Dean&Francis
ISSN 2959-6157

The subway's short-term passenger flow prediction during Morning rush hour base on ARIMA —— using Hangzhou city as an example

Dongyang Li[1*]

[1]Architecture and Agriculture Institute, Sichuan Agricultural University, Sichuan, 610000, China

Abstract:

With the development of city's population boost rapidly, people's demand on public transport rising quickly, pushing great pressure to the whole public transport system in the city. In order to relieving urban traffic, public transport like subway has increased rapidly to face this situation. This article will collect passenger flow data during morning rush hour from Hang zhou during January 1-7, 2019, testing the stationarity of data by using Scatter plot, Sequence Diagram and Augmented Dickey-Fuller test (ADF) and using the ARIMA model for prediction. The result shows that through changing different parameters to build model, analysis data and comparing the result, researcher can choose the most accurate prediction model among all these models. Thefinal result shows that the model formed in this paper can make accurate prediction result of passenger flow, which can be used as a reference indicator in citizens' route planning to help them avoid traffic jams.

Keywords: ARIMA Model, Passenger Flow Prediction, Time Series Analysis, Subway Transportation, Urban Traffic Management

1 Introduction

Nowadays, with the development of city's population boost rapidly, leaving great pressure on urban traffic system due to people's increasing demand on public transport. Among all the public transport, subway becomes the most popular urban public transportation not only because it is fast and convenient, but also subway are more friendly to city's environment comparing with others.

So far, many big cities like Shanghai or Beijing are trying to construct a huge and effective subway network system cover the whole city to support citizens' public transportation system, but still can't meet citizens' demand growing year by year: in 2019, the passenger flow of subway line 6 in Beijing has reached 63.2 thousand per hour while the number of which in Shanghai line 11 comes to 59.5 thousand per hour [1]. In order to keep subway station working effectively during special time like rush hour or holidays,

models and algorithms are needed to help the stuff getting an accurate prediction result of short-term passenger flow and avoiding oversaturation which may cause congestion, Inefficient, or even stampede. Also, short-term passenger flow prediction is also meaningful to the safety of the whole subway system's operation as well as management.
Due to the progress in computer science and Big Data field, the intelligent transport system like Advanced Traffic Management System (ATMS) and Electronic Toll Collection (ETC) has quickly developed, research can easily collecting and processing massive transportation data including passenger flow [2]. However, the challenge we have to face is that short-term passenger flow affected by massive factors like the weather, holiday, the time period you choose to study and so on, and the data are random, which means they will be affected easily when the external environment begin to change [3].Through using different algorithms and statistic methods born from the computer science and artificial intelligence, many researchers try to get accurate prediction result using different ways: researcher Zhang et al. using Kalman Filter (KF) model ameliorated by themselves to collect and analyse passenger flow data in Beijing's public transport station [4]; considering random interferential factors, researcher Li utilizing upgraded algorithm model [5]; by using pluralistic data for analyze, researcher Lu et al. try to create multi-source data fusion and genetic wavelet neural network (GA-WNN) to get more accurate result rather than single data resource [6]; researcher Zhang using maximum likelihood estimation (MLE) combined with Autoregressive Integrated Moving Average Model (ARIMA) to predict short term traffic volume data [7].
This article trys to analyze data collected from Hangzhou city's subway system, using Autoregressive Integrated Moving Average Model (ARIMA), which was founded by George Box and Gwilym Jenkins in 1970 [8]. These models are based on statistics and especially effective when handle with short-term Time series prediction with single variable rather than other traditional statistic methods [9]. Auto regressive Integrated Moving Average Model can also be used in other congestion like forecasting the trend in Stock market, for example, researcher Sun et al. Using Autoregressive Integrated Moving Average Model (ARIMA) combine with Radial Basis Function (RBF) to predict the stock movements of the Guizhou Maotai company [10].

2 Methods

2.1 Data Source

This paper uses Hangzhou subway's passenger flow data from January 1-7, 2019 for the study, with objective and accurate data sources, choosing passenger flow's data in one line from 7:00 am to 9:00 am (morning rush hour) for analysis and prediction.

2.2 Indicator Selection and Description

After pre-processing, two variables are been kept: time period and passenger flow. According to table 1 below, time period shows the time interval we choose (every 5 minutes) for collecting passenger flow's data; passenger flow data shows the accurate number of the passenger come and leave the metro line during specified time interval.

Table 1. Name and explanation of variables

Full Name	Data type	Explanation
TIME SERIES	Typedef	Duration between one moment and another
PASSENGER FLOW	INT	the number of people come and out in specified time series

2.3 Indicator Selection and Description

Autoregressive Integrated Moving Average Model (ARIMA) is a time series model using historical data to predict the possible trend in the future by finding the autocorrelation and difference among data, discovering the hidden feature in data. Generally, Autoregressive Integrated Moving Average Model including three parts: Autoregressive Model (AR), Differential processing (I) and Moving Average Model (MA). Each of the part has its own features: Autoregressive Model part considering the influence from the past and focusing on dealing with autoregressive part of the time series; Difference processing part is used to eliminate interfering factors in time series and make time series steady; Moving Average Model part can help researcher to avoid interference from the past time data and deal with the Moving Average section part in time series. Through these three parts, Autoregressive Integrated Moving Average Model can easily capture the changing trend in data and handle with data has accidential changes, makes it a perfect method to work out time series' prediction problem.
First, after colllecting and pre - processing data, Stationar-

ISSN 2959-6157

ity tests are needed to ensure the Stationarity of our data. Usually, researcher will use both scatter plot and Augmented Dickey-Fuller test (ADFtest) to get more correct result. To ensure the Autoregressive Integrated Moving Average Model produce an accurate prediction result, three important parameters must be confirmed: d, p and q. parameter d usually regarded as the order of the Differential processing part, helping series to become stable, the formula below shows how it works in the Differential processing:

$$d\_order\_y = (1-B)^d y_t \quad (1)$$

For this article, the Augmented Dickey-Fuller test's result can give scientific result on the value of parameter d.

The parameter p represent Autoregressive, describing hysteretic value before observation value, the formula below shows how it works in the Autoregressive Model:

$$AR: Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_q Y_{t-q} + \delta_t \quad (2)$$

In this formula, c is considered as constant, ϕ_q refer to model parameter and δ_t related to random time series during analysis. Parameter q related to Moving Average, describing hysteretic value after mistaken part.the formula below shows how it works in the Moving average Model:

$$MA: Y_t = \mu + ?_t + \theta_1 ?_{t-1} + \theta_2 ?_{t-2} + ... + \theta_q ?_{t-q} \quad (3)$$

In this formula, parameter μ is considered as constant and θ_q refer to model parameter. By comparing and analyising Autocorrelation function (ACF) histogram and Partial Autocorrelation function (PACF) histogram, an accurate value range of parameter p and parameter q can be selected. In order to optimize the model and get more accurate result, different combinations of parameter p and q are chosen to formed Autoregressive Integrated Moving Average Model and making prediction. After comparing the result, a more accurate prediction model for passenger's flow has been founded.

The figure 1 below shows the whole process of whole process of passenger flow prediction by using ARIMA from the Stationarity tests to model assess and prediction.

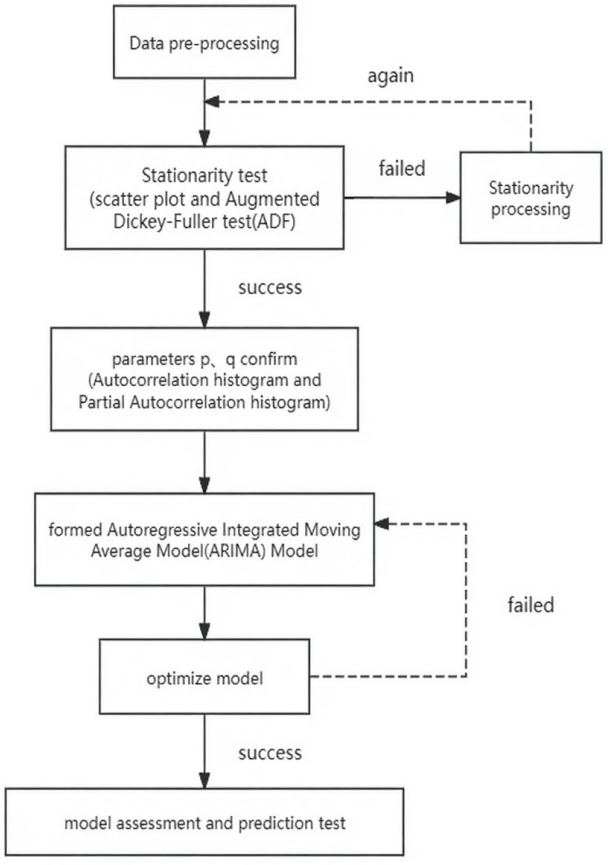

Fig. 1. whole process of passenger flow prediction by using ARIMA

3 Result and Discussion

3.1 Scatter plot, Sequence Diagram and Augmented Dickey-Fuller test

First of all, the original data is analyzed for stability by making scatter plot using passenger flow data from 7:00 am to 9:00 am January 1-7, 2019 ,and these data were collected in every 5 minutes. Usually, Non-stationary series data shows a continued upward trend in scatter plot. in order to get smoothly data for prediction, the passenger flow data need to be transformed to ensure data's stability. Beside the scatter plot, the Sequence Diagram are also needed before model foundation since it can help researchers to figure out the trend of passenger flow data. Figure 2 and Figure 3 below shows the Scatter plot of passenger flow data and the Sequence Diagram after been transformed.

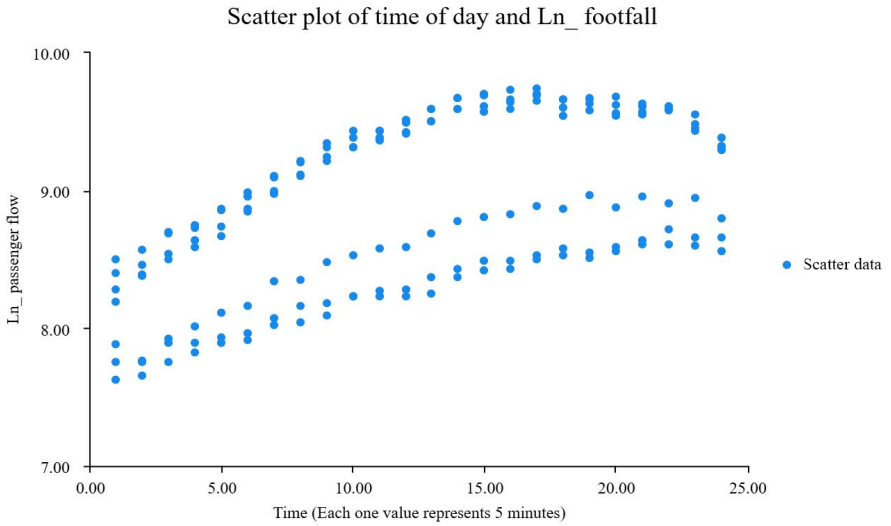

Fig. 2. the Scatter plot of passenger flow data

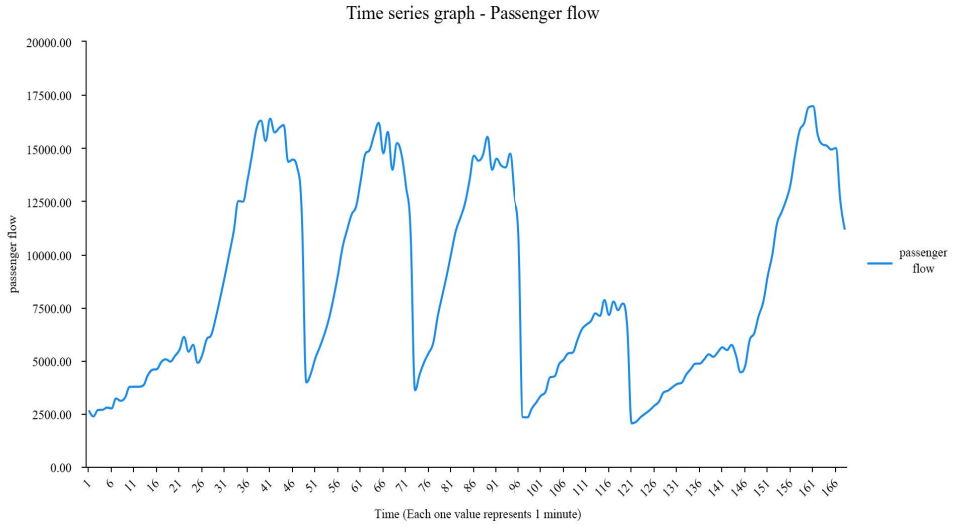

Fig. 3. the Sequence Diagram of passenger flow data

The value in figure 2 and figure 3's horizontal line means each one is equal to 1 or 5 minutes, and the value in vertical part shows the result of passenger flow data after the logarithmic operation to make data smoothly.

Also, in order to verify whether the data after pre-processing are smooth series, a Augmented Dickey-Fuller (ADF) test is needed. The purpose of Augmented Dickey-Fuller test is to make the stability by testing whether there is a unit root in time series data. If it has, it means that the data is unsmoothly, and it can't been used in next step's Autoregressive Integrated Moving Average Model(ARIMA) for analysis.The table 2 below shows the test's result.

From the ADF test results, the paper can notice that there is a certainty of rejecting the null hypothesis with greater than 99% .The result shows that the data after the first-order difference is certainly a smooth data, so the parameter d = 1.

Table 2. Results of ADF test

difference order	t	p	Critical value		
			1%	5%	10%
0	-2.995	0.035	-3.47	-2.879	-2.576
1	-11.275	0	-3.47	-2.879	-2.576
2	-8.52	0	-3.472	-2.88	-2.576

ISSN 2959-6157

3.2 Model foundation

After the Augmented Dickey-Fuller test, the next step is to identify the suitable value of parameter p and q before model foundation. To solve this problem, Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) are needed because through figure out truncated and trailing characteristics in Autocorrelation function and Partial Autocorrelation function can we analyze the autocorrelation of the time series used in ARIMA model,which helps us make sure the range of parameter p and q. The figure 4 and figure 5 below shows the result of ACF and PACF, from which we can indicate that the suitable value for parameter p = 0,1,2, and for q = 0,1,2.

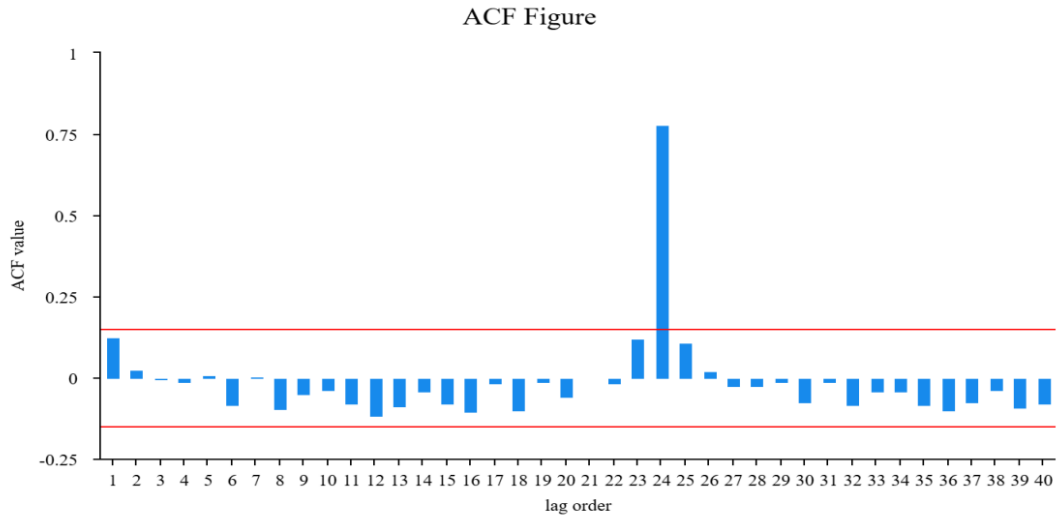

Fig 4. the result of ACF

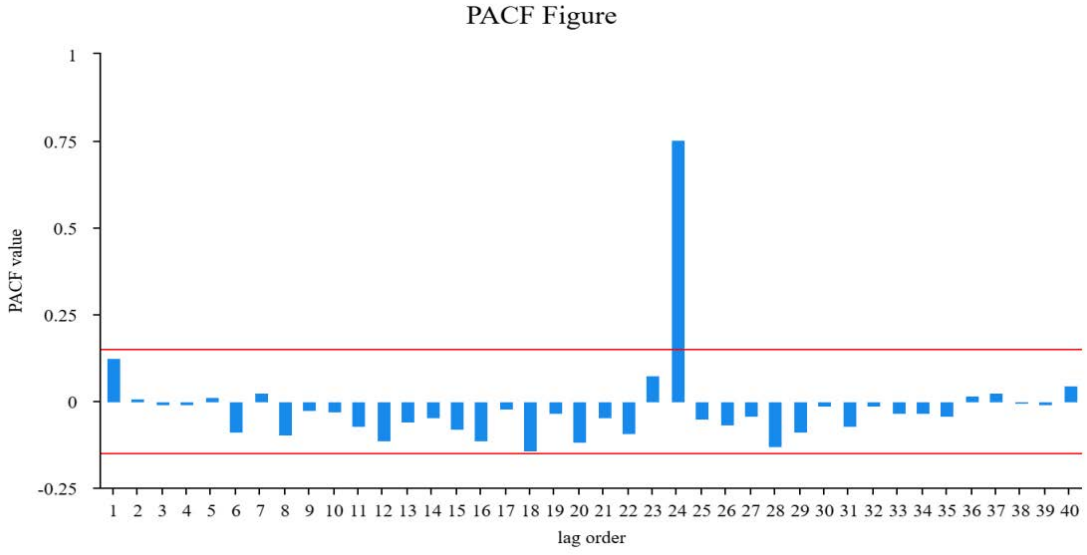

Fig 5. the result of PACF

Through different combinations of parameter p and q to formed ARIMA model, we received 9 different types of ARIMA model in total.In order to choosethe most accurate one among all these models, model parameters including Coefficient, Standard Error, value Z, value P, Akaike information criterion (AIC), Bayesian Information Criterion (BIC) and Root Mean Square Error (RMSE). The table 3 below shows the result of summary for model parameters from different ARIMA models.

Table 3. Summary for model parameters

(p,d,q)	coefficient	standard error	z value	p value	Akaike information criterion (AIC)	Bayesian Information Criterion (BIC)	Bayesian Information Criterion (RMSE)
(p=0,d=1,q=0)	0.009	0.041	0.211	0.833	-39.614	-33.378	0.2124
(p=0,d=1,q=1)	0.008	0.047	0.18	0.857	-40.215	-30.861	0.2107
(p=0,d=1,q=2)	0.008	0.048	0.175	0.861	-38.318	-25.846	0.2106
(p=1,d=1,q=0)	0.007	0.041	0.177	0.860	-40.299	-30.945	0.2106
(p=1,d=1,q=1)	0.007	0.039	0.174	0.862	-38.307	-25.835	0.2106
(p=1,d=1,q=2)	0.009	0.075	0.121	0.904	-36.319	-20.729	0.2106
(p=2,d=1,q=0)	0.007	0.041	0.175	0.861	-38.309	-25.837	0.2106
(p=2,d=1,q=1)	0.008	0.186	0.041	0.967	-36.310	-20.720	0.2106
(p=2,d=1,q=2)	0.011	0.067	0.157	0.875	-34.802	-16.094	0.2103

The value of Root Mean Square Error (RMSE) usually used as the index for measuring deviation between predicted value and actual value. Through the value of the Root Mean Square Error, researcher can easily access the model's accuracy. So, according to table 3, the best model choose for prediction is the one when parameter p = 2, and the parameter q = 2.

3.3 Result of prediction

Through using the the parameters of the ARIMA model as ARIMA (2,1,2), we successfully received passenger flow prediction in each 5 minutes from January 1-7, 2019. The figure 6 below shows a comparison between the predicted result and the true result.

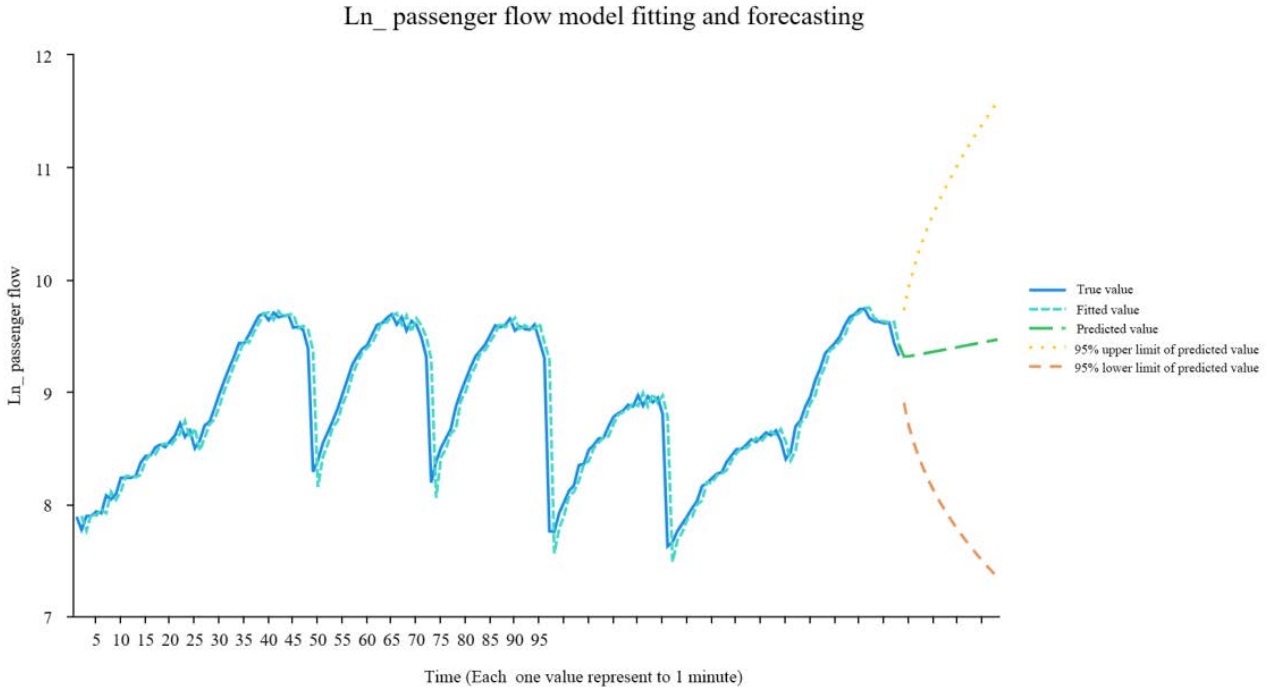

Fig 6. comparison between predicted data and true results

Figure 6 clearly shows that the prediction results of the ARIMA (2,1,2) model basically matched the actual data's trend, and the fit seems good. The result indicate that after optimizing, this model has a high accuracy to predict the subway's passenger flow during morning rush hour (7:00 am to 9:00 am) in Hangzhou City.

ISSN 2959-6157

4 Conclusion

This paper shows a complete process of making passenger flow's prediction through the ARIMA model. By using Data Visualization technology like scatter plot, Autocorrelation function (ACF), Partial Autocorrelation function (PACF) and so on to ensure data's stability and optimize the model. The prediction result has demonstrated that the true results and predicted data obtained from the ARIMA model fits well. However, there still some disadvantages and limits while using ARIMA model to make prediction for subway's passenger flow: First, the ARMIA model can's process effectively when facing with massive time series data because it takes a long time for training and optimizing the algorithm, in order to deal with this situation, researchers often combine ARIMA model with other algorithms like the Radial Basis Function (RBF), Support Vector Regression(SVR) and so on for prediction; second, ARMIA models are sensitive to outliers and confounders, it can't make accurate prediction result if the time series data is affected by outliers and confounders; third, if considering influencing factors like weather and holiday during prediction, the result from ARMIA model may not as accurate as researcher's consideration, which means a more complex and effective model like Long Short Term Memory(LSTM) system, Kalman Filter (KF) model or hybrid model are needed to face with complex environment. In the future, these algorithms and models can be widely used in apps like the AutoNavi Maps to make traffic prediction for passenger to help them avoid traffic jam and optimize their travel routes planing.

References

1. H. Chang, Research on short-term passenger flow prediction of subway based on LSTM neural network, Master Thesis, Xijing University, College of Computer Science (2022)

2. Q. Zhou, Short-term passenger flow prediction of subway based on combined features, Master Thesis, Chongqing Technology and Business University, International scientific and technological cooperation base for intelligent manufacturing services (2020)

3. L. Hai, W. Liu, Y. Liu, et al, Prediction of subway passenger flow based on ARIMA algorithm, Comput. & Digital eng. 53, 666-670, (2025) 10. 3969/j. issn. 1672-9722. 2025. 03. 010

4. Z. Zhang, D. Zhang, J. Jia, et al. Prediction of short-term passenger flow of rail transit platform based on improved Kalman filter, J. of Wuhan University of technol. (Transportation Science & Engineering), 41, 974-977, (2017) 10.3963/j. issn. 2095-3844. 2017. 06. 017

5. Z. Li, Research on Short-term Passenger Flow Prediction of Urban Rail Transit Basedon Multi-feature Fusion, Master Thesis, Southwest Jiaotong University, College of Transport and Logistics (2020)

6. B. Lu, Q. Shu, G. Ma, et al, Short-term traffic flow prediction based on multi-source traffic data fusion, J. of Chongqing Jiaotong University(Nat. Sci.). 38, 13-19, (2019) 10.3969 / j. issn. 1674-0696. 2019. 05. 03

7. T. Zhang, P. Yuan, Short-term traffic volume prediction model based on ARIMA, Int. Comput. and Application, 10, 273-278, (2020)

8. Z. Zhang, N. Chen, B. Zhu, et al, Analysis of PM2.5 Sources in Wuhan City Based on the Random Forest Model, EVS, 43, 1151-1158, (2020) 10.13227/j.hjkx.202108051.

9. Y. Zhang. Research and application of time series analysis based on ARIMA-LSTM hybrid model, Master Thesis, Yangtze University, College of Computer Science (2023)

10. D. Chen, F. Du, H.Xia, et al, Stock prediction based on the combination of ARIMA and SVR rolling residual model, Comput. Era, 05, 76-81, (2022) 10.16644/j.cnki.cn33-1094/tp.2022.05.019.