passenger flow changes, and leveraging relevant data sources. The multi-source data used in this study are described and pre-processed to capture the spatial, temporal, and other factors that contribute to subway passenger flow distribution. Utilizing the extracted features as inputs, an improved Long Short-Term Memory (LSTM) method is employed for short-term passenger flow prediction. The performance of the improved LSTM method is compared and analyzed against traditional methods. The results demonstrate that the proposed approach outperforms traditional methods in terms of prediction accuracy for the same prediction target. Furthermore, the fusion of multi-source data and the inclusion of external factors significantly enhance the prediction accuracy. This research highlights the importance of considering various factors and data sources when forecasting short-term passenger flow in urban rail transit systems. By employing an improved LSTM method and integrating multiple data dimensions, the proposed approach offers superior prediction accuracy compared to traditional methods. The findings contribute to the development of efficient and reliable prediction models for optimizing urban rail transit operations and improving passenger services.

Keywords: data mining; LSTM; prediction of short-time passenger flow; rail transit

1 INTRODUCTION

In recent years, with the increasing level of economic development, medium and large cities have attracted more and more foreigners to settle, i.e. the "siphon effect". The rapid expansion of the urban population has also caused greater pressure on rail trips. Thanks to the rapid development of infrastructure projects, subway has gradually become the main way for people to commute. Traditional private cars have the disadvantages of high pollution, high energy consumption and easily being in congestion, while public trip modes, such as buses and subways, can effectively reduce environment pollution and traffic congestion, and have been vigorously developed by various countries and governments for their advantages of high capacity, high efficiency, high speed, low carbon and environment-friendly. It promotes environment protection while serving the publics, reducing traffic pressure and improving commuting efficiency. At present, all governments are vigorously promoting the construction of urban rail transit, reasonably formulating the line planning of rail transit to ensure all-round coverage in key areas (residential areas, commercial areas and other dense areas of passenger flow), meanwhile achieving the connection and integration between the construction of rail transit and other modes of transportation, so as to truly benefit the publics and make it convenient for the people. At the same time, the public has been given travel subsidies, off-peak discounts and other preferential policies one after another to encourage more trips on public rail transportation. As a result, rail transit is growing rapidly in major cities gradually showing high density and high coverage development trend. When studying rail-related contents, the huge amount of data stored in the AFC system (Automatic Fare Collection System) are often used and analyzed. AFC is an automatic ticketing system for urban rail transit, which can achieve ticket sell and checking, billing, charging, counting, scoring and management in rail transit through automation. The AFC system stores OD (Origin-Destination) data of passengers. Based on OD data, passenger's travel behaviour and the characteristics of passenger flow patterns can be analysed, thus rail operators can be guided to maintain the internal order of the subway, improve operational efficiency and increase passenger satisfaction. Therefore, how to maximize the utility of data and how to use it to generate maximum value are the keys to current research. As more and more people use the subway as a mode of trip, it often generates situations such as crowded passenger flow during rush hours, surge in passenger flow on holidays, and abnormal passenger flow during special events. For example, the daily commute in the rush hours in the morning and evening will be crowded, big crowds on October Golden Week and other holidays, falling and trampling accidents, a short surge in passenger flow caused by large events such as concerts and soccer games, and so on. In these unconventional events, the change of passenger flow is often irregular and uncontrollable, therefore, it is necessary to have advance prediction of short-time passenger flow and make corresponding countermeasures, so that all kinds of accidents that may occur can be avoided as far as possible. This requires the rail transit operator should have an overall knowledge about the situation of the passenger flow in the first instance in all kinds of unexpected conditions, and provide early warning so as to achieve early intervention, deployment and timely de-cluttering. Therefore, it is really important to use the AFC data to summarize the historical passenger flow regularity so that the short-term passenger flow can be predicted in advance. At first, it can anticipate changes in passenger flow in future periods and prepare in advance for various possible events. Secondly, it can minimize possible risks by early warning and early deployment. At last, the current prediction of passenger flow mostly focuses on the long time forecast, and there are less very systematic methods for the change of passenger flow in a short period, and there is also a lack of knowledge about the urgent change of passenger flow in a short period. Based on the mentioned background analysis, the use of scientific and rigorous methods for short-time passenger flow forecasting, so as to improve the travel efficiency of passengers, improve the rail transit operation environment, and avoid possible risks and hazards, is currently the most important part of improving the rail transit system. However, how to establish a set of scientific and objective systems of the prediction of short-time passenger flow in rail transit, so as to make it have high anti-interference, universal applicability to deal with various kinds of emergencies, and accuracy of prediction, is still a problem worthy of in-depth study and urgent to be solved. Research objective and contribution is how to establish a set of scientific and objective methodology systems of short-time passenger flow prediction for rail transit, so as to make it have a high degree of anti-interference, universality to deal with various kinds of emergencies, and accuracy of prediction. In this paper, the influence of passenger flow changes in the characteristics of the value input to the machine learning model, divided into the training set and test set for the prediction of short-term passenger flow. In the prediction method, this paper adopts the improved LSTM algorithm for prediction. The results show that the fusion of multiple data and external factors (weather, emergencies, etc.) can well improve the prediction accuracy.

2 LITERATURE REVIEW

The current methods of prediction of the passenger flow include the following two stages: the first is the stage of prediction model based on mathematical and statistical methods, which are based on statistical theory for prediction of the passenger flow; the second is the stage of prediction model based on machine learning methods, which is based on machine learning neural networks, etc. for prediction.

2.1 The Overview of Prediction Models Based on Mathematical and Statistical Methods

Guo et al. [1] proposed an adaptive Kalman filter model to predict passenger flow by using a generalized autoregressive conditional heteroskedasticity model (GARCH) and a seasonal difference autoregressive sliding average model (SARIMA) with Kalman filter model. The results showed that the large fluctuations in flow can be well handled when using this method for prediction, and the good prediction effect of the adaptive Kalman filter was proved by a sensitivity test. Jiang [2] predicted the short-time passenger flow. Firstly, the data were fused and processed, then the time series data of highway traffic flow were synthesized using chaos theory, and the traffic flow was predicted using support vector method and radial basis function neural network method. Zhang et al. [3] used the Kalman filter method combined with the time-series lag characteristics of short-term passenger flow to predict the changes of short-term passenger flow and summarized its regularity. Xiong et al. [4] constructed the state equations for forecasting problems based on the Kalman filter principle with Beijing rail transit as the research object, modified the state transfer matrix using historical data, determined the value of the next state transfer matrix by grey-scale relationship analysis, and predicted the passenger flow at Xidan subway station based on the above process. Cai et al. [5] used ARIMA model to forecast passenger flow based on Guangzhou Metro AFC data for practical application. Qin and Dong [6] proposed a combined ARMA-RBF prediction method of passenger flow to forecast the passenger flow data of Beijing Metro Line 4 and verified that this improved method can effectively improve the accuracy of prediction of passenger flow. Milenković et al. [7] used the SARIMA model to forecast the Salvia railroad passenger flow considering the autocorrelation of seasons and verified that it performed well. Davis and Nihan [8] used the K-NN method to predict short-time passenger flows and other related problems, so that they avoided the step of using external parameters and finding the best, while comparing the results with a linear time series model; and the comparative analysis concluded that their chosen K-NN model was more advantageous for the prediction of short-time passenger flow. Lin et al. [9] selected the passenger flow of the subway station of Guangzhou railway station for short-time prediction in order to predict the change and fluctuation of passenger flow in transportation hubs, and proposed a K-order nearest neighbour pattern matching model to predict the change of passenger flow on weekdays and holidays respectively and achieved good prediction results. Similarly, Huan et al. 10] also used the K-order nearest neighbour algorithm and also selected the passenger flow of Guangzhou subway station for the case study, and the prediction model was proved to have stronger stability and higher accuracy. This kind of statistical methods has better accuracy in the prediction in the case of little change in passenger flow. However, this kind of methods often fails in scenarios where the passenger flow fluctuates drastically and the data oscillates strongly due to unexpected conditions. The model of this type of methods is relatively simple and easy to be operated, but its prediction accuracy decreases significantly once it faces to complex situations.

2.2 The Overview of Prediction Models Based on Machine Learning Methods

As machine learning techniques are constantly updated and iterated, more research has improved the original neural networks, extending from shallow to deep learning mechanisms and focusing on deep learning approaches. Qin [11] modified the traditional LSTM model and used particle swarm algorithm to optimise the parameters, and the model was significantly improved. Li [12] took Beijing urban rail transit data as the research object, considered the influence of multiple factors, and improved the LSTM model to establish a passenger flow prediction model. As the results showed, the improved method has significant superiority over other modelling methods, such as ARIMA model, historical averaging method, and traditional LSTM. Polson and Sokolov [13] proposed that changes in passenger flow were accompanied by large scale state transitions, and deep neural networks were able to identify and mine such nonlinear spatio-temporal transformation features very well, so a linear model approach of line depth structure combining regularized expressions and curvilinear tangent sequences was proposed to solve such problems. Tian and Pan [14] used passenger flow data to achieve better prediction accuracy with LSTM-RNN model and applied it to the practice in the prediction of short-time passenger flow. Impidovo et al. [15] proposed a new generative deep learning framework that can efficiently process time series data and named TrafficWave for dealing with the problem about traffic prediction. Shi et al [16] used three kinds of neural networks:

BackPropagation Neural Network, Long-Short Term Memory Neural Network and Random Forest method. By comparing the accuracy of each method, it was proved that Long-Short Term Memory outperformed the other models in all aspects of prediction, providing a high value for realistic applications. Wang et al. [17] proposed a model framework based on the prediction of short-time passenger flow, where the data are first clarified and pre-processed as required, and then the LSTM neural network in deep learning algorithm is used to predict the short-term passenger flow. This model has strong robustness and significant prediction accuracy. Wang et al. [18] analysed the spatio-temporal characteristics of short-term passenger flow by using long-short term memory neural networks and self-adaptive K-means data aggregation algorithm, and proposed a deep learning model of k-ConvLSTM to predict the passenger flow of Shenzhen Subway. The results prove that this method outperformed other methods in terms of root mean square error, mean absolute error, and absolute error percentage. Huang et al. [19] used a multi-task regression model as the top layer of the deep learning structure and used supervised learning methods to predict the data, while unsupervised learning methods were used to obtain passenger flow features based on the underlying structure of the Deep Belief Network convinced degree network model, and a weighted sharing mechanism was used, so that the prediction model obtained better prediction results overall. Li et al. [20] chose Shenzhen rail transit network as a case and used convolutional neural network to extract the association between each station and a long-short time memory network to extract the temporal association to predict all incoming and outgoing passenger flows, and achieved high accuracy. Jia et al. [21] studied the performance of DBN and LSTM in the prediction of short-term passenger flow by using rainfall as input for non-flow parameters. In conclusion, in the research of passenger flow prediction, the prediction based on machine learning methods often has stronger anti-interference and higher adaptability than traditional mathematical and statistical methods; especially the adoption of deep learning methods has shown good prediction advantages.

2.3 Literature Summary

There are many methods for the prediction of short-term passenger flow, and the prediction performance varies according to the transformation of conditions. At present, there are various research methods on short-time passenger flow prediction, from statistical-based time series models to deep learning neural network models, all of them have good prediction results. However, different methods have different effects on different prediction objects and prediction time granularity, and no model can yet be exhaustive and perform well in all aspects. Based on this situation, this paper selected multiple prediction methods for comparison, and when the prediction object changed, the prediction performance of different methods corresponded to produce different responses, so that in the face of different scenarios, the prediction accuracy was improved by complementing the strengths and weaknesses of multiple models, so that a more comprehensive solution to the problem could be achieved. In summary, this paper firstly determined the research scope as the prediction of short-time passenger flow, and used the method of mathematical and statistical verification to determine the time granularity. Meanwhile, this paper fused the data from multiple sources, fused the AFC data with weather and unexpected events, etc., studied the factors affecting the change of passenger flow and extracted the characteristics of passenger flow, and then used the multi-methods of statistics, machine learning and deep learning for passenger flow prediction, and compared the results for analysis.

3 THE MULTI-SOURCE DATA FOR PREDICTION MODEL

3.1 AFC Card Data

The metro AFC card data in this paper came from the summary of data from various platforms provided by the relevant operating companies of Beijing urban rail transit. Since the daily passenger flow of Beijing subway was very large, the large amount of data provided a solid data base for our study and facilitated further processing and analysis.

Table 1 Metro AFC data description

Field	Field Name	Field Description
TXN_DATE_TIME	Exit Time	Exit time, specified to the second
DEVICE_LOCATION	Arrival Station Number	Nine digits number,corresponding to different station name
CARD_SERIAL_NUMBER	Card Number	Card ID
TRIP_ORIGIN_LOCATION	Starting Station Number	Nine digits number,corresponding to different station name
ENTRY_TIME	Entry Time	Entry Time, specified to the second
PRODUCT_ISSUER_ID	Card Major Category (99 One Card, 1 One Card)	Card Major Category
PRODUCT_TYPE	Card Subtype	Card Subtype
PAYMENT_VALUE	Chargeback Amount	(100 times, 1 means the staff ticket or counting ticket, 0 means the free ticket, welfare ticket or station work ticket)
CARD_LIFE_CYCLE_COUNT		
RECONCILIATION_DATE	Reconciliation Date	Uniform reconciliation time for metro tickets
SETTLEMENT_DATE	Settlement Date	Settlement date of metro ticket amount
SAM_ID	Security Key	Security key
DEVICE_ID	Gate Number	Gate number, corresponding to different swipe gate entrance
SOURCE_PARTICIPANT_ID	Operating Company	Operating company code
PURSE_REMAINING_VALUE	Wallet Balance	Wallet balance
PTSN	Transaction Serial Number	Transaction serial number

At the same time, the metro gates were rich in data types of card swipe, including various types of tickets such as elderly, students and general tickets, so the collected population was more dispersed and comprehensive, and the data had objectivity and accuracy. In the actual application process, the gate swipe record of each entry/exit station was recorded and stored into the system as one record. The record included exit time, arrival station number, card number, starting station number, entry time, card major category (99 One Card, 1 One Card), card subtype, chargeback amount, reconciliation date, settlement date, security key, gate number, operating company, wallet balance, and transaction serial number. These fields covered the areas we are studying. The specific swipe record fields, field names and corresponding descriptions are shown in Tab. 1.

3.2 POI Data

POI, i.e., Point of Interest, each POI contains coordinates, name, category and other information. For example, car 4S stores, subway station stores, residential houses, etc. can be regarded as POI. In this paper, we choose the crawler results of Internet open source Gaode Map, and then correspond the POI points with Gaode Map background API classification, which is used to verify the accuracy of site clustering results.

3.3 Weather Data

Weather data from National Meteorological Information Center - China Meteorological Data Network. The collection frequency of this weather data is once every ten minutes, with a total of 144 items throughout the day, containing fields such as station number, date, one-hour precipitation, relative temperature and humidity, and 10-min wind speed. The data used in this paper includes the data of two months from March-April 2018, and the specific data fields and data are shown in Tab. 2.

Table 2 Weather data description

Field	Field Name	Data Example
STA	Station Number	54511
DATE	Date	2020/10/19
PRCP	One-hour Precipitation / mm	0
HUM	Relative Humidity / %	59.3
TEMP1	Feeling Temperature / ℃	−2.7
TEMP2	Temperature / ℃	1.9
SPD	10 min Wind Speed	5.4
WEA	Weather Condition	0 - 6
PRE	Pneumatic Pressure / HPa	1016

3.4 Daily Operation Scheduling Data

The daily operation scheduling data mainly comes from the daily metro operation reports, including the summary of daily, monthly and yearly accumulated passenger traffic data of each line, the summary of daily, monthly and yearly accumulated data of electricity consumption, and the fault records of the day, etc. The specific data structure and data are shown in Tab. 3.

Table 3 Metro Operation Dispatching Data Description

Field	Field Name	Data Example
DAYFLOW	Daily Passenger Flow	1076573
MONTHFLOW	Monthly Passenger Flow	11769371
YEARFLOW	Yearly Passenger Flow	348698303

3.5 Data Fusion

Data integration is the federation of disparate data sets. This paper gives the study of the prediction of short-time passenger flow of rail transit based on AFC data, i.e., the characteristics of subway passenger flow distribution pattern in short time period and the change of passenger flow in short time period in the future. In this process, there are many factors affecting the change of passenger flow, including temperature, humidity, wind speed, pandemic, holidays, POI, etc. In order to consider the effect of various factors on passenger flow, several data sets must be integrated and thus aggregated for further analysis and prediction. The data collection in this paper mainly used SPSS, Python, Excel and Tableau tools. While doing the linking of different datasets, the same fields are selected for association so that different data sets can be linked with common attributes to do the next step of analysis and form a unified whole. For example, weather data and AFC data are connected by using date and time; AFC data and POI data are connected by using subway stations, etc. In practice, it is especially important to pay attention to the uniformity of data formats. While using the software for automatic data set association, the same fields need to have the same expression. For example, some data of date are "YYYY-MM-DD", some are "YYYYMMDD", some are "YYYY/MM/DD"; the time data is "00:00:00" or "00:00", with slight differences in both time granularity and date format. Therefore, the process of data integration requires unifying the formats and fully understanding the actual meanings of different fields, or performing certain data deconstruction and coding according to actual needs. The combined partial data are shown in Tab. 4.

Table 4 Merged Data Sample

Time		Station	Passenger Flow	Week	Whether Workday	10-min Wind Speed / m/s	Barometric Pressure / Hpa	Relative Humidity / %	Temperature / ℃	One-hour Precipitation	Feeling Temperature / ℃	Weather Condition	Contingency
2018/3/1	5: 50	Chegongzhuang	15	7	1	3.4	1015.7	28	1.9	0	−2.767	0	0
2018/3/1	6: 00	Chegongzhuang	30	7	1	3.4	1016	29	1.2	0	−2.39869	0	0
2018/3/1	6: 10	Chegongzhuang	39	7	1	3.4	1016	29	1.2	0	−2.39869	0	0
2018/3/1	6: 20	Chegongzhuang	48	7	1	3.4	1016	29	1.2	0	−2.39869	0	0

4 DATA MINING AND EXTRACTION FOR PREDICTION MODEL

4.1 Extraction of Spatial Influencing Factors of Subway Passenger Flow

Based on the classification of subway stations, we observed the characteristics of passenger flow distribution in different categories of stations, so that we could understand the characteristics of spatial differences in the distribution of subway passenger flow. In the following, passenger flows will be classified according to workdays and non-workdays, and the daily changes of passenger flows for the four different attributes will be compared.

(1) The characteristics of workday passenger flow distribution at different types of stations.

The workdays were all selected for the passenger flow on Thursday, October 22, 2020. Type I, sights & passenger hub type station. The selected representative station is Beijing Station. The distribution of its passenger flow throughout the day has no obvious high and low peaks. The passenger flow throughout the day is at a high level and fluctuates more obviously, with large changes in passenger flow within a short period of time, corresponding to the large number of passengers entering and leaving Beijing Station each day and the random arrival time. The traffic distribution characteristics of this type of station can be called as "All-peak Type". Type II, mixed residential stations, is represented by Caishikou Station, and Type III, mixed work stations, is represented by Chegongzhuang Station. These two types of stations are relatively similar in terms of the distribution characteristics of all-day traffic, with non-significant morning peak and significant evening peak for inbound passenger flow, and significant morning peak and non-significant evening peak for outbound passenger flow. Except for the morning and evening peaks, all other periods are low peak hours, and both of these two types of stations are mixed stations. By summarizing the distribution characteristics of passenger flow throughout the day, the inbound and outbound passenger flows are divided into morning and evening peaks, so the two types of station traffic type are called "Twin-peak Type". Type IV, suburban residential stations, the representative station is Tiantongyuan North Station. Its inbound passenger flow shows a significant morning peak and outbound passenger flow shows a significant evening peak. Except for the morning and evening peaks, the passenger flow at other times is relatively calm and there are no other peaks. The inbound and outbound passenger flows have only a single peak, so the distribution of these flows is said to be "Single-peak Type". Fig. 1 shows the distribution of inbound and outbound passenger flows of the four types of stations throughout the day on weekdays.

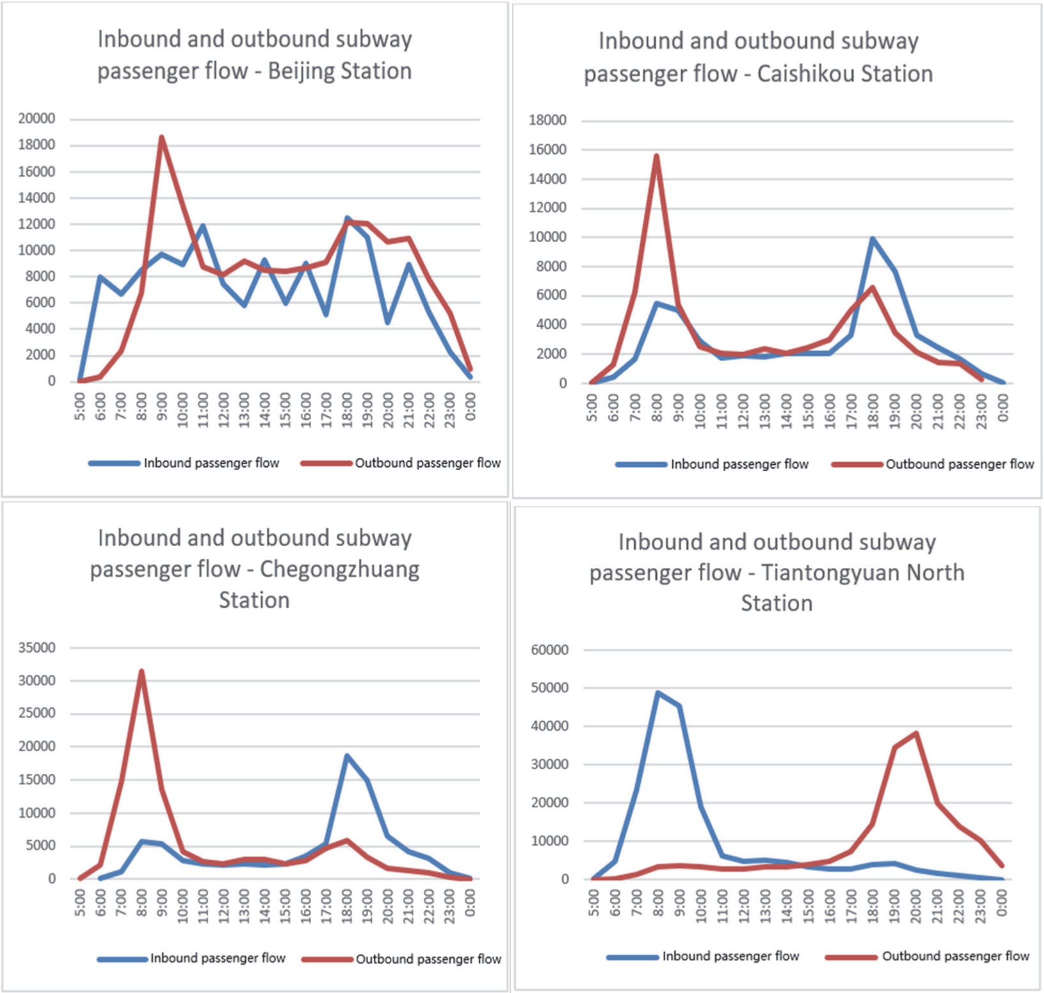


Figure 1 Passenger flow distribution of four types of stations on weekdays

(2) The characteristics of holiday passenger flow distribution at different types of stations.

The workdays were all selected for the passenger flow on Saturday, October 24, 2020. Fig. 2 shows the same spatial variability in passenger flow on holidays for stations with different attributes. Type I, sights & passenger hub type stations, the selected representative station is Beijing Station. Its holiday traffic still shows the characteristics of "All-peak", with high traffic throughout the day, and no significant peak. Type II, mixed residential stations, is represented by Caishikou Station, and Type III, mixed work stations, is represented by Chegongzhuang Station. Such sites are also in line with the " Twin-peak Type " distribution characteristics, but the flow of passengers throughout the day compared to weekdays shows a significant decline. For Type IV, suburban residential stations, the representative station selected is Tiantongyuan North Station. The passenger flow distribution is in line with the "Single-peak Type", but there is a significant drop in passenger flow compared to weekdays.

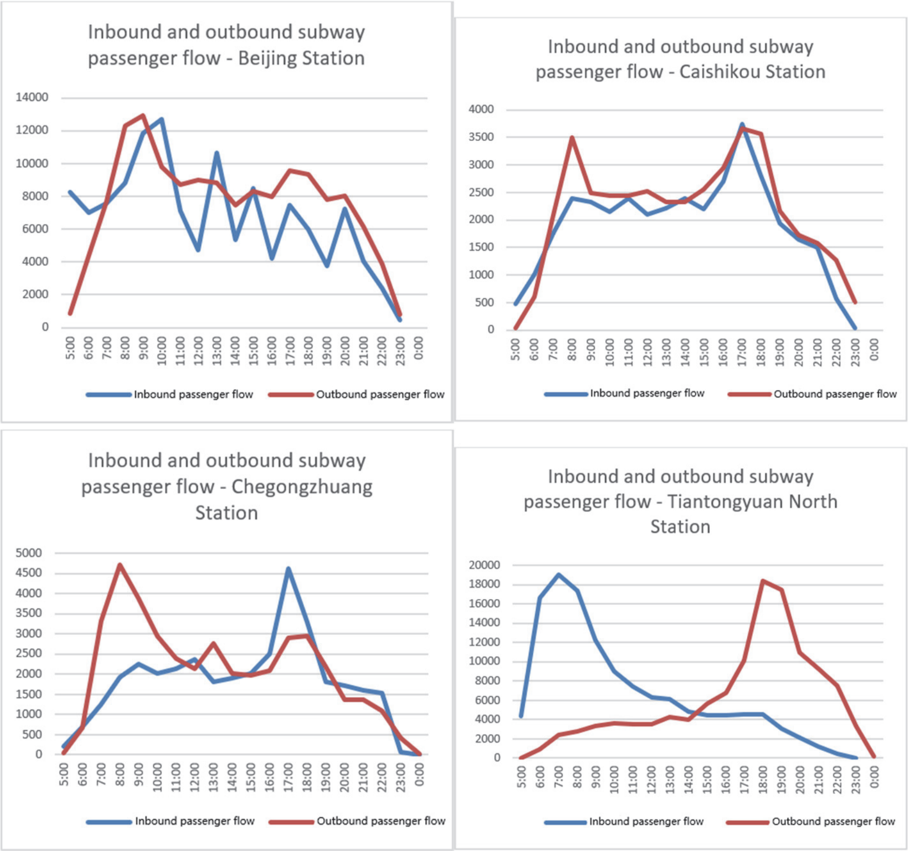


Figure 2 Passenger flow distribution of four types of stations on holidays

Table 5 Extraction and description of spatial features of subway passenger flow

Feature Name	Feature Field	Feature Description	Sample Data
Characteristics of spatial influencing factors of subway passenger flow	SpatialFactor	Classification of subway stations	1

In summary, there are significant differences in the characteristics of passenger flow distribution of stations with different spatial locations. This section analyzes the spatial distribution characteristics of metro passenger flow by selecting representative stations of subway stations with different attributes and plotting their passenger flow distribution. Therefore, the categories of subway stations (I, II, III, and IV) are input as feature values into the model for the prediction of short-time passenger flow (Tab. 5).

4.2 Analysis and Extraction Features of Time Distribution of Subway Passenger Flow

According to common sense and observation, in general, the change of subway passenger flow is gradual and mostly cumulative, with very special cases of sharp increase of passenger flow in a short period, so it can be inferred that there is a significant correlation between the current passenger flow and the passenger flow of the previous period. In order to verify this situation, this paper selected the four types of stations derived from 4.2.3 clustering, and selected representative stations among them to plot the change of passenger flow at 5 min time granularity during morning peak hours, as shown in Fig. 3. Based on this, this study performs the mathematical verification of the correlation of passenger flow in adjacent time periods by Pearson correlation analysis. Using SPSS bivariate correlation Pearson correlation analysis, the table shown in Tab. 6 was derived as a table of Pearson correlation coefficients between the current subway passenger flow and the historical passenger flow 5 min ago.

As shown in the table above, the Pearson correlation coefficient is 0.934 and the Sig is less than 0.05; therefore, they are significantly correlated. This shows a significant correlation between the passenger flows in adjacent time periods, which becomes a temporal proximity similarity characteristic of metro passenger flows. Based on this characteristic, we use the historical near-neighbourhood periods as input values when making short-time passenger flow forecasts, so that we can predict the future short-time passenger flow. The summarized input values of the subway temporal nearest neighbour characteristic are shown in Tab. 7.

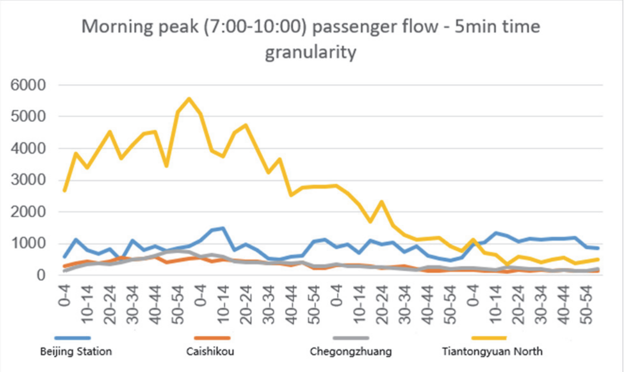


Figure 3 Morning peak (7:00 - 10:00) passenger flow 5 min time granularity

Table 6 Pearson correlation coefficient of current and historical passenger flow factors

		Current Passenger Flow	Passenger Flow After 5 Minutes
Current Passenger Flow	Pearson Correlation	1	0.937**
	Significance (two-tailed)		0.000
	Number of cases	48	47
Passenger Flow After 5 Minutes	Pearson Correlation	0.934**	1
	Significance (two-tailed)	0.000	
	Number of cases	47	47

** Significant correlation at the 0.01 level (two-tailed).

Table 7 Extraction and description of time proximity characteristics of subway passenger flow

Feature Name	Feature Field	Feature Description	Sample Data (first six periods of passenger traffic)
Characteristics of Subway Passenger Flow Time Proximity	Time Proximity	Passenger flow time series for the current time nearest historical period	[5, 8, 20, 56, 77, 90]

4.3 Analysis and Extraction of Other Influencing Factors of Subway Passenger Flow

4.3.1 Extraction of Contingency Impact Factors

Contingencies can have a large impact on subway passenger flow, and their effects can be divided into two categories: positive and negative effects. Positive impacts include holidays, concerts, and sporting events, all of which will have an increase in passenger flow compared to general conditions. Negative impacts include natural man-made disasters such as earthquakes, fires, and pandemics, and there will be less passenger flow under this type of event.

The changes of passenger flow under the influence of positive and negative events are shown in Fig. 4, respectively. The figure on the left shows the passenger flow of a site in February 2018 (regular) and February 2020 (pandemic). After observing and passing the t-test, the significance is less than 005, so it is identified that there is a significant difference between the pandemic passenger flow and the regular passenger flow, and the level of pandemic passenger flow is significantly lower than the regular passenger flow.

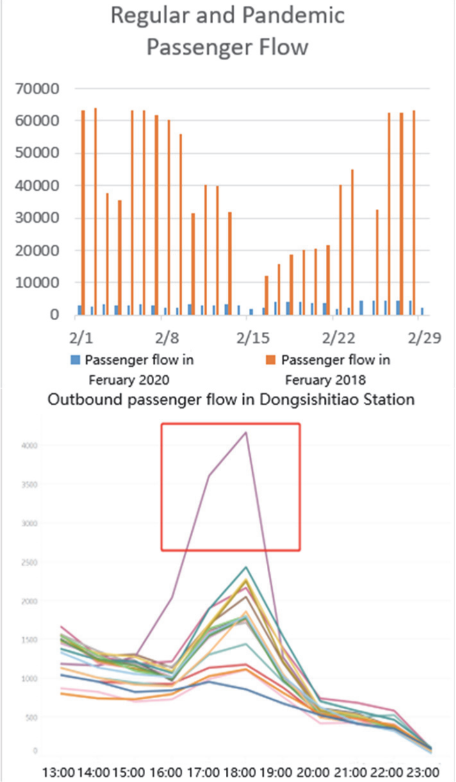


Figure 4 Comparison of passenger flow under emergency and normal conditions

The figure on the right shows the change in passenger flow from 13:00 to 23:00 in April 2018 at the DongsishitiaoStation, a subway station adjacent to the Worker Stadium, where most of the passengers taking the subway enter and exit the station when large activities are held at the Workers' Stadium. It can be seen that the passenger flow around 18:00 on April 22nd shows a different trend from other dates, with the passenger flow doubling, which is due to the large Chinese Super League match held at 19:22 in the Worker Stadium on that day, so the peak flow of passengers out of the Dongsishitiao Station at the time of 18:00. Correspondingly, after the match, the inbound passenger flow increased suddenly at the station. From this we can conclude that some unconventional events can cause unanticipated and irregular changes in rail passenger flow. There are also differences in the impact of the various types of emergencies analysed above on passenger flow. Some contingencies such as public health events, new coronavirus, natural disasters, etc. will have a greater impact on the passenger flow of the entire road network.

4.3.2 Extraction of Weather Influencing Factors

Weather is also an important factor affecting changes in passenger flow, with passengers choosing to travel less in times of bad and extreme weather. Weather can have a different impact on passenger travel depending on the time of day and route. For rigid travel, such as commuting to work and school, the impact of weather is relatively small; while for non-rigid travel, such as shopping, sightseeing and recreation, extreme weather tends to reduce passenger trips. We obtained real-time data on daily weather changes on the Internet. In order to verify the correlation between weather conditions and passenger flow changes, we used Pearson correlation coefficients to verify selected temperature, feeling temperature, relative humidity, one-hour precipitation, ten-minute wind speed, and barometric pressure values for numerical verification analysis. Tab. 8 shows the results of Pearson correlation statistics. Some of the variables that are not significantly correlated are removed and the remaining variables will be applied in the passenger flow prediction example.

Table 8 Verification of person correlation between weather and passenger flow

Variable	Pearson Correlation Coefficient	Sig (Bilateral)
One-hour Precipitation / mm	−0.356**	0.00
Relative Humidity / %	−0.092**	0.02
Feeling Temperature / °C	0.085**	0.01
Temperature / °C	0.076**	0.02
Ten-minute Wind Speed / m/s	0.123**	0.02
Weather Condition	0.453**	0.00
Barometric Pressure / HPa	−0.12*	0.05

5 THE IMPROVED LSTM MODEL FOR PREDICTION MODEL AND RESULTS

5.1 Passenger Flow Prediction Model

For the nonlinear and volatile characteristics of passenger flow sequences, a single forecasting method has certain shortcomings, so we have to make different forecasting method choices for different feature values. First of all, for the time features, the LSTM network is used for feature extraction and prediction by time series, and the attention mechanism is added, which can better reflect the weighting relationship of different features. Then, the predicted results are fused with other spatial and external features, and the data are processed to a uniform order of magnitude using normalization. After that, the results are predicted using a fully connected network and the data are transformed into actual predicted values of passenger flow by inverse normalization.

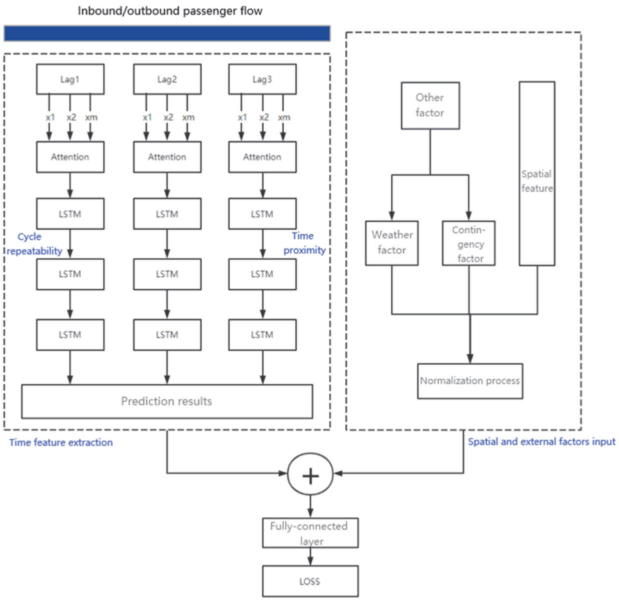


Figure 5 Improved LSTM prediction model diagram

In this method: (1) The time-series features of passenger flow over time are extracted by using LSTM. (2) Using attention mechanism, we further explore the dependency of passenger flow on different features and assign different weights to make the results more accurate. (3) Regarding the external and spatial factors of passenger flow, the data are normalized so that all data are in similar dimensions to avoid prediction errors due to order-of-magnitude differences. (4) Finally, all the features are fused and predicted with the full connection layer, fully considering all the influencing factors of passenger flow. The overall framework of the model is shown schematically in Fig. 5.

5.2 Model Parameter Setting and Implementation

(1) Implementation of the model.

1) LSTM-based time-dependent mining.

When the research problem involves the prediction of time series correlation, it is mostly handled by the LSTM method. Because LSTM has a strong long-term memory capability, it records the state information, mines the nested information of the time series, and propagates backward through the hidden nodes. Not only that, the LSTM-specific input, oblivion and output gates are determined by the activation function control. The input gate is responsible for the input and controls the effect of the new input on the predicted outcome. The oblivion gate is responsible for deciding whether the information at moment t-1 can be forgotten. When the oblivion gate is in the open state, the information from the previous moment will be retained, and when the oblivion gate is in the closed state, the information from the previous moment will be forgotten. The output gate is responsible for the output, whether the absolute last state is passed to the output or not. According to the correlation analysis of the current passenger flow and the current passenger flow in the previous order, it is found that they have strong correlation, and LSTM can be used to mine the potential information and the adjacent time passenger flow dependence of the passenger flow. The correlation process of transmitting passenger flow time series information is shown in Fig. 6. The time dependence of passenger flow is mined by a series of time-step data inputs, so that the corresponding prediction results can be output. The following is the LSTM feature mining framework, where k denotes the prediction input of the previous k stages and q denotes the predicted passenger flow of the post q stages.

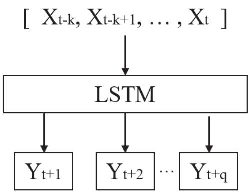


Figure 6 LSTM feature mining framework

2) Introduction of attention mechanism.

As the popularity of deep learning grows, more and more studies introduce attention mechanisms to increase the accuracy of prediction, especially in text mining with wide applications. In recent years, scholars have gradually introduced attentional mechanisms into transportation passenger flow. The attention mechanism mainly simulates human observation habits and will have different levels of attention to different observed things. This form of distributing different attention for a goal is called the attention mechanism. This study uses an improved attention mechanism that is currently commonly used, by replacing intermediate elements with a series of vectors. The essence of this is the assignment of different weights; not all elements are equally important, therefore, the more important elements will be assigned higher weights and play a greater role in classification and prediction, while non-important elements are assigned a lower weight, which gives more accurate results in the prediction process. In this paper, when extracting time-dependent features using LSTM, we introduce an attention layer to better assign weights to the passenger flow of each time sequence, so that we can better capture and extract effective information, and increase the weight of effective information, thus improving the prediction results. $H \in R^{d \cdot N}$ denotes the matrix consisting of the LSTM network output vectors [h_1, h_2, …, h_N], N can be considered here as the prediction period T. The attention mechanism eventually produces a vector of attention weights α and γ.

$$M = \tanh\left(W_h H\right) \tag{1}$$

$$\alpha = \mathrm{softmax}\left(w^T M\right) \tag{2}$$

$$\gamma = H\alpha^T \tag{3}$$

Among them, $W \in R^{d \cdot N}$, $\alpha \in R^N$, $\gamma \in R^d$. $W_h \in R^{d \cdot d}$ and $w^T \in R^d$ are the parameter matrices for the training process. With the introduction of the attention mechanism, the final result is a stitching of H and r. The output vector is represented as follows:

$$h^* = \tan h\left(W_p r + W_x h_N\right) \tag{4}$$

Among them, $h^* \in R^d$, W_p and Wx are the parameter matrices that need to be trained by the subsequent model; the output variable h^* finally passes through a fully connected layer to achieve the prediction of passenger flow. Namely, the time series data is used as the input of the Attention layer, and after being processed by the Attention mechanism layer, the data is put into the LSTM layer for prediction, and the resulting result is the result after being assigned weights by the Attention mechanism.

3) Feature processing and fusion for fully connected layers.

The fully connected neural network has better adaptability for different features on passenger flow and can map the nonlinear features better. Therefore, after LSTM, a fully connected neural network is again applied to fuse all features and predict them again. Meanwhile, in the process of feature fusion, prediction errors due to order-of-magnitude differences occur in the prediction process because of the different representational meanings of the data and the different forms of data presentation. In order to avoid such situation, all feature data are normalized, and after the prediction of the fully connected layer, the data are then denormalized to finally output the actual predicted passenger flow results.

(2) Parameter setting of LSTM model

When using the LSTM model approach for prediction, the parameters are set as follows:

1) Automatic parameter setting. The parameters of this class are selected as reasonable optimization functions, etc. After waiting for the input training data, the network changes the parameters such as weights by itself according to the magnitude of the error value until it meets the error requirement or proceeds to the maximum number of iterations. The parameters of this class are automatically adjusted by iterative training on the data. The specific settings are shown in Tab. 9.

Table 9 Parameter setting table

Parameter	Setting
Objective Function	MSE
Activation Function	Tanh
Optimization Function	Adam

2) Artificially set parameters. Such parameters are considered to be set, including the number of neurons, the number of hidden layer nodes, the number of training iterations, etc. By setting appropriate parameters, the model is able to achieve optimal prediction while ensuring operational efficiency. Among them, the hidden layer is used to obtain data features, and the setting of the hidden layer affects the prediction effect. The more the number of hidden layers, the more complex the structure of the neural network, the stronger the ability to extract features, the larger the computation, the longer the time consuming, and the easier the overfitting phenomenon; on the contrary, the prediction effect decreases. With reference to similar studies and experimental experiences, this paper sets the hidden layers to three layers, and at the same time adds Dropout function after each hidden layer and sets the parameter to 0.2, so that the purpose of discarding non-essential information and reducing overfitting can be achieved. The memory function of the collective long and short term memory neural network is used to select the appropriate number of nodes. Regarding the setting of step size, the time granularity of this paper is 10 min, and combined with other literature, a step size of 6 is selected, which is just one hour of passenger flow time series for the first six moments. By referring to similar studies and combining specific operational results, the final adjustment parameters were set as shown in Tab. 10.

Table 10 Parameter setting table

Parameter	Setting
LSTM Layers	3
Neuron	128
Dropout	20%
Epochs	20
Batch_size	32

5.3 Prediction Results and Analysis

The predicted results are fitted as shown in Fig. 7. It can be seen that the improved LSTM fits extremely well, and small peak fluctuations are shown. The LSTM prediction error MAE with the introduction of the attention mechanism is 13 and the RMSE is 19. In particular, the prediction results in the smoother segments are very well fitted, and the prediction accuracy remains high in the case of more and larger fluctuations. Overall, the prediction is good. In addition, since external factors such as days of the week, holidays, weekday breaks, weather, and unexpected events (pandemics, large events, etc.) are included in the input values, the method can effectively give a response when the input external conditions change, and show better stability. On the one hand, the method can perform well with other prediction methods to accurately predict passenger flow change data in a stable state; on the other hand, the method can effectively combine external conditions to respond in a fluctuating state of passenger flow. With reference to the model construction process and algorithm implementation process of the above mixed work-oriented station - Chegongzhuang Station, we used the same steps to predict the passenger flow of the representative stations of the other three types of stations, namely, the attraction & passenger hub station - Beijing Station, the mixed residential station - Caishikou Station, and the suburban residential station - Tiantongyuan North Station, respectively. The results of the prediction using the improved LSTM model with the introduced attention mechanism and the fit to the true values are shown in Fig. 8.

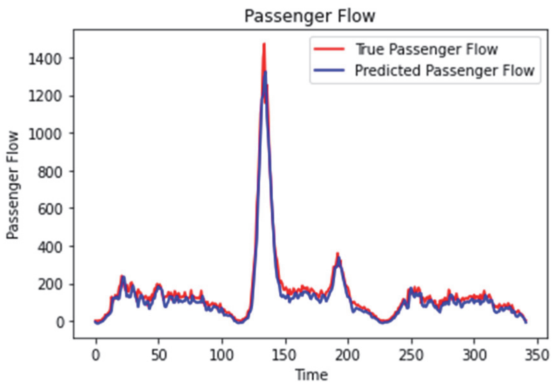
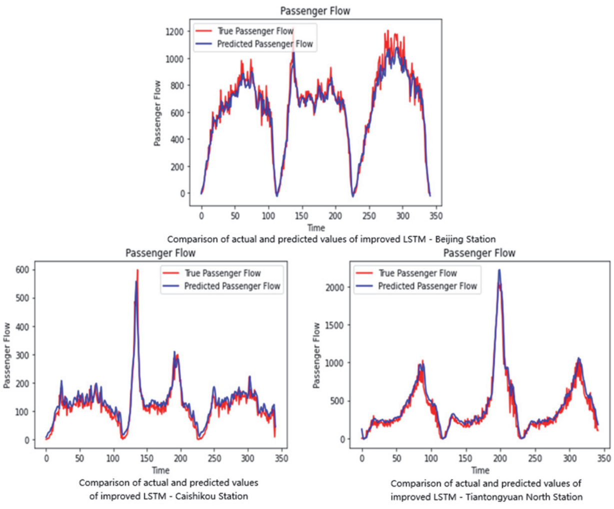


Figure 7 Fitting diagram of improved LSTM predicted value and real value (Chegongzhuang station)



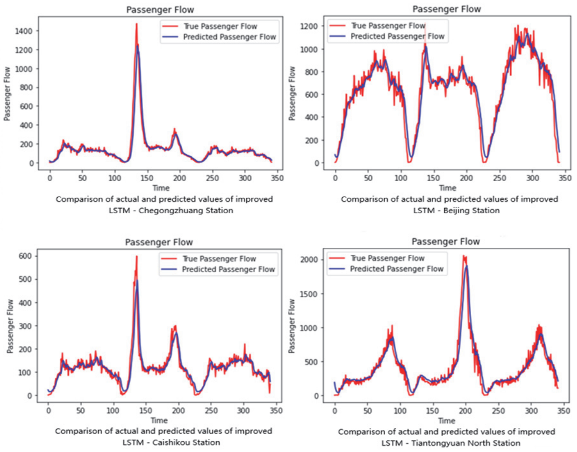
Comparison of actual and predicted values of improved LSTM - Beijing Station

Comparison of actual and predicted values of improved LSTM - Caishikou Station

Comparison of actual and predicted values of improved LSTM - Tiantongyuan North Station

Figure 8 Improved fitting diagram of LSTM predicted value and real value (Beijing Station, Caishikou station and Tiantongyuan North Station)

In order to compare the performance difference between the improved LSTM model with the attention mechanism introduced and the unimproved LSTM model, we ran the experiment again using the traditional unimproved LSTM method to predict the passenger flow at representative stations for each of the four types of stations. The results of the predicted passenger flow at the four types of stations using the traditional unimproved LSTM model and the fit to the true values are shown in Fig. 9. After calculating the error values of the predicted results of passenger flow at each station, the comparison results are obtained as in Tab. 11. By using the table data of the error results of MAE and RMSE, it can be seen that the final prediction accuracy is Chegongzhuang > Caishikou > Tiantongyuan North > Beijing Station, no matter using the improved LSTM prediction method or the traditional LSTM prediction method. It can be concluded that the deep learning LSTM prediction model has better prediction effect in the case of small passenger flow, slow change in passenger flow and obvious passenger flow pattern; as the prediction effect decreases significantly when the passenger flow is large, the passenger flow changes rapidly, and the passenger flow fluctuates significantly, but in general it maintains a good prediction accuracy, and still shows the superior performance of the method even when dealing with multi-fluctuation passenger flow problems. Meanwhile, the improved LSTM outperforms the traditional LSTM model under each prediction object and each prediction index, which illustrates the effectiveness of the improved method.



Comparison of actual and predicted values of improved LSTM - Chegongzhuang Station

Comparison of actual and predicted values of improved LSTM - Beijing Station

Comparison of actual and predicted values of improved LSTM - Caishikou Station

Comparison of actual and predicted values of improved LSTM - Tiantongyuan North Station

Figure 9 Fitting diagram of LSTM predicted value and real value (Chegongzhuang station, Beijing station, Caishikou station and Tiantongyuan North Station)

Table 11 Comparison of prediction errors of LSTM model

Prediction Method	Name of the Station	MAE	RMSE
Improved LSTM	Beijing Station	49	55
	Caishikou Station	15	20
	Chegongzhuang Station	13	19
	Tiantongyuan North Station	35	51
Unimproved LSTM(Baseline model)	Beijing Station	52	58
	Caishikou Station	15	22
	Chegongzhuang Station	15	22
	Tiantongyuan North Station	32	46

Based on the results in Tab. 11, the use of improved LSTM methods for short-time passenger flow forecasting can improve the travel efficiency of passengers, improve the rail transit operation environment, and avoid possible risks and hazards. Rail transportation often produces peak hour passenger traffic congestion, holiday passenger traffic surge, and special events under the abnormal passenger flow and so on. For example, the morning and evening peaks of daily commuting will be congested, holiday travel crowded, concerts and ball games and other large-scale activities resulting in a short-term surge in passenger flow. In the face of such unconventional events, the changes in passenger flow are often irregular and uncontrollable. The research method mentioned in this paper can predict these short-time passenger flows in advance and help rail transit operators make corresponding countermeasures, thus avoiding all kinds of possible accidents to the greatest extent possible.

6 CONCLUSION

In this paper, we study the analysis of urban rail transit passenger flow characteristics and short-time passenger flow prediction based on deep learning. By processing and analyzing the historical passenger flow data, we classify the metro stations and summarize and analyze the regular characteristics of passenger flow in spatial dimension and temporal dimension. Multiple datasets are fused, predictions are made using deep learning methods, and ultimately the full-text findings are applied in practice.

The contributions of this paper can be summarized as follows:

(1) Description and processing of multi-source data sets. The multi-source dataset is first described and pre-processed, including data cleaning, integration, transformation, and statute. By describing the data, a more in-depth and a tangible understanding of the research problem is obtained; by processing the data, the data meets the data format requirements for subsequent modelling and prediction. A data foundation is laid for the next study.

(2) Analyze and extract the factors affecting the change of subway passenger flow in three dimensions: time dimension, spatial dimension and other influencing factors. From the spatial dimension, the subway stations are clustered according to the passenger flow characteristics, divided into four categories and the distribution of POI to verify the reasonableness of the clustering results, followed by the distribution characteristics of passenger flow in each category separately. From the time dimension, the temporal nearest neighbour similarity characteristics and cycle repetition characteristics of passenger flow are analyzed respectively. Based on this, the impact of factors other than temporal and spatial factors on passenger flow, such as unexpected events and weather factors, is judged, and the eigenvalues affecting the change in passenger flow are extracted and used for input into the next passenger flow prediction model.

(3) Predict the short-time passenger flow of rail transit by using the forecasting method and compare the results. The values of the previously extracted features affecting passenger flow changes are selected and input to the machine learning model, which is divided into a training set and a test set for short-time passenger flow prediction. In the prediction method, the improved LSTM algorithm is used for prediction in this paper. The results show that the fusion of multiple data and external factors (weather, contingencies, etc.) can improve the prediction accuracy very well. For the same prediction object, the improved LSTM prediction is better than the traditional method, and can be adapted to more unexpected events when predicting passenger flow with complex and volatile situations.

Future research can consider more factors and machine learning methods along with practical applications to refine the research model.

7 REFERENCE

[1] Guo, J., Huang, W., & Williams, B. M. (2014). Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C: Emerging Technologies*, *43*, 50-64. https://doi.org/10.1016/j.trc.2014.02.006

[2] Jiang, X. (2016). Short-time traffic flow prediction based on chaos theory and data fusion. *Chongqing University of Posts and Telecommunications*, 2016.

[3] Zhang, C. H., Song, R., & Sun, Y. (2011). Short-time passenger flow prediction at bus stops based on Kalman filter. *Transportation System Engineering and Information*, *11*(4), 154-159.

[4] Xiong, J., Guan, W., & Sun, Y. X. (2013). Kalman filter-based prediction of subway interchange passenger flow. *Journal of Beijing Jiaotong University*, *37*(3), 112-116.

[5] Cai, C. J., Yao, E. J., Wang, M. Y. & et al. (2014). Passenger flow prediction of urban rail transit inbound and outbound stations based on multiplicative ARIMA model. *Journal of Beijing Jiaotong University*, *38*(2), 135-140.

[6] Qin, L. N. & Dong, L. X. (2021). Research on passenger flow prediction algorithm for urban rail transit. *Transportation Engineering*, *21*(1), 40-47.

[7] Milenković, M., Švadlenka, L., Melichar, V., Bojović, N., & Avramović, Z. (2018). SARIMA modelling approach for railway passenger flow forecasting. *Transport*, *33*(5), 1113-1120. https://doi.org/10.3846/16484142.2016.1139623

[8] Davis, G. A. & Nihan, N. L. (1991). Nonparametric regression and short-term freeway traffic forecasting. *Journal of Transportation Engineering*, *117*(2), 178-188. https://doi.org/10.1061/(ASCE)0733-947X(1991)117:2(178)

[9] Lin, P. Q., Chen, L. T., & Lei, Y. W. (2018). Short-time prediction of subway passenger flow based on K-nearest neighbour pattern matching. *Journal of South China University of Technology* (*Natural Science Edition*), *46*(1), 56-63.

[10] Huan, N., Xie, Q., & Ye, H. X. (2018). Real-time prediction of inbound passenger flow of urban rail based on improved KNN algorithm. *Transportation Systems Engineering and Information*, *18*(5), 125-132.

[11] Qin, Y. F. (2019). Short-time passenger flow prediction of urban rail transit stations based on AFC data. *Beijing Jiaotong University*. https://doi.org/10.1061/9780784483053.292

[12] Li, R. Y. (2019). Short-time prediction of OD passenger flow in urban rail transit system based on improved spatio-temporal LSTM model. *Beijing Jiaotong University*.

[13] Polson, N. G. & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1-17. https://doi.org/10.1016/j.trc.2017.02.024

[14] Tian, Y. & Pan, L. (2015). Predicting short-term traffic flow by long short-term memory recurrent neural network. *2015 IEEE international conference on smart city/SocialCom/SustainCom (SmartCity)*, 153-158. https://doi.org/10.1109/SmartCity.2015.63

[15] Impedovo, D., Dentamaro, V., Pirlo, G., & Sarcinella, L. (2019). TrafficWave: Generative deep learning architecture

for vehicular traffic flow prediction. *Applied Sciences*, *9*(24), 5504. https://doi.org/10.3390/app9245504

[16] Shi, X. R., Wang, C. H., Liu, D. J., Zhang, X., & Zhang, B. (2020). Deep neural network-based rail traffic passenger flow prediction and visualization. *Electronic Technology and Software Engineering*, *2020*(19), 182-185.

[17] Wang, X. X., Xu, L. H., & Chen, K. X. (2019). Data-driven short-term forecasting for urban road network traffic based on data processing and LSTM-RNN. *Arabian Journal for Science and Engineering*, *44*, 3043-3060. https://doi.org/10.1007/s13369-018-3390-0

[18] Wang, Q. W., Chen, Y. R., & Liu, Y. C. (2021). Short-time passenger flow prediction for urban rail transit based on convolutional long and short term memory neural network. *Control and Decision Making*, *36*(11), 11.

[19] Huang, W., Song, G., Hong, H., & Xie, K. (2014). Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, *15*(5), 2191-2201. https://doi.org/10.1109/TITS.2014.2311123

[20] Li, X. C., Peng, Y. Z., Wu, Z. X., & Chen, Z. W. (2020). Short-time prediction of passenger flow at subway stations based on deep spatio-temporal network. *Traffic and Transportation*, 33(S2), 55-61.

[21] Jia, Y., Jia, Y., Wu, J., Ben-Akiva, M., Seshadri, R., & Du, Y. (2017). Rainfall-integrated traffic speed prediction using deep learning method. *IET Intelligent Transport Systems*, *11*(9), 531-536. https://doi.org/10.1049/iet-its.2016.0257

Contact information:

Jing XUAN
Beijing Infrustructure Investment Inc.,
Beiing 100101, China
E-mail: xuanjing@bii.com.cn

Jiulin SONG
Northeastern Universit, Boston, America
E-mail: song.jiu@northeastern.edu

Jingya LIU
School of Economics and Management, Beijing Jiaotong University,
Beijing 100044, China
E-mail: 19120615@bjtu.edu.cn

Qiuyan ZHANG
School of Traffic and Transportation, Beijing Jiaotong University,
Beijing 100044, China
E-mail: qyzhang1@bjtu.edu.cn

Gang XUE
(Corresponding author)
School of Economics and Management, Tsinghua University,
Beijing 100084, China
E-mail: xuegang@sem.tsinghua.edu.cn