

Research Article

Influential Factor Analysis and Prediction on Initial Metro Network Ridership in Xi'an, China

Tao Lyu,¹ Mingfei Xu,¹ Jia Zhang,² Yuanqing Wang ,¹ Liu Yang,³ and Yanan Gao^{1,4}

¹Department of Traffic Engineering, College of Transportation Engineering, Chang'an University, Xian, China

²Xi'an Rail Transit Group Co., Ltd, Xian, China

³Department of Urban Planning, College of Urban and Environmental Sciences, Northwest University, Xian, China

⁴Urban Planning and Transportation Group, Eindhoven University of Technology, Eindhoven, Netherlands

Correspondence should be addressed to Yuanqing Wang; wyqing@chd.edu.cn

Received 7 September 2021; Revised 10 February 2022; Accepted 23 February 2022; Published 5 May 2022

Academic Editor: Rocío de Oña

Copyright © 2022 Tao Lyu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To satisfy the adaptability of forecasting the short-term and abrupt volume of the initial metro network, we build the multiple enter linear regression (MELR) model to explore the determinants and forecast the intensity during the twice expansion of the initial metro network in Xi'an. We further compare the prediction of the metro transport capacity between the MELR models with exponential smoothing and autoregressive integrated moving average (ARIMA) models. Results show that the passenger intensity significantly fluctuates with the months and days, and MELR model is more adapted for the short-term prediction of the abrupt volume than the ARIMA model during the new metro line opening and the old line expands, which avoids the drawback of time series models that need a huge database. This study provides a guide for the prediction of initial metro network volume and accurate purchase of the rail vehicles during the metro planning and expands stages.

1. Introduction

With the rapid development of urbanization, the metro becomes the mainstream of public transportation and is a powerful countermeasure to reduce urban traffic congestion and build a low-carbon transit system, on account of its advantages of high efficiency, large capacity, less land occupation, and convenience [1, 2].

This study defines the urban initial metro lines are the first and second operated lines, and the network formed by these two lines is called the initial metro network. They were approved in the first phase of metro construction. The passenger volume collected in this period is the initial metro passenger volume. There are two obvious characteristics of the initial metro passenger volume: (1) the extreme volume appearing during the holidays and (2) as the operation time continues, the gap of passenger volume between stations gradually narrows, except for the hot stations. The metro network passenger volume intensity was defined as the ratio of daily passenger volume and the operation mileage, which is an important indicator to reflect the operation efficiency of the network [3].

When there is a lack of scientific forecast of passenger volume, high metro construction standards will lead to investment waste, while low construction standards cannot meet traffic demand. The planning of the metro should be based on the maximum capacity of the station or line, irrespective of the stage of its construction. The prediction of the metro volume or carrying capacity was based on the land-use intensity of the city. Then, the potential rail vehicle quantity can be decided by the predicted volume. This traditional method used in the Project Feasibility Study Report may meet challenges during the metro network operation, namely, the land-use characteristics and intensity will change with the operation of the metro. It will lead to passenger volume increase around the metro lines and stations [4–7]. The metro passenger volume always fluctuates in different months and on different days. Meanwhile, the opening and expansion of the new and old metro lines will cause the volume experiencing an abrupt increase. Subsequently, as the metro operation days increase, the transport capacity of the metro may not be able to meet the demand of the actual metro passenger volume [8]. We must

adjust the time schedule or the operation interval of the metro to satisfy the needs. Based on this, there exist two vital issues before initial metro network transit system planning and construction: (1) analyzing the metro lines and networks' passenger volume intensity and their spatial-temporal distribution characteristics, and the influential factors, and (2) research on the accurate prediction method of metro lines and networks' passenger volume intensity [2, 9], which is part of the short-term prediction. The most commonly used short-term forecast method is the time series model [10], but it cannot characterize the jumpy variation of volume in the crucial date. In view of the characteristics of a sudden change of passenger volume in the pivotal construction stage of the initial metro network, it is necessary to find a valid short-term passenger volume forecasting method. But the analysis on metro daily volume prediction focuses on the metro station-level, which cannot reflect the operation efficiency of the network, lacking the studies based on metro lines and networks to characterize the abrupt change of the volume at different operation stages [10]. Furthermore, the significant temporal variation in the initial metro volume has seldom gotten attention and been verified theoretically. So, to fill in the research gap, taking the initial metro network in Xi'an as an example, aiming at the twice expansion of the network, we build the multiple enter linear regression (MELR) model to explore the temporal determinants of the metro volume intensity and applied MELR in forecasting short-term metro volume intensity. The MELR model is the extension of the ordinary least square (OLS) regression model, and it has the strength of understanding the relationship between multiple independent and dependent variables easily [5, 11]. Because the optimized time series model has high accuracy in the short-term prediction of passenger volume, 8 time series regression models were compared with MELR for daily passenger volume intensity prediction in the line operation stage (Line-S: from September 16, 2011 to September 14, 2013), initial network stage 1 (Network-S1: from September 15, 2013 to June 15, 2014), and initial network stage 2 (Network-S2: from June 16, 2014 to May 31, 2015). Furtherly, we compare the prediction of the transport capacity between the MELR and Autoregressive Integrated Moving Average (ARIMA) model. The time series regression models include the Simple Nonseasonal Exponential Smoothing model (SNES), Holt Nonseasonal Exponential Smoothing model (HNES), Brown Nonseasonal Exponential Smoothing model (BNES), Damped Nonseasonal Exponential Smoothing model (DNES), Simple Seasonal Exponential Smoothing model (SSES), Winters Additive Exponential Smoothing model (WAES), Winters Multiplicative Exponential Smoothing model (WMES), and ARIMA model. Results find that the concise and valid MELR model is more adapted for the short-term prediction especially toward the jumpy volume than the ARIMA model. The results provide a rapid and convenient way for initial passenger volume intensity prediction. It can guide urban metro plan, design, facility equipment configuration, metro train purchase, and operation

schedule management in the metro planning and expand stages in urban cities [10].

The major contribution of this study can be summarized as following three aspects:

- (1) We discover the temporal influential factors and predict the initial metro volume intensity in different operation stages in Xi'an initial metro network, instead of at the metro station level.
- (2) Furthermore, the initial metro network intensity on each expansion stage is easily influenced by the temporal determinants including the month of the year, day of the week, and the days since the metro opened.
- (3) Compared with the ARIMA model, the MELR model has higher accuracy and is more adapted to the influential factors' exploration and the short-term prediction of the jumpy initial metro passenger volume.

This paper is organized as follows. Section 2 reviews literature referring to methodologies of short-term metro passenger volume prediction and influential factors. Section 3 describes the research flowchart, dependent and independent variables, and the MELR and ARIMA models. The regression and prediction results were shown and discussed in Section 4. Finally, the conclusions are given in Section 5, including a summary of findings and limitations, as well as the future works.

2. Literature Review

2.1. Prediction Methodologies. In terms of passenger volume, forecasting is mostly based on the traditional four-step (trip generation, trip distribution, mode split, and trip assignment) method, multiple regression model, or their variants [12–14]. The four-step method is more applicable to the long-term passenger volume prediction on a regional or traffic zone scale at the planning stage, and a large set of explanatory variables need to be calibrated [12, 14, 15]. Meanwhile, there are some difficulties in data collection, authenticity, and reliability. The defects involve low accuracy and response, imprecise data, insensitivity to land use, institutional obstacles, and high cost [15]. It consequently cannot meet the accuracy requirements of the short-term passenger volume prediction.

Recently, the direct demand model has gotten attention as an alternative to the traditional four-step model. It is advantageous and makes up for deficiencies of the traditional four-step model. It estimates ridership via regression models and treats ridership as a function of its influencing factors which helps to predict the rail passenger volume [4, 5]. The OLS regression is the widely used direct demand model, and it assumes that the parameters are stable [5, 11]. Liu et al. [11] proposed the Direct Ridership Model (DRM) which can provide an estimation of rail station ridership without relying on a complicated transportation demand model and extensive data collection. He et al. [1, 8] systematically reviewed and summarized the related studies on direct demand models for metro ridership prediction.

Compared with the traditional four-step demand forecasting method, the main strength of the direct demand model in ridership modeling is simple usage, easy interpretation, quick response, and low expense [8].

From another classification perspective, the short-term traffic passenger volume forecasting methods can be generally divided into two categories: parametric and nonparametric regression algorithms. The characteristics of widely used models were summarized by Vlahogianni et al. [16]. They provide a comprehensive thought before building short-term volume forecast methods. The parametric regression algorithm includes the time series forecasting model [17], linear regression model [18], ARIMA model [19, 20], and so on. The nonparametric regression algorithm includes Kalman filtering model [21], support vector regression (SVR) [22–24], neural network model (NN) [3, 16, 25, 26], Genetic Algorithm (GA) [27], and so on. And, robust statistics should work well on both parametric and nonparametric methods to avoid their misuse [17]. To give intuitive and reliable results for readers, researchers are accustomed to making a comparison between their proposed method with the traditional parametric and nonparametric methods, in terms of the study on metro passenger volume prediction.

Besides the methods mentioned above, some research tried to integrate both the parametric and nonparametric methods to achieve better performance. A hybrid EMD-BPN forecasting approach combines empirical mode decomposition (EMD) and back-propagation neural network (BPN). Results showed that the proposed approach performs well and stably in forecasting the short-term metro passenger volume in Taipei; the prediction accuracy of the neural network is better than that of the ARIMA model and Seasonal Autoregressive Integrated Moving Average (SARIMA) model [9]. But, the problem of mode mixing caused by the intermittency of metro volume is reducing the predicted capability of EMD-BPN [9]. Sun et al. [23] constructed a hybrid method of wavelet and SVM to predict Beijing subway passenger volume, especially in the morning and evening peak hours. Results found that it is the most promising and robust method among Wavelet-NN and EMD-NN by overcoming the shortcomings of Wavelet and SVM, respectively. Compared with several conventional statistical algorithms and computational intelligence algorithms in the emergent event, Li et al. [26] concluded that the Artificial Neural Network (ANN) model has the highest accuracy and shortest training time in evaluating passenger volume, but it does not take the transfer passenger demand from neighboring bus stops into consideration. The mathematical and neural network models (ANN, long short-term memory (LSTM)) used to predict metro stations' passenger volume in Qingdao found the LSTM model to be better and suitable for capturing the long-term and short-term characteristics of metro passenger information [6]. Based on a deep recurrent neural network (DRNN), a time series prediction model was proposed for the short-term metro passenger volume prediction in Shanghai, and it has good robustness compared with the traditional SVR and the BPN method, in the processing of time series data [2].

However, we also found some shortcomings in traditional forecasting methods. For example, the Artificial Neural Network (ANN) trains models by using a large amount of historical data to obtain a more accurate mapping relationship between output and input, so it has a strong dependence on the data [28]. It always has issues with overtraining, local optima, and a high computational burden. But, the SVR is an alternative to ANN for short-term prediction problems when the amount of data is less or when the training data was not a good representative sample of the testing data [24]. In addition, understanding the results is a major challenge in terms of the interpretability of the function modeled by the machine-learning algorithm.

Besides, the time series method is based on the trend of available operational historical data, combined with the current situation to calculate the growth coefficient through various regression analyses (e.g., exponential smoothing model and multiple regression model). Particularly, ARIMA has become one of the common parametric forecasting approaches since the 1970s. Rui et al. [29] found that the time series model has defects in predicting the short-term passenger volume but it is more suitable for predicting the long-term passenger volume. There also have some studies demonstrating that ARIMA is superior to other models in short-term metro volume forecasting. By investigating the effect of temporal and spatial features, as well as the influence of weather on metro station passenger volume, the time series model ARIMA, linear regression, and SVR are employed for forecasting short-term passenger volume in Shenzhen metro stations [18]. In Beijing, an ARIMA model was established to carry out the short-term prediction of rail transit stations' passenger volume, which has high prediction accuracy and can characterize periodic changes in time series data [20]. By comparing with the autoregressive (AR) model, SVR, and back-propagation (BP) neural network model, results show that the optimized time series model of ARIMA has high forecast accuracy in the short-term prediction of passenger volume in Xi'an metro [19]. ARIMA performs well and robustly in modeling linear and stationary time series. These studies give us the inspiration to conduct analysis on the initial metro line and network volume intensity prediction in Xi'an by using the ARIMA model.

In summary, the aforementioned research widely focuses on making comparisons between parametric and nonparametric regression algorithms, to determine what method has the best prediction accuracy and robustness. Although these methods can capture the stable and regular fluctuation of metro passenger volume, they are seldom designed to forecast extremely irregular fluctuation in passenger demand due to the new and old lines opening and expanding which may attract a huge volume. Furthermore, most of the previous methods mainly focus on metro volume prediction at the station-level; there has been little research directly on the initial line or network level [10, 30].

2.2. Influential Factors. The metro passenger volume is affected by different land-use patterns around the metro station [4, 6]. Particularly, Zhao et al. [31] have shown the

relationship of station ridership on Sundays at obvious peaks with different types of land use. Kuby et al. [5] evaluated the impact of the surrounding environment of stations on light rail occupancy in the United States, and proposed significant influencing factors including employment, population, and the rental ratio of the house within light rail walking distance. Liu et al. [11] found that the bus connection at station and employment density have significant impacts on rail station ridership. Other factors affecting metro passenger volume include accidents, crimes, road safety, weather, etc. [32].

Thus, many factors have impacts on the metro station passenger volume [7], which is mainly divided into four categories including land-use variables, station characteristics, socioeconomic and demographic characteristics, and intermodal connection or transit accessibility [33–37]. The OLS model is the most widely used method of ridership modeling and influencing factor analysis [5, 11, 15, 33, 37, 38]. Combined with OLS and other regression models, the analysis on metro volume influencing factors investigation obtained fruitful findings in cities all over the world. Based on OLS regressions, Zhao et al. [38] found 11 variables related to land use, external connectivity, inter-modal connection, and station context may significantly correlate to metro station-level ridership in Nanjing. Furthermore, they utilized OLS and spatial error models (SEM) to explore the determinants of transit ridership, found that land-use type, transit accessibility, income, and density are strongly significant predictors [33].

By multiple regression analyses, Loo et al. [34] found that major interchange rail station and car ownership are significant and positively associated with railway ridership in New York City and Hong Kong. In addition, Wang et al. [36] demonstrated that shopping and recreational factors have a statistically significant relationship with metro trips during the afternoon and evening peak hours in the commercial districts of Hong Kong by linear regression. The regression model also discovered that land-use density and station-level accessibility (the number of bus routes at rail stations, and the number of stations' entrances or exits) are positively related to rail transit ridership in Seoul [7]. Based on the logistic regression model, dwelling density was found to be an important factor in increasing the Hamilton street railway share in Canada [35].

For the sake of inquiring the local and global determinants of metro station-level ridership, the OLS and geographically weighted regression (GWR) were used by Cardozo et al. [15]; they found that the number of metro lines, workers, employment, and suburban bus lines have a statistically positive influence on both models. He et al. [8] explored the local potential influencing factors of the metro station ridership in Shenzhen by the GWR model; they found population, network degree centrality, betweenness, days since metro opening, shopping land use, and distance to the city center have a positive or negative impact on metro station ridership, to a certain extent. An adapted geographically weighted LASSO (Ada-GWL) model was used to explore the influence factors (land use, network structure, social economics, and inter-modal traffic access) of

Shenzhen metro stations' ridership from the spatial perspective, which demonstrates high interpretability and goodness-of-fit [1].

In a word, scholars have done a lot of research on the spatial influential factors of metro station passenger volume and have made fruitful results. The independent variables include land use (residential, restaurant, retail, shopping, office, banks, hospital, and hotels), network structure (distance to the city center, degree centrality, betweenness centrality), social economics (population, days since opened), and inter-modal traffic accessibility (number of bus stations), but it does not specifically consider the temporal variation in the current methodology [1, 8]. Research on investigating the temporal influential factor of the initial metro network passenger volume especially remains insufficient in the literature. In the past, more research focused on passenger volume forecasting and influencing factors at the station level, and few pieces of research were directly aimed at the level of the initial metro networks. So, this study specializes in exploring the temporal determinates on passenger volume intensity based on the MELR model.

3. Dataset and Methodologies

Xi'an is the fourteenth city to build the metro and the thirteenth city to open and operate in China. The population of Xi'an in 2015 was 8,705,600 and there are 1,984,561 employees [39]. The number of metro passengers in 2015 was 342,093,500 (person-times), which took up to 8.66% of the main urban public transportation modes in 2015 [39]. By the end of 2016, its operating mileage ranked 11th among the 28 metro operation cities in China. From September 2011 to November 2016, before the opening of Line 3, it formed the initial network and transferred between Line 1 and Line 2. The network has 39 stations in total. The locations and names of these stations are shown in Figure 1. The development of Xi'an Metro has undergone the following three stages.

Line-S: The 1st phase of Line 2 (BEI KE ZHAN-HUI ZHAN ZHONG XIN) was operated on September 16, 2011, with a length of 20.5 km and 17 stations.

Network-S1: The 1st phase of Line 1 (FANG ZHI CHENG-HOU WEI ZHAI) was operated on September 15, 2013, with a length of 25.4 km and 19 stations. Form an initial network and transfer at the BEI DA JIE station on Line 2.

Network-S2: The 2nd phase of Line 2 (HUI ZHAN ZHONG XIN-WEI QU NAN) was operated on June 16, 2014, with a length of 6.3 km and 4 stations. The total operation length of Line 2 is 26.8 km at this moment.

The data of this study were collected from the Xi'an Metro Limited Liability Company, which is calculated from the automatic fare collection (AFC) transaction data. The data of Line 2 was collected from the daily passenger volume report from September 16, 2011, to May 30, 2015 (1354 days in total). The data of Line 1 were collected from the daily passenger volume report from September 15, 2013, to May

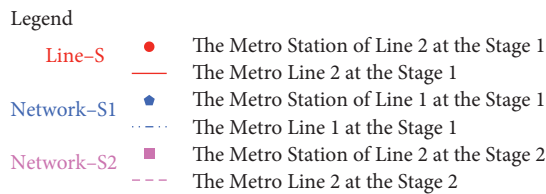


FIGURE 1: Map of the initial metro network in Xi'an.

30, 2015 (623 days in total). So, we divided the research periods into 3 stages: stage 1 (S1) is from September 16, 2011 to September 14, 2013, stage 2 (S2) is from September 16, 2011 to June 15, 2014, and stage 3 (S3) is from September 16, 2011 to May 31, 2015.

To show the research content and steps logically, we first elaborate on the perspicuous research flowchart (Figure 2). The independent variables were identified from the temporal distribution characteristics of daily and monthly passenger volume intensity, including days since the metro opened, months of the year, and days of the week. The dependent variables are daily passenger volume intensity of the initial

network in S1, S2, and S3 stages. The methodologies are mainly used to investigate the influencing factors and volume intensity prediction. The Pearson correlation coefficient of the independent variables in MELR models can avoid multicollinearity with significant factors, and the model explainability can be measured by Adjusted R^2 , Variance Inflation Factor (VIF), and Durbin-Watson (U). MELR model was compared with exponential smoothing and ARIMA models to predict volume intensity in S1, S2, and S3. In the process of prediction, 6 model comparison metrics and 2 model comparison tests were conducted to verify the accuracy and robustness of forecasting models. Finally, we

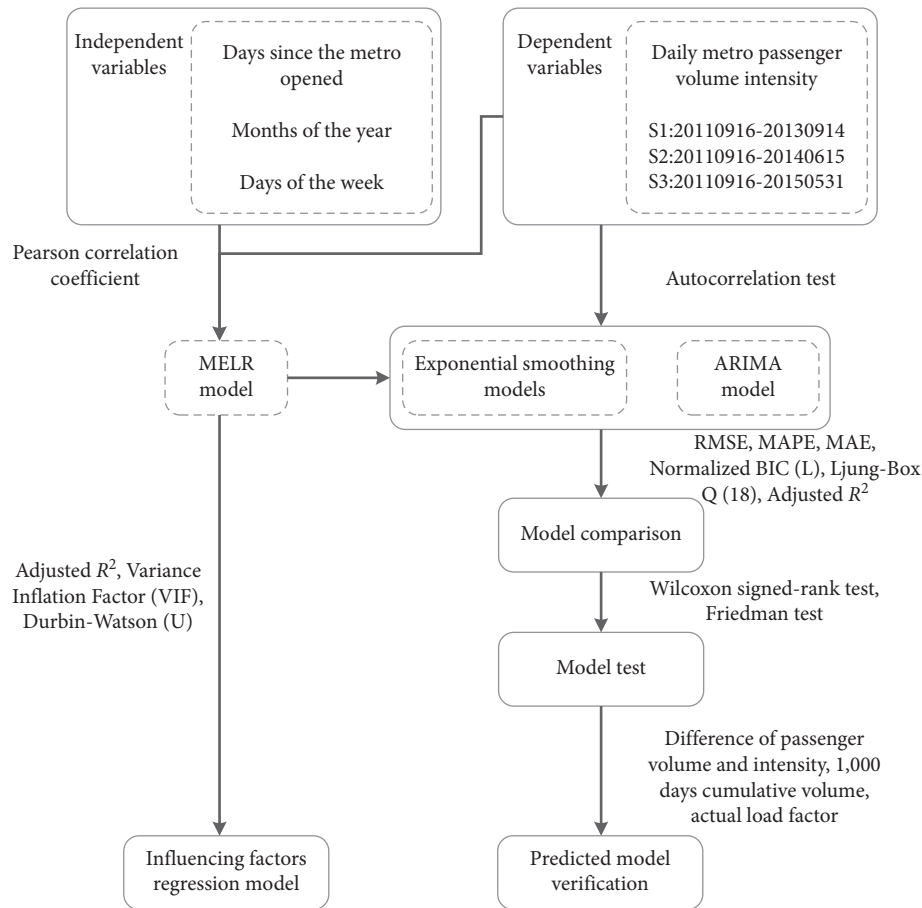


FIGURE 2: The research flowchart.

verified the predicted model by the difference of passenger volume and intensity, 1,000 days cumulative volume, and the actual load factor in 6-carriage B2 trains.

3.1. Dependent and Independent Variables. By exploring the temporal distribution characteristics of metro passenger volume intensity, we can initially determine the candidate independent variables.

3.1.1. Days since the Metro Opened. Figure 3 shows the temporal distribution of the daily passenger volume intensity of the initial network in 3 operation stages. A big change in daily passenger volume intensity exists among Line-S and Network-S1, indicating that the opening of Line 1 in the initial network may bring about an abrupt variation of passenger volume. Then, the extension of the metro Line 2 in Network-S2 leads to the mileage increase in the network; it also brings a great increase in passenger volume intensity in Network-S2. Namely, there exists a linear upward tendency between the accumulated operation days and the volume intensity of the initial network. Thus, “days since the metro opened” can be used to explore the influential factors of the initial metro network passenger volume intensity in Xi’an.

3.1.2. Months of the Year. Months likely lead to changes and function as decision factors of passenger volume intensity. In July and August, the growth of tourists in the summer holidays led to an increase in metro passenger volume intensity. In addition, months with special holidays also can influence passenger volume. In most cases, the Mid-Autumn Festival and National Day are celebrated in September and October, respectively. Therefore, tourists gather rapidly on holidays, resulting in the metro passenger volume intensity growing. In Figure 4, the monthly average daily passenger volume intensity shows a sharp and moderate increase along with the opening and extension of the metro Line 1 and Line 2 in Network-S1 and Network-S2, respectively. The passenger volume intensity has an increase of 67.75% in Network-S1 with the opening of Line 1, then the increasing slowdown in Network-S2 to 20.51%. The monthly average daily passenger volume intensity in Line-S, Network-S1, and Network-S2 were 8601, 14428, and 17387 persons/km*day, respectively. There has been an increasing trend in passenger volume intensity with the months, except for the special months of February and March with Spring Festival involved. The intensity keeps relatively stable around September, so the independent variable of September is chosen as a comparative variable in the MELR model, making the comparison between other months with September in Section 4.1.

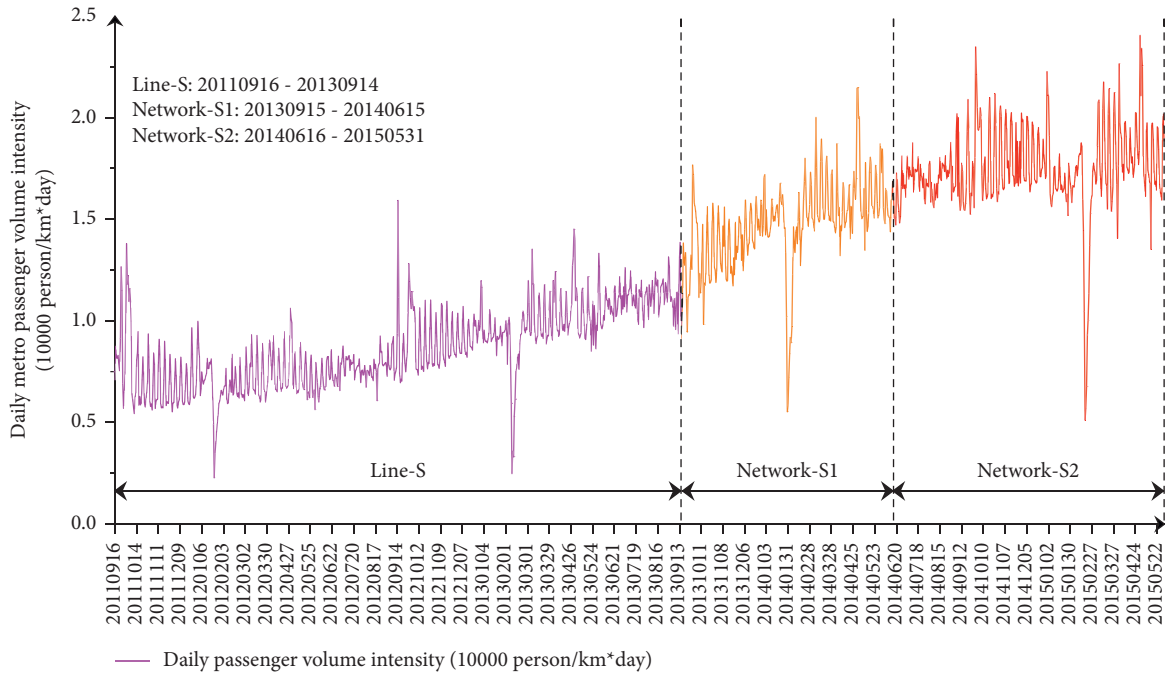


FIGURE 3: The temporal distribution of daily passenger volume intensity.

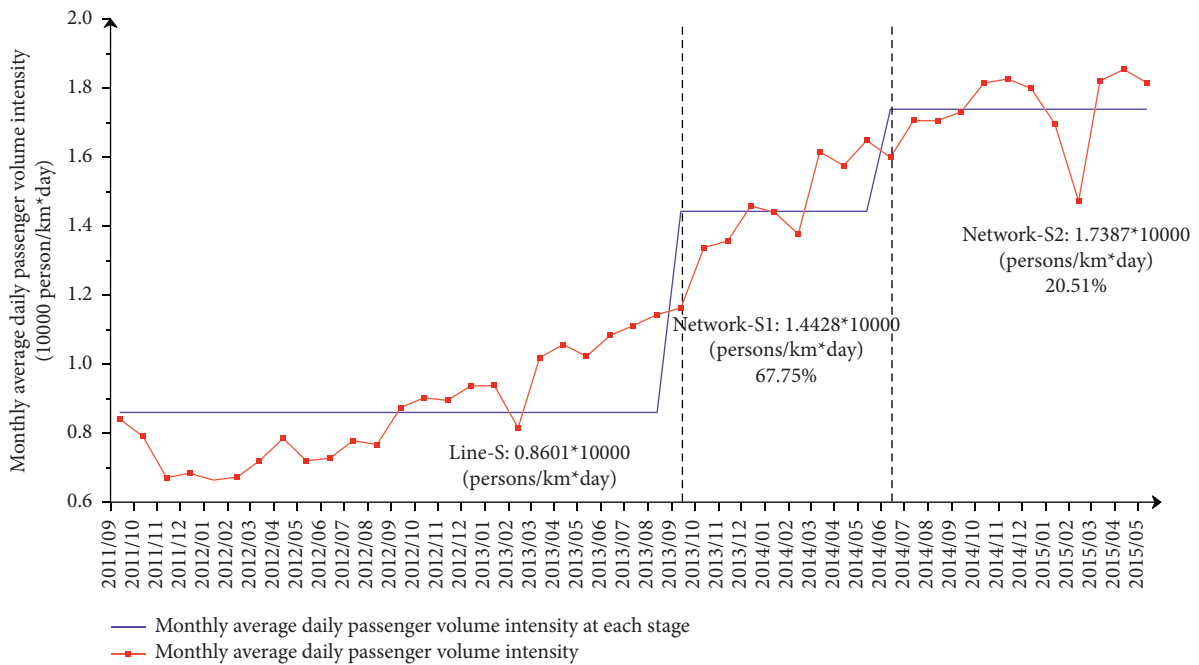


FIGURE 4: The temporal distribution of monthly average daily passenger volume intensity.

3.1.3. *Days of the Week.* The daily passenger volume intensity maintains at a stable level from Monday to Thursday; then it experiences a great increase from Friday to Sunday (Figure 5). Especially, the intensity on Friday is 11.11% higher than Thursday. So, the daily passenger volume intensity often varies in days of the week or between the weekdays and weekend [10]. Xi'an is a tourism city with deep

historical and cultural heritage; citizens tend to take a short tour during weekends, resulting in the increase of metro passenger volume. Therefore, it is necessary to comprehensively consider the influence of temporal factors on ridership intensity and the prediction of metro network passenger volume. It is evident that the daily passenger volume intensity remains stable on Wednesday. Hence, the

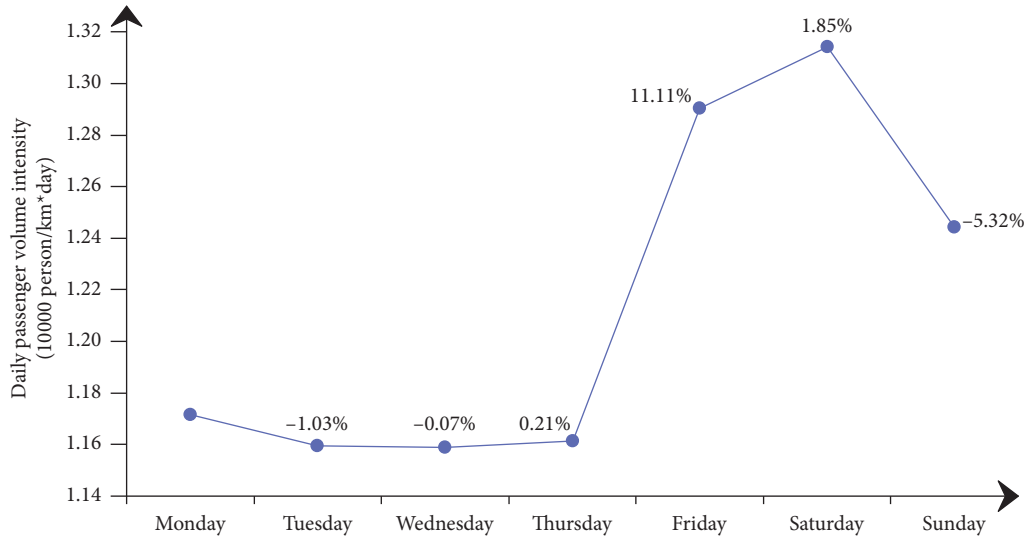


FIGURE 5: The daily passenger volume intensity on weekdays and weekends.

independent variable of Wednesday is chosen as a comparative variable in the MELR model to investigate the temporal influencing factor of intensity in Section 4.1. The regression coefficients of other days can be compared with Wednesday and show the difference of the influencing intensity.

3.2. Methodology. The purpose of this research is to find out the significant influencing variables of the daily passenger volume intensity of initial metro network and predict its short-term volume intensity at S1, S2, and S3. Based on some indices and statistic tests, exponential smoothing and ARIMA models were compared with MELR model for the fitting and prediction accuracy of S1, S2 and S3's daily passenger volume intensity. Specially, all the models were processed with the help of SPSS 23.

3.2.1. MELR Model. The general form of the MELR model is shown in (1). In terms of the response variable in direct demand models, daily ridership has been the most common concern [5, 34, 37, 40]. In order to effectively show the transport efficiency of the initial network, the dependent variable Y indicates the daily passenger volume intensity (10000 persons/km*day) of S1, S2, and S3; there are k independent variables, which are X_1, X_2, \dots, X_k ; β_0 is the constant; $\beta_1, \beta_2 \dots \beta_k$ are partial regression coefficients; and ε is the error term. The estimators of equation (1) can be obtained by using the least square method based on the estimation of the multiple linear regression equation obtained from the sample data.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon. \quad (1)$$

We can find the daily passenger volume intensity mainly influenced by temporal factors in Section 3.1. Firstly, the days since the metro opened can be used to characterize the process of metro volume accumulation. The longer the operation mileage of the subway, the more people it takes.

Secondly, the indicator of periodical variation is expressed by the month of the year and the day of the week.

In the process of regression, the factor of days since the metro opened can be measured, but the day of a week and the month of a year cannot be measured quantitatively. To verify the influence of these unqualified factors on the dependent variables, it is necessary to introduce dummy variables (e.g., 0 or 1). We take two dummy variables of September and Wednesday as reference variables in the MELR models. The regression coefficients of the rest of the independent variables are compared with them. The detailed statistical description of the dependent and independent variables is summarized in Table 1.

3.2.2. ARIMA Model. The traditional statistical time series forecasting approaches include exponential smoothing, moving average, and ARIMA, in which future values are constrained to be a linear function of past observations. The ARIMA model is easy to understand and implement, and is computationally tractable [41]. It has been widely applied in forecasting short-term traffic volume [10, 41, 42]. In this section, the basic model of moving average and ARIMA are briefly reviewed. The ARIMA model originates from the autoregressive (AR) model, moving average (MA) model, and the combination of AR and MA (ARMA) models [43]. For the AR model of order p , known as an AR(p) model, the current value of time series can be expressed based on (2).

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t. \quad (2)$$

The MA(p) model, which expresses the current value of time series as a current and q previous values of random errors, can be expressed as follows:

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}. \quad (3)$$

Thus, the general expression for a combined AMRA(p, q) process can be defined as (4), where y_t is the predicted value, ϕ_i represents coefficients associated with each previously

TABLE 1: Statistic description of dependent and independent variables.

Variable	Notation	Description	Mean	Standard deviation	Min	Max
Daily metro passenger volume intensity of S1	Y_1	Ten thousand person/km * day	0.865	0.195	0.227	1.593
Daily metro passenger volume intensity of S2	Y_2	Ten thousand person/km * day	1.029	0.337	0.227	2.148
Daily metro passenger volume intensity of S3	Y_3	Ten thousand person/km * day	1.214	0.441	0.227	2.405
Days since the metro opened	T_i	Number of days/day ($i = 1, 2, \dots, 1354$)	668	391.0	1	1354
Month	X_j	Month of the year ($j = \text{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, } (X_j \text{ is 0,1 variable})$)	—	—	—	—
Day	X_k	Day of the week ($k = \text{Mon, Tues, Wed, Thur, Fri, Sat, Sun, } (X_k \text{ is 0,1 variable})$)	—	—	—	—

observed value, y_{t-i} are the previously observed values, θ_i are coefficients associated with previous white noises, ε_t is a normal white noise process with zero mean and variance σ^2 , and ε_{t-1} are previous noise terms.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}. \quad (4)$$

Generally, the ARMA model is applied to stationary time series. However, if the series are nonstationary, these series are transformed into a stationary time series using the d th difference process, and the difference d is usually 0, 1, or at most 2 [43]. Therefore, the ARIMA(p, d, q) can be obtained as (5). Where $w_t = \nabla^d y_t$. Note that if we replace w_t by y_t ; that is to say, when $d=0$, (5) represents a mixed ARMA model.

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \dots + \phi_p w_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}. \quad (5)$$

From the above equation, it is clear that the past passenger volume usually influences the present and future volume. We take the independent and dependent variables as same as MELR model for volume prediction in time series models.

3.2.3. Model Evaluation Index. The results of the root mean square error (RMSE), the mean absolute percentage error (MAPE), the mean absolute error (MAE), Ljung-Box Q(18), and Adjusted R^2 were used to examine the model robustness and measure the prediction accuracy and reliability. Bayesian information criterion (BIC) is also a criterion for selecting a more fitted model based on the likelihood function and penalty term. Its penalty is heavier than the Akaike information criterion (AIC) for the complexity of the model [44]. The model with a smaller RMSE, MAE, MAPE, and BIC, and a larger Adjusted R^2 indicates the model is more stable and accurate. The Equations of MAE, RMSE, and MAPE are defined as (6)–(8). Where y_i is the actual

passenger volume intensity, \hat{y}_i is the predicted passenger volume intensity, and n is the sample size.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (7)$$

$$\text{MAPE} (\%) = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| * 100. \quad (8)$$

In addition, the Wilcoxon signed-rank test and Friedman test are carried out to ensure the significance of the superiority of the compared model for metro volume intensity prediction [45]. We also make a comparison between the true passenger volume and predict values in S1, S2, and S3. As shown in (9), the passenger volume is the product of the intensity and the operation mileage. The unit of the passenger volume is million persons/day. The unit of the operation mileage is kilometers. In addition, the unit of the passenger volume intensity is 10000 persons/km*day in this study.

$$\text{Passenger volume} = \frac{\text{passenger volume intensity} * \text{operation mileage}}{100} \quad (9)$$

The vehicle used in the initial metro network is 6-carriage B2 trains, with a rated passenger capacity of 1468 people. The number of departures, the actual capacity, and the actual load factor in 6-carriage B2 trains can be obtained, respectively, from (10) to (12). In (10), the departure interval of the train is set as 6.5 minutes. It operates from 6:00am to 23:00 pm, 17 hours in total. We think one more train as an auxiliary vehicle. In (11), the initial metro network includes two lines and the train runs in both directions. We consider the actual load factor of the train to be no more than 90% as the comfort limit that passengers can endure.

$$\text{The number of departures} = \frac{(23 - 6) * 60}{6.5} + 1, \quad (10)$$

$$\text{The actual capacity} = \frac{(\text{daily passenger volume}/\text{the number of departures} * 1000000)/2}{2}, \quad (11)$$

$$\text{The actual load factor} = \frac{\text{the actual capacity}}{1468} * 100\%. \quad (12)$$

4. Results and Discussions

4.1. MELR Results of Influencing Factors. There are interactions among candidate independent variables, so the Pearson correlation coefficients of 18 independent variables in Table 1 are calculated. The Pearson correlation coefficients of all independent variables are less than 0.5 in S1, S2, and S3, and the Variance Inflation Factor (VIF) are less than 5 (range from 1.338 to 4.151) in Table 2, suggesting that these independent variables are well selected and the multicollinearity issues can be avoided in the regression models [38]. The detailed regression results are discussed as follows:

- (1) The goodness-of-fit of the 3 MELR models are 0.957, 0.970, and 0.976, reflecting the highly linear relationship between the dependent variable and the independent variables with a good explanatory ability and no constant. The explanatory ability of S3 is the highest compared with S1 and S2, indicating that the MELR model has high accuracy with the dataset increase in the later initial metro network stage.
- (2) The days since the metro opened can be used to characterize the process of metro volume accumulation. It has a significant positive impact on daily passenger volume intensity as shown by He et al. [1, 8]. This study shows that it has lower influencing intensity than that reported in He et al. [8] which takes Shenzhen as an example. The metro is much more maturely developed in Shenzhen, with 5 lines and 118 stations and the least operation days of 839, whereas, in Xi'an initial metro network, the quantity of lines and stations are way less than Shenzhen. This causes the influencing difference of operation mileage between Shenzhen and Xi'an.
- (3) In Figure 6(a), the coefficient in other months is positive and significantly higher than that of February in S1, S2, and S3. Among them, March, April, July, and October have relatively high significant positive impact intensity. This is mainly because the return after the Spring Festival in March and April. Another factor playing a role is the tourists promoting the growth of metro volume during these two tourist golden months of the summer holidays and the National Day holiday. For S1, S2, and S3, the coefficient on Friday, Saturday, and Sunday is significantly higher than that on Wednesday in Figure 6(b). Studies find that the metro passenger volume shows an increasing trend on Friday and

Saturday [10]. Among them, the impact intensity on Saturday and Sunday is higher than the weekday, because a short tour around Xi'an may increase the metro volume.

- (4) The values of the *F*-statistic are 912.715, 1783.871, and 3068.325. The independent variable significance is less than 0.001, indicating that the dependent variables have a significant impact on the independent variables. In addition, the value of Durbin-Watson(*U*) is 0.627, 0.523, and 0.533 of the 3 MELR models, and they are all close to zero, showing the regression residuals of the MELR models have autocorrelation. Hence, some other time series models are further chosen to predict the daily passenger volume intensity and a comparison of them is investigated in Section 4.2.

4.2. Comparison of the Volume Intensity Prediction

4.2.1. The Overall Comparison Results of the Intensity. The exponential smoothing series models are relatively simple and effective methods for time series forecast in the short-term range. Nowadays, ARIMA and exponential smoothing models have been used for comparison purposes whenever a new forecasting model is proposed for short-term traffic [46]. To explore the accuracy of MELR models for metro volume prediction, we compare them with the 8 commonly used time series models.

Before conducting the time series model, the dependent variable needs to be tested for autocorrelation. Afterward, the Box-Ljung test is applied to verify the correlation between the squared residuals of the dependent variables in Table 3 [44]. The autocorrelation of the dependent variables is high from 0.497 to 0.954, the Box-Ljung all pass the significance test at the level of 0.05, demonstrating that there is a certain relationship between these dependent variables, that is, the past data can affect the following data, and we can use the time series prediction model.

After comparing 8 time series models, 6 better time series and 1 MELR model are retained eventually in Table 4 and Figure 7. The Adjusted R^2 of SNES and DNES models are close to 0. Hence, these two models are excluded from the original prediction models. In Table 4, the Adjusted R^2 values of the MELR model are greater than those of the corresponding time series regression models, demonstrating that the MELR model has the strongest explanatory power and stability among these time series models. The closer the value is to 1, the better the performance. But the value of

TABLE 2: Results of the MELR models in S1, S2, and S3.

Variables	MELR-S1			MELR-S2			MELR-S3		
	B	Pvalue	VIF	B	Pvalue	VIF	B	Pvalue	VIF
T_i	0.001	<0.001	4.151	0.001	<0.001	3.746	0.001	<0.001	3.605
X_{Jan}	0.293	<0.001	1.411	0.282	<0.001	1.504	0.237	<0.001	1.491
X_{Feb}	0.209	<0.001	1.389	0.189	<0.001	1.476	0.105	<0.001	1.456
X_{Mar}	0.293	<0.001	1.473	0.309	<0.001	1.567	0.273	<0.001	1.535
X_{Apr}	0.322	<0.001	1.486	0.304	<0.001	1.563	0.271	<0.001	1.520
X_{May}	0.249	<0.001	1.520	0.260	<0.001	1.607	0.217	<0.001	1.560
X_{Jun}	0.241	<0.001	1.558	0.214	<0.001	1.466	0.218	<0.001	1.352
X_{Jul}	0.261	<0.001	1.613	0.214	<0.001	1.338	0.256	<0.001	1.365
X_{Aug}	0.237	<0.001	1.660	0.185	<0.001	1.354	0.219	<0.001	1.383
X_{Sep}									
X_{Oct}	0.425	<0.001	1.359	0.382	<0.001	1.446	0.370	<0.001	1.443
X_{Nov}	0.334	<0.001	1.357	0.308	<0.001	1.453	0.305	<0.001	1.450
X_{Dec}	0.322	<0.001	1.401	0.322	<0.001	1.499	0.304	<0.001	1.482
X_{Mon}	0.201	<0.001	1.684	0.162	<0.001	1.673	0.176	<0.001	1.665
X_{Tues}	0.201	<0.001	1.683	0.157	<0.001	1.681	0.162	<0.001	1.671
X_{Wed}									
X_{Thur}	0.202	<0.001	1.683	0.155	<0.001	1.681	0.161	<0.001	1.677
X_{Fri}	0.300	<0.001	1.678	0.267	<0.001	1.677	0.295	<0.001	1.674
X_{Sat}	0.338	<0.001	1.682	0.300	0.041	1.678	0.317	<0.001	1.675
X_{Sun}	0.267	<0.001	1.671	0.233	0.084	1.671	0.247	<0.001	1.669
F(Sig.)		912.715			1783.871			3068.325	
Adjusted R^2		0.957			0.970			0.976	
Durbin-watson(U)		0.627			0.523			0.533	

MAE, RMSE, and MAPE in the MELR models are larger than the corresponding time series regression models. Thus, the results show that the MELR models generally perform better in understanding the determinants, and the time series regression models perform better in modeling linear volume forecasting. The same observation is drawn in references [19, 44]. Based on the value of MAPE, ranging from 6.597% to 7.011% in ARIMA models, it can be seen from Table 4 and Figure 7 that the prediction result of S3 is the most accurate among S1 and S2. Because the days since the metro opened in S1 are shorter than that of S2, the ARIMA models have advantages in the short-term passenger volume intensity prediction with rich datasets. The observations are the same as Ma et al. [19], but opposite to Chen et al. [44]. Except for the model of ARIMA in S1 and S2, the Ljung-Box Q(18) rejects the test and the p -value is all less than 0.05, showing that the residuals in the past affect the present residuals. It is believed that the residuals are not white noise sequences, and the model needs to be improved because it cannot fully recognize the real data. So, only the differences between the fitting/predicted and actual values of the intensity for MELR and ARIMA models of S1, S2 and S3 in Xi'an are compared, see Figures 8(a) and 8(b). If the difference changes around 0, it means that the fitting and prediction performance of the model is better. Otherwise, it means that the model should be improved with a bad prediction capacity. Figure 8(a) shows that the fitting effect of ARIMA is higher than MELR. While the MELR model predicts better results than ARIMA, as indicated in Figure 8(b), because the differences of the true and the predicted value in MELR model fluctuate around 0. Compared to the predicted value in S1 and S2, the

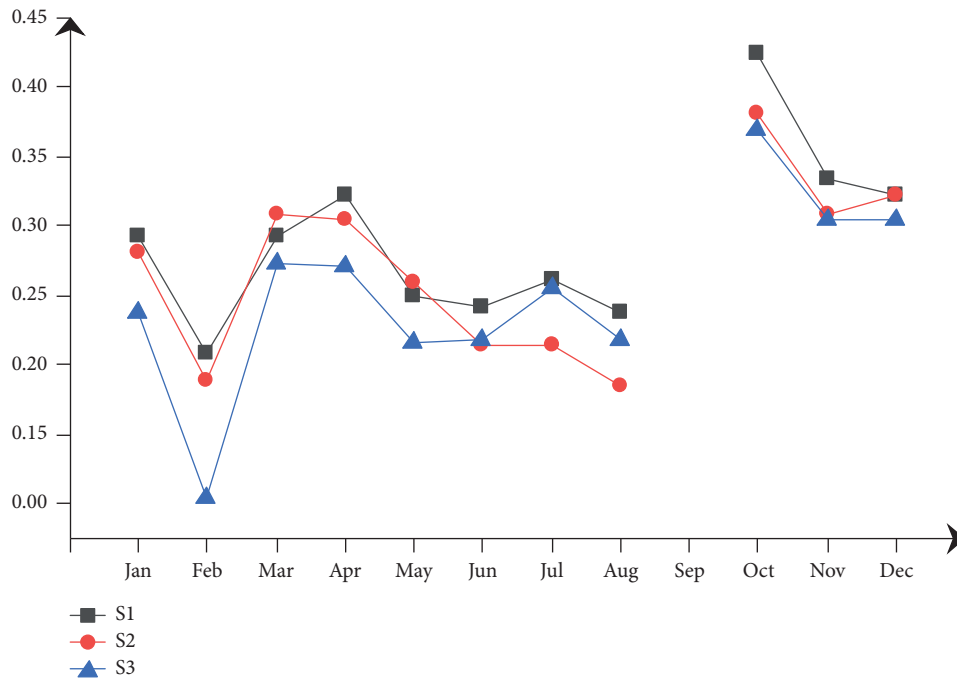
difference in MELR models remains stable, but the ARIMA models perform better in S2. This reveals the accuracy of ARIMA model is more dependent on the dataset size.

Because the model performance of SNES and DNES is less satisfactory, the Wilcoxon signed-rank test and Friedman test are not conducted for them. Only the significant comparison results of the statistical tests are shown in Table 5. Under the two-tail-test with one significance level of $\alpha=0.05$, it is further found that the MELR model significantly outperforms the other models in metro passenger volume intensity prediction, except for the model comparison of MELR vs. SSES/WAES/WMES in S1.

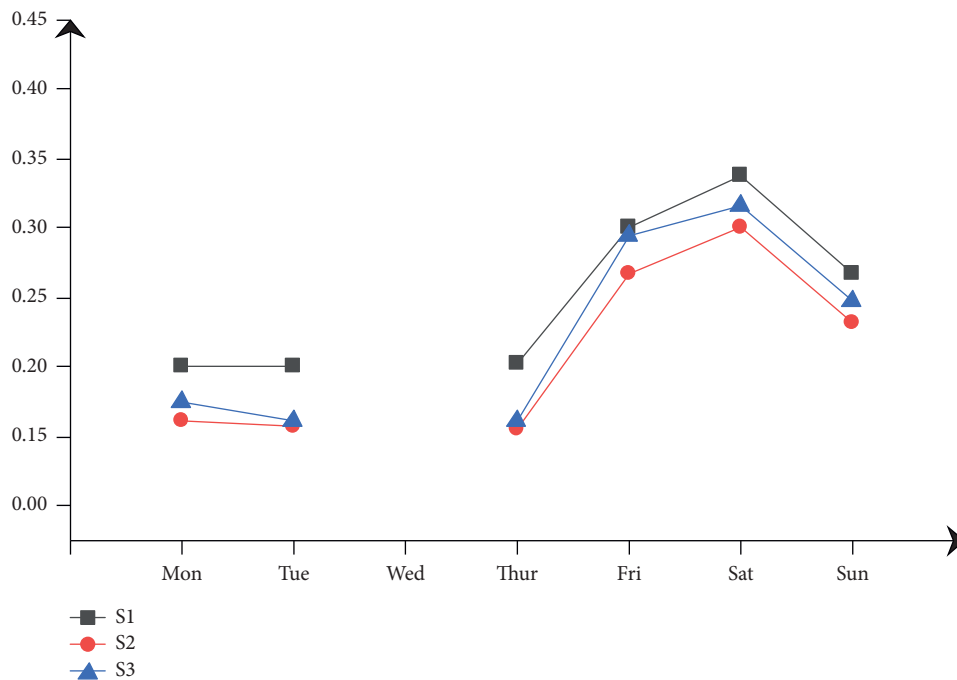
Eventually, compared with the ARIMA models, the MELR models have good performance in metro volume determinants exploration. They also show sufficient expandability and robustness in short-term initial metro volume prediction.

4.2.2. The Further Verification of the Prediction Results.

Table 6 shows the accumulated differences between the fitting/predicted value and the actual value of the intensity and volume by the MELR and ARIMA models in Line-S, Network-S1, and Network-S2. Results show that the MELR and ARIMA models have good fitting and predictability in each stage. The ratio of accumulated fitting differences in ARIMA models is less than MELR models, whether the model is focused on S1 or S2. However, the ratio of accumulated predicted differences in MELR models is less than ARIMA models, when the model is focused on the S1. As the amount of dataset increases, the ratio of accumulated predicted differences in MELR and



(a)



(b)

FIGURE 6: Variation of the regression coefficients in months (a) and days (b).

ARIMA models becomes lessened in S2. Unless the ARIMA-S1 model of the intensity, all of the ratio of accumulated predicted differences in other models are less than 9%. We can conclude the MELR model can adapt to the abrupt change volume when the new and old lines open and expand. Especially, the MELR model can avoid the drawback of insufficient datasets. The MELR and ARIMA models have good predictability in short-term daily passenger intensity.

Based on the predicted intensity, the passenger volume can be calculated. According to the data recorded in Baidu Encyclopedia, the 1,000 days (from June 11, 2014 to September 16, 2011) cumulative passenger volume of Xi'an Metro has reached 313 million person-times. The predicted result of the MELR-S1 model is 291 million person-times, which is only 7.01% less than the actual value. While the predicted result of the ARIMA-S1 model is 291 million person-times, which is 10.68% less than the

TABLE 3: Autocorrelation test of the dependent variables.

Delay	S1		S2		S3	
	Autocorrelation	Box-Ljung test Sig	Autocorrelation	Box-Ljung test Sig	Autocorrelation	Box-Ljung test Sig
1	0.843	0.000	0.935	0.000	0.954	0.000
2	0.669	0.000	0.869	0.000	0.907	0.000
3	0.575	0.000	0.829	0.000	0.876	0.000
4	0.538	0.000	0.810	0.000	0.861	0.000
5	0.552	0.000	0.811	0.000	0.866	0.000
6	0.637	0.000	0.838	0.000	0.885	0.000
7	0.689	0.000	0.850	0.000	0.893	0.000
8	0.624	0.000	0.820	0.000	0.869	0.000
9	0.535	0.000	0.784	0.000	0.842	0.000
10	0.504	0.000	0.765	0.000	0.828	0.000
11	0.497	0.000	0.764	0.000	0.827	0.000
12	0.525	0.000	0.777	0.000	0.840	0.000
13	0.613	0.000	0.807	0.000	0.861	0.000
14	0.659	0.000	0.821	0.000	0.872	0.000
15	0.599	0.000	0.800	0.000	0.855	0.000
16	0.514	0.000	0.770	0.000	0.834	0.000

TABLE 4: The comparison of passenger volume prediction in MELR and time series models.

Model	RMSE	MAPE (%)	MAE	Normalized BIC (L)	Ljung-Box Q (18)	Adjusted R^2
MELR-S1	0.181	15.271	0.123	—	—	0.957
MELR-S2	0.187	14.237	0.128	—	—	0.970
MELR-S3	0.199	13.506	0.139	—	—	0.976
ARIMA (2, 0, 0) (1, 1, 1)-S1	0.088	7.011	0.057	-4.825	0.161	0.667
ARIMA (1, 0, 6) (1, 1, 1)-S2	0.100	6.884	0.065	-4.577	0.012	0.690
ARIMA (0, 0, 4) (1, 1, 1)-S3	0.108	6.597	0.071	-4.406	0.000	0.702
HNES-S1	0.107	9.225	0.077	-4.443	0.000	0.464
HNES-S2	0.119	8.794	0.085	-4.243	0.000	0.489
HNES-S3	0.131	8.234	0.092	-4.048	0.000	0.491
BNES-S1	0.125	11.742	0.094	-4.148	0.000	0.273
BNES-S2	0.138	10.683	0.101	-3.958	0.000	0.314
BNES-S3	0.152	10.029	0.110	-3.757	0.000	0.315
SSES-S1	0.094	7.660	0.063	-4.721	0.000	0.376
SSES-S2	0.105	7.331	0.070	-4.500	0.000	0.417
SSES-S3	0.115	6.908	0.077	-4.314	0.000	0.398
WAES-S1	0.094	7.690	0.063	-4.709	0.000	0.375
WAES-S2	0.105	7.337	0.070	-4.492	0.000	0.417
WAES-S3	0.115	6.917	0.077	-4.308	0.000	0.398
WMES-S1	0.093	7.693	0.064	-4.714	0.000	0.378
WMES-S2	0.105	7.414	0.071	-4.491	0.000	0.416
WMES-S3	0.114	6.932	0.077	-4.325	0.000	0.408

actual value, the MELR-S2 model predicts 305 million person-times, which is only 2.70% less than the actual value. While the predicted result of the ARIMA-S2 model is 309 million person-times, which is 1.16% less than the actual value. Therefore, if there are sufficient historical data, it is further verified that the MELR and ARIMA models have superiority in short-term passenger volume forecasting for the initial metro network.

To further verify the models' predicted ability, we choose the beginning day and the National Day in the Network-S1 and Network-S2 as the comparative days. The daily passenger volume and intensity, the actual capacity, and load factor in 6-carriage B2 trains can be retrieved from Table 7. It is found that the real passenger volume can meet the needs in

the Network-S1, and the actual load rate reaches 85% in the National Day in Network-S1. But after entering the Network-S2 stage, the actual load factor increased to 90.7%, almost in a saturated situation and up to the comfort limit on October 11, 2014. The actual load factor calculated by the MELR-S2 and ARIMA-S2 models is closer to the actual value. Only when the amount of data is large, the ARIMA model can predict the value of holidays more accurately, while the MELR model has lower requirements for the dataset size and can be more accurate. The MELR model owns good predictions for normal and holidays. It shows that the MELR model has good adaptability to the initial passenger volume prediction with jumpy characteristics in each stage.

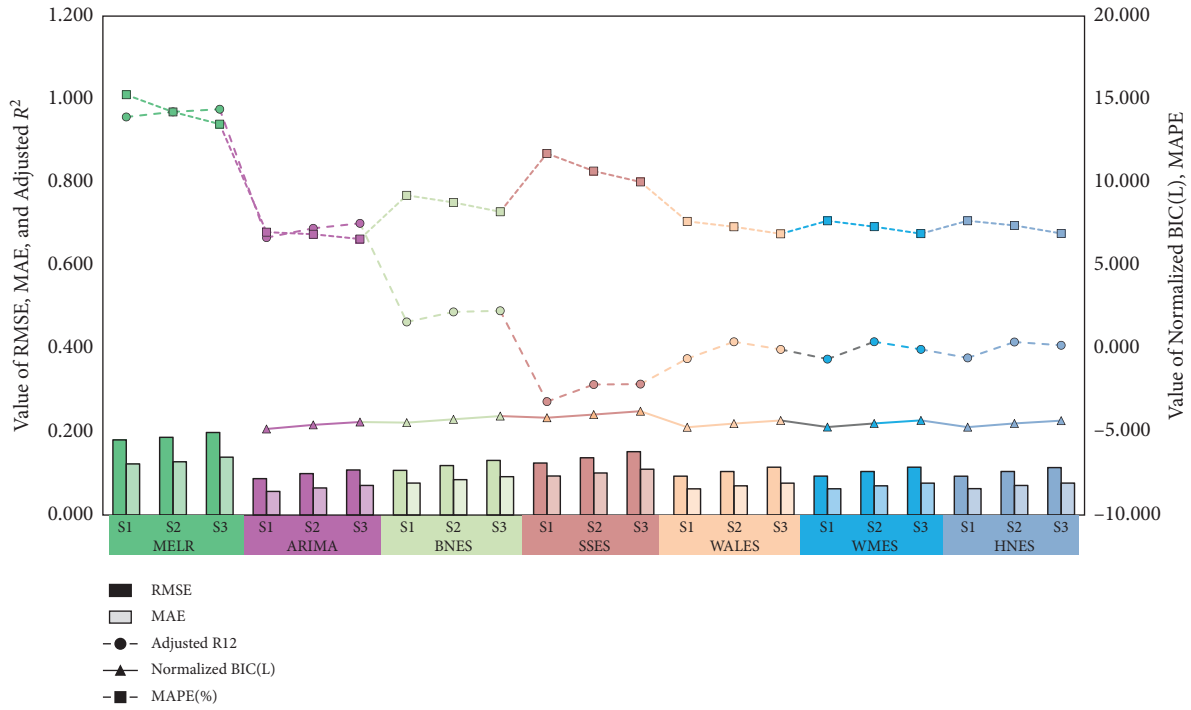
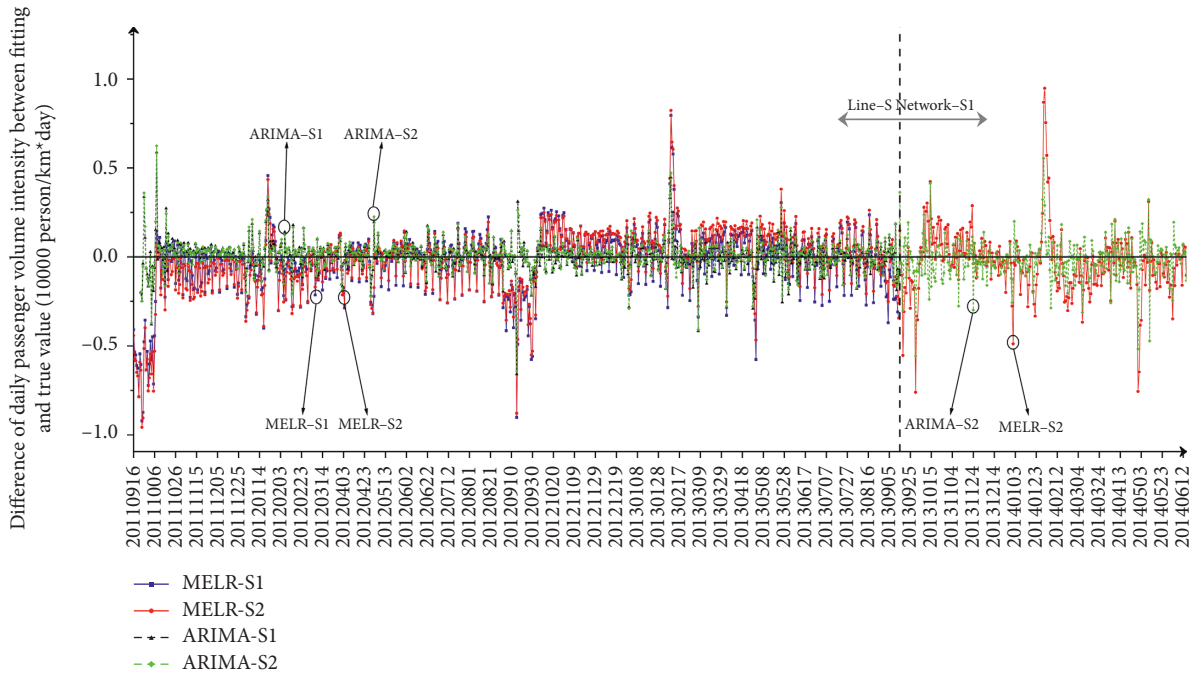


FIGURE 7: The comparison of key evaluation indices in MELR and time series models.



(a)

FIGURE 8: Continued.

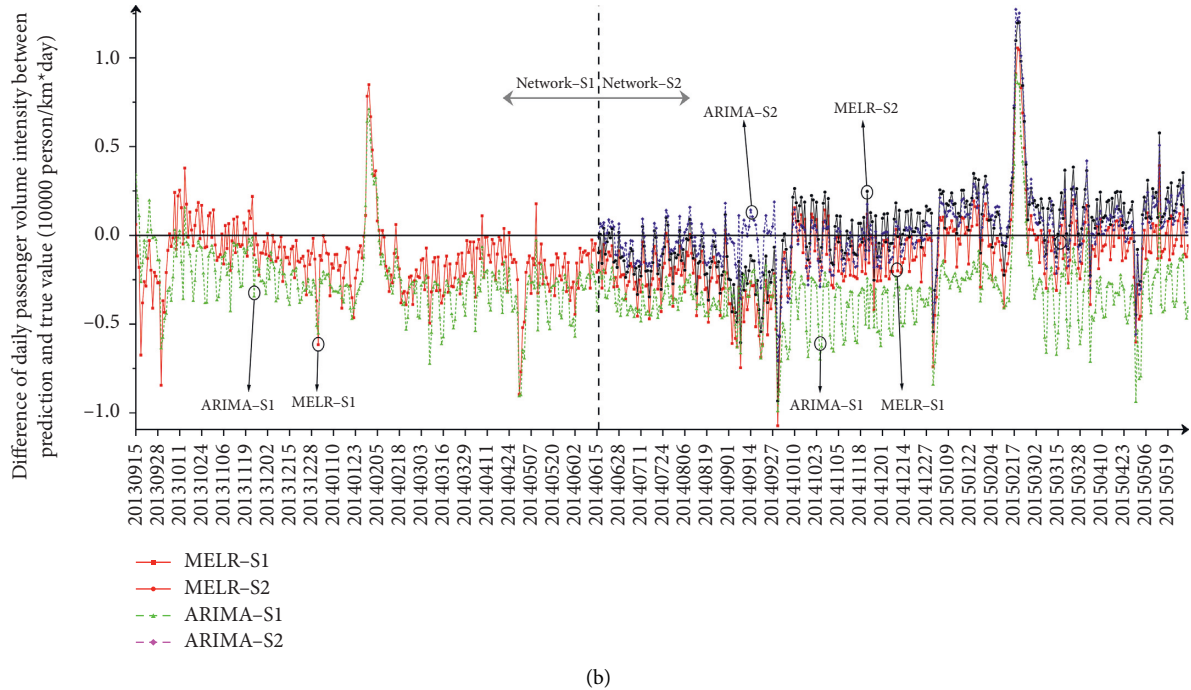


FIGURE 8: (a) The difference between fitting and actual daily passenger volume intensity in the initial network for MELR and ARIMA models. (b) The difference between predicted and actual daily passenger volume intensity in the initial network for MELR and ARIMA models.

TABLE 5: Results of Wilcoxon signed-rank test and Friedman test in S1, S2, and S3.

Dataset stages	Compared models	Wilcoxon signed-rank test $\alpha = 0.05$ P value	Friedman test
S1	MELR vs. HNES	0.0060	$F = 98.515$ $P = 0.000$ (Reject H_0)
	MELR vs. BNES	0.0280	
	MELR vs. SSES	0.1020	
	MELR vs. WAES	0.0780	
	MELR vs. WMES	0.0730	
	MELR vs. ARIMA	0.0380	
S2	MELR vs. HNES	0.0140	$F = 310.634$ $P = 0.000$ (Reject H_0)
	MELR vs. BNES	0.0040	
	MELR vs. SSES	0.0150	
	MELR vs. WAES	0.0080	
	MELR vs. WMES	0.0120	
	MELR vs. ARIMA	0.0040	
S3	MELR vs. HNES	0.0000	$F = 388.945$ $P = 0.000$ (Reject H_0)
	MELR vs. BNES	0.0000	
	MELR vs. SSES	0.0010	
	MELR vs. WAES	0.0000	
	MELR vs. WMES	0.0010	
	MELR vs. ARIMA	0.0010	

TABLE 6: The difference between fitting/predicted and actual daily passenger intensity and volume in MELR and ARIMA models.

Stages and models	Difference of daily passenger intensity (and the ratio)				Difference of daily passenger volume (and the ratio)			
	MELR-S1	MELR-S2	ARIMA-S1	ARIMA-S2	MELR-S1	MELR-S2	ARIMA-S1	ARIMA-S2
Line-S	-20.8254 (-3.30%)	-14.4570 (-2.29%)	1.7207 (0.27%)	6.5048 (1.03%)	-4.2692 (-0.03%)	-2.9637 (-0.02%)	0.3527 (0.00%)	1.3335 (0.01%)
Network-S1	-34.7964 (-8.65%)	-7.8562 (-1.95%)	-68.0916 (-16.93%)	-4.2372 (-1.05%)	-15.9715 (-0.13%)	-3.6060 (-0.03%)	-31.2541 (-0.25%)	-1.9449 (-0.02%)
Network-S2	-42.1463 (-6.90%)	8.0836 (1.32%)	-118.7175 (-19.44%)	9.5106 (1.56%)	-22.0003 (-0.12%)	4.2196 (0.02%)	-61.9705 (-0.34%)	4.9645 (0.03%)

Note. The bold text indicates the fitting value and the normal text indicates the predicted value. Values in parentheses indicate the ratio of the accumulated difference to the true intensity or volume.

TABLE 7: The comparison of daily passenger volume and intensity on special days.

Model	Stage	Date	Daily passenger volume (million persons/day)	Daily passenger volume intensity (10000 persons/km)	The actual capacity in 6- carriage B2 trains	The actual load factor in 6-carriage B2 trains (%)
True value	Network-S1	20130915	0.419	0.913	663	45.2
		20131001	0.789	1.718	1248	85.0
	Network-S2	20140616	0.842	1.612	1332	90.7
MELR-S1	Network-S1	20141011	0.924	1.770	1462	99.6
		20130915	0.447	0.974	708	48.2
		20131001	0.619	1.350	980	66.8
	Network-S2	20140616	0.739	1.415	1169	79.6
		20141011	0.965	1.849	1527	104.0
		20130915	0.487	1.061	771	52.5
MELR-S2	Network-S1	20131001	0.636	1.385	1006	68.5
		20140616	0.791	1.515	1251	85.2
	Network-S2	20141011	1.020	1.954	1614	109.9
ARIMA-S1	Network-S1	20130915	0.576	1.255	911	62.1
		20131001	0.523	1.140	828	56.4
	Network-S2	20140616	0.671	1.285	1061	72.3
		20141011	0.736	1.409	1164	79.3
ARIMA-S2	Network-S1	20130915	0.586	1.277	927	63.2
		20131001	0.753	1.642	1192	81.2
	Network-S2	20140616	0.837	1.603	1324	90.2
		20141011	0.947	1.814	1498	102.0

5. Conclusion

This study aims to explore the influencing factors and predict the passenger volume of the initial metro network, which makes up for this insufficiency based on the station level. Using the MELR and time series models, we build the MELR model to explore the determinants and forecast the intensity during the twice expansion of the initial metro network, based on the 1354 and 623 days historical operation data of the Xi'an initial metro network. We further compare the prediction of the metro transport capacity between the MELR models with exponential smoothing and ARIMA models. The following observations are found: (1) The MELR model exhibits high explanatory ability of temporal influencing factors exploration in Xi'an initial metro network. (2) The days since the metro opened can be used to characterize the process of metro volume accumulation. It has a significant positive impact on daily passenger volume intensity. (3) The passenger intensity fluctuates with the months and days. The return volume after the Spring Festival in March and April, and the tourists promote the growth of metro volume during the summer holidays and the National Day holiday. And, a short tour around Xi'an may increase the metro volume increase on Saturday. (4) The MELR is a concise and valid model to predict the abrupt volume during the new metro line opening and the old line expanding, which avoids the drawback of time series models that need a huge database. The results of short-term passenger volume forecasting can provide useful information for decision-makers of metro systems. They can appropriately adjust the operation plans (e.g., headway and train dispatching), activate the station passenger crowd regulation plan and emergency response plan provided that the predicted passenger volume is higher than the predetermined threshold.

The proposed MELR models can be used to examine the influencing factors of ridership or its density in various transportation modes. Furthermore, ARIMA and exponential smoothing models can be used for comparison purposes whenever a new forecasting model for short-term traffic volume or time is proposed in the future.

We also summarize the limitations, as well as the future works as follows. At first, we mainly consider temporal influencing factors, and some results can be explained from a deeper spatial perspective (e.g., urban structure, built environment) in the initial line and network level. Secondly, a suitable fine-grained time granularity is studied to predict subway passenger volume, which provides decision-makers with accurate schedule and operation management results. Finally, more other models that can simultaneously explore the influencing factors and predict the passenger volume of metro lines with high accuracy should be considered in the future.

Data Availability

Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request. (Date (day/month/year), Daily passenger volume, Days since the metro opened).

Additional Points

Discovering the temporal influential factors of the initial metro network volume intensity. The MELR model is more adapted for the short-term prediction of the abrupt volume. The ARIMA model's predicted accuracy is more dependent on the dataset size.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Authors' Contributions

Tao Lyu and Mingfei Xu are co-first authors and were contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant number: 51878062 and 51908462], the Natural Science Basic Research Program of Shaanxi (Program No. 2020JQ-387), the Fundamental Research Funds for the Central Universities, CHD (Program No. 300102341307), and the Higher Education Discipline Innovation Project 111 [grant number: B20035]. The authors confirm contributions to the paper as follows: Tao Lyu was involved in formulating the methodology, visualization, data curation, formal analysis, writing—original draft, review, & editing. Mingfei Xu was concerned with conceptualization, writing—review & editing. Jia Zhang contributed to formulating the methodology, data curation, formal analysis, and writing—review & editing. Yuanqing Wang was involved in funding acquisition, project administration, supervision, writing—review & editing. Liu Yang and Yanan Gao were involved in writing—review & editing. All listed authors have contributed to the manuscript substantially and have agreed to the final submitted version.

References

- [1] Y. He, Y. Zhao, and K. L. Tsui, "An Adapted Geographically Weighted LASSO (Ada-GWL) Model for Predicting Subway Ridership," *Transportation*, vol. 48, pp. 1–32, 2020.
- [2] D. Liu, Z. Wu, and S. Sun, "Study on Subway passenger flow prediction based on deep recurrent neural network," *Multi-media Tools and Applications*, vol. 1, 2020.
- [3] Baidu, "Passenger Intensity," 2021, <https://baike.baidu.com/item/%E5%AE%A2%E8%BF%90%E5%BC%BA%E5%BA%A6/15508423?fr=aladdin>.
- [4] R. Cervero, "Alternative approaches to modeling the travel-demand impacts of smart growth," *Journal of the American Planning Association*, vol. 72, no. 3, pp. 285–295, 2006.
- [5] M. Kuby, A. Barranda, and C. Upchurch, "Factors influencing light-rail station boardings in the United States," *Transportation Research Part A: Policy and Practice*, vol. 38, no. 3, pp. 223–247, 2004.
- [6] C. Lin, K. Wang, D. Wu, and B. Gong, "Passenger flow prediction based on land use around metro stations: a case study," *Sustainability*, vol. 12, no. 17, pp. 1–23, 2020.
- [7] H. Sung, K. Choi, S. Lee, and S. Cheon, "Exploring the impacts of land use by service coverage and station-level accessibility on rail transit ridership," *Journal of Transport Geography*, vol. 36, pp. 134–140, 2014.
- [8] Y. He, Y. Zhao, and K. L. Tsui, "Geographically modeling and understanding factors influencing transit ridership: an empirical study of Shenzhen metro," *Applied Sciences*, vol. 9, no. 20, pp. 1–23, 2019.
- [9] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 148–162, 2012.
- [10] Z. Qian, X. Liu, S. Spurgeon, and D. Yu, "A two-layer modelling framework for predicting passenger flow on trains: a case study of london underground trains," *Transportation Research Part A: Policy and Practice*, vol. 151, pp. 119–139, 2021.
- [11] C. Liu, S. Erdogan, T. Ma, and F. W. Ducca, "How to increase rail ridership in Maryland: direct ridership models for policy guidance," *Journal of Urban Planning and Development*, vol. 142, no. 4, Article ID 4016017, 2016.
- [12] H. Bar-Gera and D. Boyce, "Origin-based algorithms for combined travel forecasting models," *Transportation Research Part B: Methodological*, vol. 37, no. 5, pp. 405–422, 2003.
- [13] S. C. Wirasinghe and A. S. Kumarage, "An aggregate demand model for intercity passenger travel in Sri Lanka," *Transportation*, vol. 25, no. 1, pp. 77–98, 1998.
- [14] M. G. McNally, *The Four-step Model*, *Handbook of Transport Modelling*, pp. 35–52, Elsevier, Oxford, UK, 2000.
- [15] O. D. Cardozo, J. C. Garcia-Palomares, and J. Gutiérrez, "Application of geographically weighted regression to the direct forecasting of transit ridership at station-level," *Applied Geography*, vol. 34, pp. 548–558, 2012.
- [16] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: overview of objectives and methods," *Transport Reviews*, vol. 24, no. 5, pp. 533–557, 2004.
- [17] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006.
- [18] L. Tang, Y. Zhao, J. Cabrera, J. Ma, and K. L. Tsui, "Forecasting short-term passenger flow: an empirical study on Shenzhen metro," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3613–3622, 2019.
- [19] C. Ma, P. Li, and C. Zhu, "Short-term passenger flow forecast of urban rail transit based on different time granularities," *Journal of Chang'an University (Natural Science Edition)*, vol. 40, no. 3, pp. 75–83, 2020.
- [20] P. Zhao and L. Li, "Prediction of urban rail transit station inflows based on ARIMA model," *Journal of Chongqing Jianshu University*, vol. 39, no. 1, pp. 40–44, 2020.
- [21] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.
- [22] M. Castro-Neto, Y.-S. Jeong, M.-K. Jeong, and L. D. Han, "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6164–6173, 2009.
- [23] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, 2015.
- [24] L. Vanajakshi and L. R. Rilett, "Support Vector Machine Technique for the Short Term Prediction of Travel Time," in *Proceedings of the 2007 IEEE Intelligent Vehicles Symposium*, pp. 600–605, Istanbul, Turkey, June 2007.
- [25] X. Jiang and H. Adeli, "Dynamic wavelet neural network model for traffic flow forecasting," *Journal of Transportation Engineering*, vol. 131, no. 10, pp. 771–779, 2005.
- [26] Y. Li, X. Wang, S. Sun, X. Ma, and G. Lu, "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 306–328, 2017.

- [27] B. Abdulhai, H. Porwal, and W. Recker, "Short-term traffic flow prediction using neuro-genetic algorithms," *Intelligent Transportation Systems Journal*, vol. 7, no. 1, pp. 3–41, 2002.
- [28] D. Li, C. Zhang, and J. Cao, "Short-term passenger flow prediction of a passageway in a subway station using time space correlations between multi sites," *IEEE Access*, vol. 8, Article ID 72471, 2020.
- [29] X. Rui, S. Daniel, and C. Shukai, "Short-term bus passenger demand prediction based on time series model and interactive multiple model approach," *Discrete Dynamics in Nature and Society*, vol. 2015, Article ID 682390, 11 pages, 2015.
- [30] T. Lu, E. Yao, S. Liu, and W. Zhou, "Short-time forecast of entrance and exit passenger flow for new line of urban rail transit during growth period," *Journal of the China Railway Society*, vol. 42, no. 5, pp. 19–28, 2020.
- [31] X. Zhao, Y.-p. Wu, G. Ren, K. Ji, and W.-W. Qian, "Clustering analysis of ridership patterns at subway stations: a case in Nanjing, China," *Journal of Urban Planning and Development*, vol. 145, no. 2, Article ID 4019005, 2019.
- [32] X. Cao and J. Schoner, "The influence of light rail transit on transit use: an exploration of station area residents along the Hiawatha line in Minneapolis," *Transportation Research Part A: Policy and Practice*, vol. 59, pp. 134–143, 2014.
- [33] A. Chakraborty and S. Mishra, "Land use and transit ridership connections: implications for state-level planning agencies," *Land Use Policy*, vol. 30, no. 1, pp. 458–469, 2013.
- [34] B. P. Y. Loo, C. Chen, and E. T. H. Chan, "Rail-based transit-oriented development: lessons from New York city and Hong Kong," *Landscape and Urban Planning*, vol. 97, no. 3, pp. 202–212, 2010.
- [35] M. Moniruzzaman and A. Páez, "Accessibility to transit, by transit, and mode share: application of a logistic model with spatial filters," *Journal of Transport Geography*, vol. 24, pp. 198–205, 2012.
- [36] W. L. Wang, S. M. Lo, and S. B. Liu, "Aggregated metro trip patterns in urban areas of Hong Kong: evidence from automatic fare collection records," *Journal of Urban Planning and Development*, vol. 141, no. 3, Article ID 5014018, 2015.
- [37] J. Zhao, W. Deng, Y. Song, and Y. Zhu, "What influences Metro station ridership in China? Insights from Nanjing," *Cities*, vol. 35, pp. 114–124, 2013.
- [38] X. Qian and S. V. Ukkusuri, "Spatial variation of the urban taxi ridership using GPS data," *Applied Geography*, vol. 59, pp. 31–42, 2015.
- [39] Xi'an Municipal Bureau of Statistics, "Xi'an statistical year-book in 2016," 2016, <https://tjj.xa.gov.cn/tjnj/2016/zk/indexch.htm>.
- [40] D. Zhang and X. Wang, "Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC," *Journal of Transport Geography*, vol. 41, pp. 107–115, 2014.
- [41] M. M. Hamed, H. R. Al-Masaeid, and Z. M. B. Said, "Short-term prediction of traffic volume in urban arterials," *Journal of Transportation Engineering*, vol. 121, no. 3, pp. 249–254, 1995.
- [42] S. Lee and D. B. Fambro, "Application of subset autoregressive Integrated Moving Average Model for short-term freeway traffic volume forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1678, no. 1, pp. 179–188, 1999.
- [43] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Wiley, Hoboken, NJ, USA, 1994.
- [44] E. Chen, Z. Ye, C. Wang, and M. Xu, "Subway passenger flow prediction for special events using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1109–1120, 2020.
- [45] Z. Zhang and W.-C. Hong, "Application of variational mode decomposition and chaotic grey wolf optimizer with support vector regression for forecasting electric loads," *Knowledge-based System*, vol. 228, pp. 1–16, 2021.
- [46] B. Park, C. J. Messer, and T. Urbanik, "Short-term freeway traffic volume forecasting using radial basis function neural network," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1651, no. 1, pp. 39–47, 1998.