Chapter 36
Application of Time Series Method to the Passenger Flow Prediction in the Intelligent Bus Transportation System with Big Data

Yinna Ye, Ruoxi Liu, and Feng Xue

Abstract Based on the real data collected from the bus IC card payment devices, first a time series plot on the daily passenger volume was obtained and then three kinds of time series models were proposed to do the prediction. The results show that the ARMA model with quadratic trend is the most suitable to the current data and performs the most effectively in the prediction.

36.1 Introduction

With the current development of the intelligent urban public transportation system in China, the investigation on the bus passenger flow has become a key research subject (see [1] for instance). In order to maintain the competitiveness in the transportation market and provide services with high-level quality to the passengers, the bus transportation companies need to grasp the change rules of the passenger demand sustainably [2]. However, the passenger flow in the bus system is influenced by many factors, including commuting, holiday, weather, temperature, etc. [2]. For example, the volume would experience a sudden increase during low temperature and snowy days, which would lead to the inability of bus transport capacity to meet passenger demand and brings tremendous pressure to the bus transportation management. Considering the limited bus resources, some popular routes are often in short supply, which might result in the problems of passenger flow detention and reduced-quality service. The bus companies might thus lose competitiveness in the transportation market. Therefore, it is necessary to find an effective solution to the problems

Y. Ye (✉) · R. Liu
Department of Mathematical Sciences, Xi'an Jiaotong—Liverpool University, Suzhou, China
e-mail: yinna.ye@xjtlu.edu.cn

R. Liu
e-mail: ruoxiliu15@outlook.com

F. Xue
Xiongdi Shenzhen Emperor Technology Company, Shenzhen, China
e-mail: xuefeng@xiongdi.cn

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_36

caused by such burst of passenger flow and adjust the current management policies in support to the optimal bus resource allocation, line planning and bus scheduling. The solution is of great importance to improve both the service capability and the working efficiency in the public transportation system.

The driving motivation of this work is to find a reliable method to solve the problems above. Undoubtedly, this piece of work is socially significant and important since the urban transport plan and policy could be well designed or adjusted with adapting the market demand. The implementation of this work involves a combination of big data processing, time series modeling and analysis. The primary objective of the work is to apply the time series models and data analytics to explore the passenger demand based on the real data and then to predict the daily passenger volume in a given bus line. The study will mainly focus on the following two aspects:

- Descriptive statistics on the trip characteristics of passengers, including riding date and time, and on the volume and variation characteristics of transit passenger flow at different stations in a given bus line.
- Time series parameters estimation and passenger volume prediction are based on the bus tick sale records.

In this work, SAS (version 9.4) (see [3] for instance) will be used to obtain the descriptive statistics, to do time series analysis and predictions.

The rest of the paper is organized as follows. Section 36.2 reviews the development of time series analysis and recent works on the application of the time series to the public transportation systems. Section 36.3 presents time series related concepts and methods, as well as our data analysis process. Section 36.4 summarizes and evaluates the empirical results. And finally the conclusion is discussed in Sect. 36.5 and certain open questions and some future improvements are proposed in Sect. 36.6.

36.2 Literature Review

Prior to 1920, the time series was limited to drawing lines through a mass of data. In 1927, Yule [4] first introduced the concept of ‘autoregressive’ that the variables are time related and time is not a causal factor, and pioneered the autoregressive (AR) Model of order two when studying the number of sunspots and exploring the period of the disturbed sequence. The autoregressive model he established is a special kind of stationary time series. In 1931, Walker [5] expanded and generalized the AR model to higher orders. While, Slutsky [6] was interested in the randomness of the time series, regarding them as the perturbations and then the moving average (MA) model was proposed. In 1938, Wold [7] proved that the discrete stationary process consists of implicit periodicity and linear regression. The hidden cycle is a deterministic component, while the linear regression part consists of a moving average and an autoregressive process, which are non-deterministic components of random perturbations. Any stationary time series, whose deterministic components are eliminated, can be reduced to a linear combination of random perturbations. This well-known time series decomposition idea is the theoretical basis for the idea of the autoregressive moving average (ARMA) model. By taking non-stationary into consideration, the autoregressive integrated moving average (ARIMA) model was proposed in the landmark work [2]. The book provided a systematic approach to analyze and forecast the time series and discussed how to identify, estimate and diagnose the ARIMA model.

The application of time series models in the modern society has rapidly widespread, as the application was extended to non-stationary process (see for instance [8]). A large number of empirical results show that most time series established based on the socio-economic phenomena are non-stationary and have a trend (see for instance [9]). According to Xia [9], there are two types of time trend, one is deterministic and another one is random. Deterministic time trend is the one that can be characterized by a function of the time. The commonly used trend functions are linear functions, quadratic parabola functions, exponential functions and logarithmic functions. By contrast, the time series with stochastic trend cannot be expressed by the deterministic functions of time. In this case, multiple differences are operated to the original process and then the ARIMA model is used to fit the data.

In the literature, the existing researches suggest that the time series analysis has been properly utilized in studying different public transportation systems. For the subway systems in Shanghai, Zhu [10] constructed an ARIMA model for the daily passenger flow by comparing the change rate of daily volume with that of '7-day' average volume. For the airport terminal departure passenger traffic, Li et al. [11] took daily periodicity of the process into consideration and proposed a seasonal autoregressive integrated moving average (SARIMA) model to predict the passenger flow in Kunming Changshui International Airport. For the railway passenger flow forecast, a time series model was established in [12] with the combination of the long-term trend, the seasonal and the weather factors. To achieve an accurate real-time taxi passenger hotspot prediction, Jamil and Akbar [13] proposed an automatic ARIMA model to determine the value of the model order automatically. The algorithm designed by them overcame the common obstacle, subjectivity and complexity. All these applications make use of the knowledge of passenger flow and provide instructive insight to the management of the public transportation system, which has a referential significance for our investigation.

36.3 Methodology

36.3.1 Stationary Time Series Models

The time series analysis aims to reveal the underlying dynamics and structures that affect observable data, thus establishing a suitable theoretical model for monitoring and predicting data. For the definition of stationary time series (or simply called 'time series'), one can refer for instance to the Definition 1.3.2 in [14]. In this book, the daily passenger flow volumes $\{Z_t\}$ at any unit of time t will be regarded as a discrete-time stochastic process. Roughly speaking, assuming that $\{Z_t\}$ is a stationary time series with mean 0 and Z_t depends only on its historical records $Z_{t-1}, Z_{t-2}, \ldots$ then we can use the observed historical data to estimate the dynamic properties, create optimal models and then use these models to do the prediction. In this project, we construct discretely sampled time series based on the actual daily records of passenger volume in a given bus line. The detailed description about the database can be found in Sect. 36.4.1. In the rest of this subsection, some related fundamental concepts will be introduced. One may refer to [8] for the details.

Autoregressive Model: AR (*p*). The autoregressive (AR) model is a very common time series. The general *p*-order autoregressive model, denoted as AR(*p*), is given by:

$$Z_t = \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \cdots + \varphi_p Z_{t-p} + a_t, \tag{36.1}$$

where the parameters $\varphi_1, \varphi_2, \ldots, \varphi_p$ are called **autoregressive coefficients** and they are to be estimated. The random error terms $\{a_t\}$ is the white noise, i.e., a sequence of i.i.d. random variables, $a_t \sim N\left(0, \sigma_a^2\right)$ and $\{a_t\}$ is mutually independent with $Z_{t-1}, Z_{t-2}, \ldots, Z_{t-p}$.

Moving Average Model: MA (*q*). The general *q*-order moving average model, denoted as MA (*q*), is given by:

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}, \tag{36.2}$$

where $\theta_1, \theta_2, \ldots, \theta_q$ are called **moving average coefficients** and they are to be estimated.

Autoregressive Moving Average Model: ARMA (*p*, *q*). The autoregressive moving average (ARMA) combines an AR model with a MA model to produce a new process that simulates the time series. The general ARMA model, denoted as ARMA (*p*, *q*), is given by

$$\begin{aligned} Z_t &= \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \cdots + \varphi_p Z_{t-p} \\ &\quad + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}. \end{aligned} \tag{36.3}$$

Autoregressive Integrated Moving Average Model: ARIMA(*p*, *d*, *q*). Notice that the AR, MA, and ARMA models are stationary time series. However, sometimes the time series are not necessarily stationary. It may have a linear trend component. For non-stationary time series, it is necessary to transform it into a stationary one through the backward shift operator. Such a non-stationary time series is called **ARIMA process**, denoted as ARIMA (*p*, *d*, *q*), and is given by

$$\left(1-\varphi_1 B-\cdots-\varphi_p B^p\right)(1-B)^d Z_t=\left(1-\theta_1 B-\cdots-\theta_q B^q\right)a_t, \tag{36.4}$$

where B is the backward shift operator (lag) defined as $(1-B)Z_t = Z_t - Z_{t-1}$ and d is the number (order) of the difference to make the process stationary.

ARMA Model with a Quadratic Function Trend. Indeed, besides considering a linear trend component in the time series, some other trend forms may also be taken into account. If the trend of a time series has a shape as a quadratic function, then it can be fitted by a quadratic function. The ARMA model with a quadratic function trend is given by

$$\begin{aligned} Z_t &= \text{quadratic function} + \text{ARMA process} \\ &= at + bt^2 + \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \cdots + \varphi_p Z_{t-p} + a_t \\ &\quad - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \end{aligned} \tag{36.5}$$

In the rest of this section, the application of time series method to the passenger flow prediction will be introduced. This can be achieved by the descriptive and inferential studies on the current data.

36.3.2 Time Series Analysis

According to [8], the main steps of time series analysis and modeling are:

1. Stationarity and white noise test
2. Model identification (i.e., specifying the lag order)
3. Model selection and parameter estimation
4. Diagnostic checking
5. Prediction based on the optimal model.

Stationary Test. The first step of time series analysis is to verify whether the series is stationary. There are two main methods: one is the graph test, which illustrates the features shown in the time series plots and autocorrelation diagrams, while the other one is the unit root test.

Graph Test.

1. Time series plot

According to the property that mean and variance of a stationary time series are constant, the time series plot should show that the process fluctuates randomly near a constant value and the ranges of fluctuation are similar. The time series is usually not stationary if there exists a significant trend or periodicity.

2. Autocorrelation function (ACF) plot

ACF is used to describe the degree of linear correlation between different observations in time series. It is proven that the stationary time series usually have short-term correlation. The time series is stationary if the autocorrelation function declines rapidly to zero and all the values fall into the confidence interval by lag 3. In contrast, the autocorrelation of a non-stationary series declines slowly.

Unit Root Test. The unit root test is used to check whether a time series needs to be differenced. The procedure is described in [15]. Among the unit root tests, the most widely used one is the Dickey–Fuller (DF) test, which is applicable to the AR(1) model:

$$Z_t = \varphi_1 Z_{t-1} + a_t = (1 - \varphi_1 B)^{-1} a_t = \sum_{k=0}^{\infty} \varphi_1^j a_{t-k}, \tag{36.6}$$

where $|\varphi_1| < 1$. Since the root of the characteristic equation $1 - \varphi_1 B = 0$ is φ_1^{-1}, another equivalent statement of the stationary form is that the root must be outside the unit circle. So it suffices to test whether the root of the characteristic equation is outside the unit circle, with, respectively, null and alternative hypothesis:

$$H_0 : \{Z_t\}\text{is non - stationary}, |\varphi_1| = 1, \text{a regular difference is needed}$$
$$H_1 : \{Z_t\}\text{is stationary}, |\varphi_1| < 1, \text{the series donot need to be deferenced}$$

The DF test is only applicable to the AR(1) model. In order to generalize the DF test and make it widely applicable to AR(p) processes, an augmented Dickey–Fuller (ADF) test was proposed in [16] with the same hypothesis and the decision rules and includes two other new terms: drift and trend.

White Noise Test. In order to verify whether a process is worth further time series modeling and analysis, it is needed to perform the white noise test. From the definition of the white noise, for any lag k, its autocorrelation coefficient is given by $\rho_k = 0$. It should be noted that this is the ideal situation. While in practice, most of the autocorrelation coefficients $\hat{\rho}_k$ are not equal to zero due to the finiteness of the sample sequence, but they fluctuate randomly around a value of 0 with a small float. According to the methods summarized by Wei [17], instead of considering each autocorrelation individually, the first m autocorrelation coefficients as a whole are considered and an index to determine whether a sequence is white noise or whether there exists a correlation between observations is constructed. The null and alternative hypotheses for the white noise test are, respectively:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_m = 0, \forall m \geq 1, \text{so}\{Z_t\}\text{is a white noise sequence}$$
$$H_1 : \text{for}\forall m \geq 1, \exists k \leq m \,\text{and}\, k \neq 0\text{that}$$
$$\rho_k \neq 0, \text{so}\{Z_t\}\text{is not a white noise sequence}$$

This is an approximate statistical hypothesis test that none of the autocorrelations of the series up to a given lag are significantly different from 0. If this is true for all m lags, then there is no information in the series to model and no ARIMA model is needed.

Methods of Order Specification. To determine the order (p, q) of ARMA models, SAS provides a list of the order combinations, which is mainly referred to ESACF, SCAN and MINIC methods.

The extended sample autocorrelation function (ESACF) method. Since the ACFs and PACFs of ARMA(p, q) model are all trailing, these two functions cannot be jointly used to determine the order (p, q). Considering this situation, Tsay and Tiao [18] proposed a general iterative regression method, and used the ESACF to estimate the order of the model. The method is applicable if the time series y_t belongs to ARMA(p, q) process, then by fitting AR(p) model to it, the estimate of the autocorrelation regression coefficients $\hat{\varphi}_i, i = 1, 2, \ldots, p$ will be inconsistent. Therefore, the residual error of regression must be introduced into the model as an explanatory variable, and when such process goes on until the q times the estimated model is as follows:

$$Z_t = \sum_{i=1}^{p} \varphi_i^{(q)} Z_{t-i} + \sum_{i=1}^{q} \alpha_i^{(q)} \hat{e}_{t-i}^{(q-i)} + e_t^{(q)}. \tag{36.7}$$

Now the estimator $\widehat{\varphi_i}^{(q)}$ will be consistent. Based on this idea, let $m = 0, 1, 2, ...,$ $\widehat{\varphi_i}^{(j)}$ is the jth iteration estimated autoregressive coefficient of the AR (m) model, then $\widehat{\rho_i}^{(m)}$ is defined as the sample autocorrelation function of the following model:

$$y_t = \left(1 - \widehat{\varphi_1}^{(j)} B - \widehat{\varphi_2}^{(j)} B^2 - \cdots - \widehat{\varphi_m}^{(j)} B^m\right) z_t. \tag{36.8}$$

Regarding the ESACF, there exists the following probabilistic convergence:

$$\hat{\rho}_j^{(m)} \xrightarrow{p} \begin{cases} 0, & 0 \le m - p \le j - q; \\ X \neq 0, & \text{otherwise} \end{cases}. \tag{36.9}$$

Because of this property, the distribution of the ESACF for ARMA (1,1) model can be displayed as in Table 36.1, which is characterized by the fact that all zeroes form a triangle with the vertex (1,1). Similarly for the general ARMA (p, q), the vertex of all zeroes is located at (p, q), which is the rule of identifying the order of the model. In fact, SAS provides two tables, one is for the estimate of ESACF and the other one is for the significance test.

The smallest canonical correlation coefficient (SCAN) method. Tsay and Tiao [19] firstly put forward this idea, and Choi [20] gave the concrete method of solving and judging ARMA(p, q) model. Only the conclusion of this method is given here. First, the SCAN of each model with different order combination is calculated, and then the table of SCAN similar to that of ESACF is formed. The only difference is

Table 36.1 ESACF for ARMA (1, 1) model, where *X* is a non-zero number

MA	0	1	2	3	…
AR					
0	*X*	*X*	*X*	*X*	…
1	*X*	0	0	0	…
2	*X*	*X*	0	0	…
3	*X*	*X*	*X*	0	…
…	…	…	…	…	…

that the judgment is based on the rectangle with zeroes being vertices so that the corresponding vertex position is the order of the model. In our project, SAS gives two tables, one for the estimate of SCAN coefficients and the other for chi-square test results of the coefficient significance.

The minimum information criterion (MINIC) method. The minimum information criterion (MINIC) method, proposed by Hannan and Rissanen [21], can tentatively identify the order of a stationary and invertible ARMA process. The MINIC table is constructed by computing Bayesian information criterion (BIC) for various autoregressive and moving average orders. Suppose L is the value of the likelihood function evaluated at the parameter estimates of ARMA(p, q), N is the number of observations, and k is the number of estimated parameters, the BIC of ARMA(p, q) model can be calculated as:

$$\mathrm{BIC}(p, q) = k\ \ln(N) - 2\ln(L) \tag{36.10}$$

Values of BIC(p, q) that cannot be computed are set to missing. For large autoregressive and moving average test orders with relatively few observations, a nearly perfect fit can result. This condition can be identified by a large BIC(p, q) negative value. The MINIC table can be in the form in Table 36.2. The model with the minimum BIC value is chosen as the best fitted one.

Methods of Parameters Estimation. There are various ways to estimate the parameters, such as moment estimation, least squares estimation, maximum likelihood

Table 36.2 MINIC table

MA	0	1	2	3	…
AR					
0	BIC(0,0)	BIC(0,1)	BIC(0,2)	BIC(0,3)	…
1	BIC(1,0)	BIC(1,1)	BIC(1,2)	BIC(1,3)	…
2	BIC(2,0)	BIC(2,1)	BIC(2,2)	BIC(2,3)	…
3	BIC(3,0)	BIC(3,1)	BIC(3,2)	BIC(3,3)	…
…	…	…	…	…	…

estimation and so on. In this work, method of maximum likelihood estimation is adopted, which is recommended by most experts using SAS for prediction.

Maximum likelihood method. According to the maximum likelihood method of time series analysis discussed by Guidolin and Pedio [22], under the maximum likelihood criterion, it is considered that the sample comes from the population with the highest probability of occurrence of this sample. Therefore, the maximum likelihood method for the unknown parameter's estimation is to make the likelihood function $L(\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q)$ reach the maximum, suppose $p(z_1, z_2, \ldots, z_n, \varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q)$ is the joint density function, L can be written as:

$$L(\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q) = p(z_1, z_2, \ldots, z_n, \varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q) \quad (36.11)$$

The distribution function of the population must be known to use the maximum likelihood. However, in the time series analysis, the distribution of population is often unknown. In order to facilitate calculation and analysis, it is usually assumed that the sequence follows multivariate normal distribution:

$$\begin{aligned} Z_t = \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \cdots + \varphi_p Z_{t-p} + a_t - \theta_1 a_{t-1} \\ - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}, \tilde{z} = (z_1, z_2, \ldots, z_n)', \end{aligned} \quad (36.12)$$

$$\tilde{\beta} = (\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q)', \quad (36.13)$$

$$\sum_n = E(\tilde{z}'\tilde{z}) = \Omega\sigma_a^2. \quad (36.14)$$

The likelihood function of $\tilde{z}$ is

$$L(\tilde{\beta}) = p\tilde{\beta} = (2\pi)^{-n/2} \left| \sum_n \right|^{-1/2} \exp\left\{ -\frac{\tilde{z}' \sum_n^{-1} \tilde{z}}{2} \right\}. \quad (36.15)$$

The log likelihood function is

$$l(\tilde{\beta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma_a^2) - \frac{1}{2}\ln|\Omega| - \frac{1}{2\sigma_a^2}\left[\tilde{z}'\Omega^{-1}\tilde{z}\right]. \quad (36.16)$$

The system of likelihood equations can be obtained by computing the partial derivatives of the unknown parameters of the logarithmic likelihood function.

Theoretically, solving the likelihood equations yields the maximum likelihood of the unknown parameter. However, since $\tilde{z}'\Omega^{-1}\tilde{z}$ and $\ln|\Omega|$ is not an explicit expression of the parameter, the likelihood equations are actually composed of $p + q + 1$ transcendental equations, which usually requires a complex iterative algorithm to find the maximum likelihood of the unknown parameter.

The maximum likelihood method makes full use of the information provided by each observation, so its estimation accuracy is high, and it also has good statistical properties such as consistency and progressive validity.

Diagnostic test. In this test, the goodness of fit and the accuracy of the model are measured and the correlation test and the normality test on the residual series are performed. The following two kinds of criterion will be used to measure the goodness of fit for a model:

Akaike's information criterion (AIC). Akaike [23] defined AIC as

$$\mathrm{AIC} = -2\ln(L) + 2k, \tag{36.17}$$

where L is the value of the likelihood function evaluated at the parameter estimates, N is the number of observations and k is the number of estimated parameters. The first term of the AIC measures the goodness of fit of the ARMA model to the data, and the second term is called the penalty function of the criterion because it penalizes a candidate model by the number of parameters used. Therefore, the model with the minimum AIC value should be chosen.

Schwarz's Bayesian information criterion (SBC). Schwarz [24] defined AIC as

$$\mathrm{SBC} = -2\ln(L) + \ln(N)k, \tag{36.18}$$

Similarly, the model with the minimum SBC value should be chosen. The penalty for each parameter is 2 for AIC and $\ln(N)$ for SBC, so compared to AIC, SBC tends to select a lower-order model when sample size is moderate or large.

There are other two kinds of criterion to measure the accuracy of a model's predictions will be used. One can refer to [25] for the detailed description.

Mean absolute percentage error (MAPE). The MAPE is a common measure of forecast error in time series analysis. It usually expresses accuracy as a percentage and is defined by the formula:

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n} |Z_t - F_t|/Z_t, \tag{36.19}$$

where Z_t is the actual value and F_t is the forecast value.

Mean square error (MSE). The MSE is measure of the differences between prediction values and the actual values. It is defined by:

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(Z_t - F_t)^2, \tag{36.20}$$

where Z_t is the actual value and F_t is the forecast value.

36.4 Empirical Results

36.4.1 Introduction to Database

In this work, the original data was provided by the bus companies in the city of Jiaozuo in China, including the bus IC card payment records among the six bus lines in the city, during the period of time January 01, 2018 to March 31, 2018. Table 36.3 shows a part of the originally collected data. The whole dataset consists of 2,874,878 rows (records), each with eight component variables. The meaning of the variables used in this work is shown in Table 36.4.

36.4.2 Data Preprocessing

The first phase of the data analysis is to process the data in order to construct a time series. The steps to obtain the descriptive statistics and the time series plot are the follows.

Step 1: **Standardization of the raw data**. The variables which are not straightforward numeric or character, such as 'SITE_TIME', need to be standardized in the format that the SAS can recognize and interpret.

Step 2: **Data extraction**. The daily passenger volume from the original database is extracted. For this, the 'DATA' and 'PROC' procedures are mainly used in SAS to create the datasets.

Step 3: **Construct the time series**. After the datasets, including daily passenger flow volume in each bus line, are constructed, the graphical procedure in SAS is used to plot the time series for each bus line. In this work, the line No. 18 is chosen for case study. The time series plot of the daily passenger volume in the line No. 18 during the period January 01, 2018 to March 31, 2018 is shown in Fig. 36.1.

36.4.3 Model Building

Case 1: ARMA modeling with the original time series. According to the results of ADF test shown in Table 36.5, the *p*-value is less than 0.05 for a lag of 0, indicating that the null hypothesis can be rejected and the sequence is stationary, so the ARMA model is suitable to the original data. After calculating the BIC of the models with different order combinations, SAS shows the optimal order for the order selection by the ascending order of BIC value. And three candidate models with minimum BIC values, namely $\mathrm{AR}(3)$, $\mathrm{ARMA}(1, 1)$and $\mathrm{ARMA}(1, 3)$ are chosen. The results of the parameter estimation and fitting statistics for each candidate models are summarized in Tables 36.6 and 36.7, respectively.

Table 36.3 Part of raw data in the database

	LINE_NO	BUS_NO	MACH_NO	IS_UP_DOWN	LABEL_NO	UP_PASSENGER	DOWN_PASSENGER	SITE_TIME
28750	18	2420	1649572	1	7	8	0	02JAN2018:19:02:51
28751	18	2420	1649572	1	8	2	0	02JAN2018:19:04:12
28752	18	2420	1649572	1	9	2	0	02JAN2018:19:05:51
28753	18	2420	1649572	1	10	0	0	02JAN2018:19:06:45
28754	18	2420	1649572	1	11	1	0	02JAN2018:19:08:30
28755	18	2420	1649572	1	12	1	0	02JAN2018:19:09:49
28756	18	2420	1649572	1	13	0	0	02JAN2018:19:12:35
28757	18	2420	1649572	1	14	0	2	02JAN2018:19:14:42
28758	18	2420	1649572	1	15	0	0	02JAN2018:19:16:38
28759	18	2420	1649572	1	16	2	0	02JAN2018:19:18:05
28760	18	2420	1649572	1	17	0	0	02JAN2018:19:19:38
28761	18	2420	1649572	1	18	0	0	02JAN2018:19:20:36
28762	18	2420	1649572	1	19	0	0	02JAN2018:19:22:00

Table 36.4 Descriptions of the variables used in this work

Variable	Meaning
LINE_NO	Bus line number
BUS_NO	Bus number in a certain bus line
IS_UP_DOWN	Direction of the bus (1: up direction; 0: down direction)
LABEL_NO	Bus station order number in a certain bus line
UP_PASSENGER	Number of passengers getting on the bus
SITE_TIME	Date and time when the record is collected

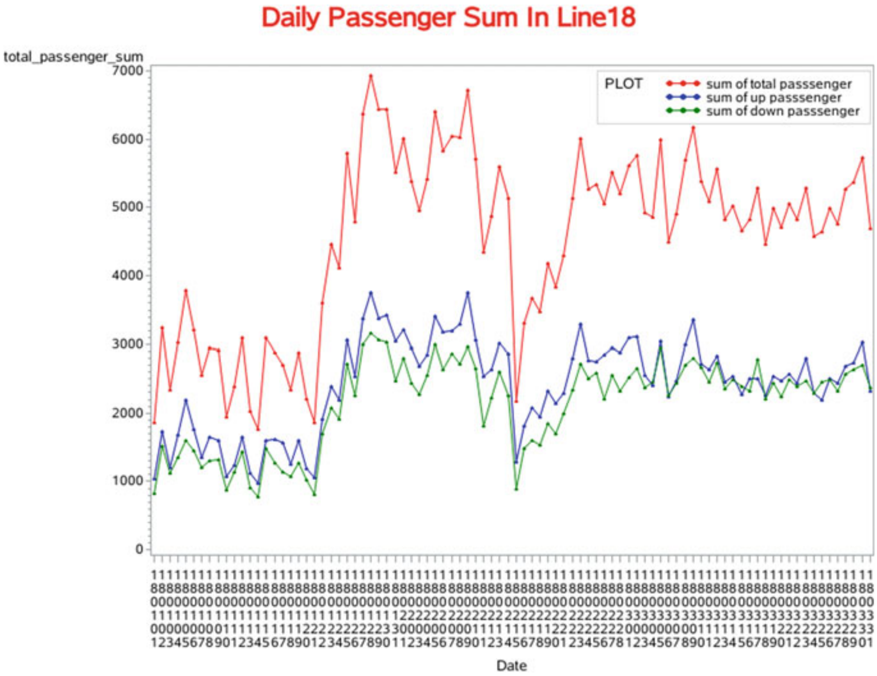

Fig. 36.1 Time series plot of daily passenger volume in line No. 18, including total passenger volume and the ones in two directions

Table 36.5 ADF test results

Augmented Dickey–Fuller unit root tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero mean	0	−0.7683	0.5129	−0.50	0.4972		
	1	−0.3326	0.6052	−0.28	0.5817		
Single mean	0	−17.2117	0.0180	−3.28	0.0189	5.46	0.0283
	1	−10.8949	0.0984	−2.40	0.1441	2.95	0.3284
Trend	0	−21.7232	0.0361	−3.50	0.0454	6.33	0.0577
	1	−14.4621	0.1790	−2.61	0.2773	3.49	0.4875

Table 36.6 Parameter estimation results of three candidate models (in Case 1)

(p, q)	MU	φ_1	φ_2	φ_3	θ_1	θ_2	θ_3
(3, 0)	2227.2	0.65563	0.05421	0.27748			
(1, 1)	2412.4	0.98886	–	–	0.37098	–	–
(1, 3)	2154.8	0.98111			0.35491	0.08107	−0.30409

Table 36.7 Fitting statistics of three candidate models (in Case 1)

(p, q)	AIC	SBC	MAPE
(3,0)	1445.263	1455.262	13.57413
(1,1)	1446.71	1454.209	13.82645
(1,3)	1445.193	1457.693	13.81004

Although ARMA(1,1) and ARMA(1,3) have the smallest SBC and the smallest AIC value, respectively, their MAPE values are much higher than that of AR(3). In comparison, AR(3) model has the smallest MAPE value indicating the highest prediction precision, and either its AIC or BIC value is slightly higher than the minimum indicating a relatively good goodness of fit. Therefore, AR(3) is chosen as the optimal model in this case. After implementing the parameter estimation by goodness of fit test for AR(3) model in SAS, all the AR coefficients are significant, so the optimal model is determined as below.

$$Z_t = 0.65563Z_{t-1} + 0.05421Z_{t-2} + 0.27748Z_{t-3} + a_t \tag{36.21}$$

As shown in Fig. 36.2, the residual diagnostics in SAS shows ACF value of the residual sequence is almost 0, and the white noise probability is greater than 0.05, which indicates that there is no dependence between the residuals and the AR(3) model has extracted all the useful information from the historical time series. Besides, the histogram and QQ-plot of residuals (Fig. 36.3) show that the residual sequence follows normal distribution, indicating the model is adequate. The ten-step ahead prediction of the passenger flow volume by using the model (36.21) and the comparison between actual and prediction are shown, respectively, in Table 36.8; Fig. 36.4.

Although the diagnostic results show that AR(3) is adequate for the sequence fitting, 95% confidence interval of the prediction is very wide. Since the wider the confidence region is, the lower the prediction accuracy is, the prediction especially in the long term may not be accurate.

Case 2: ARMA modeling with the first-order differenced time series. According to the ACF plot shown in Fig. 36.5, even the autocorrelation decreases exponentially, it does not fall into the confidence interval until lag 5. Considering that the ACF decays gradually, not rapidly to zero, the time series is regarded as non-stationary and needs to be differenced, so the ARIMA model is applied to fit the data. And two models with minimum BIC value are chosen as candidate models,

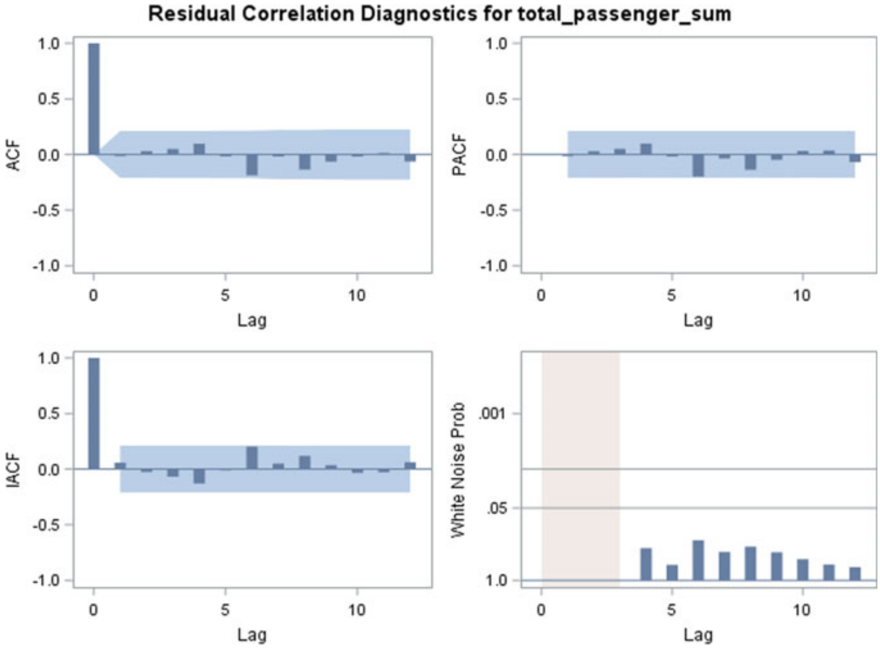

Fig. 36.2 Residual correlation diagnostics for AR(3) model

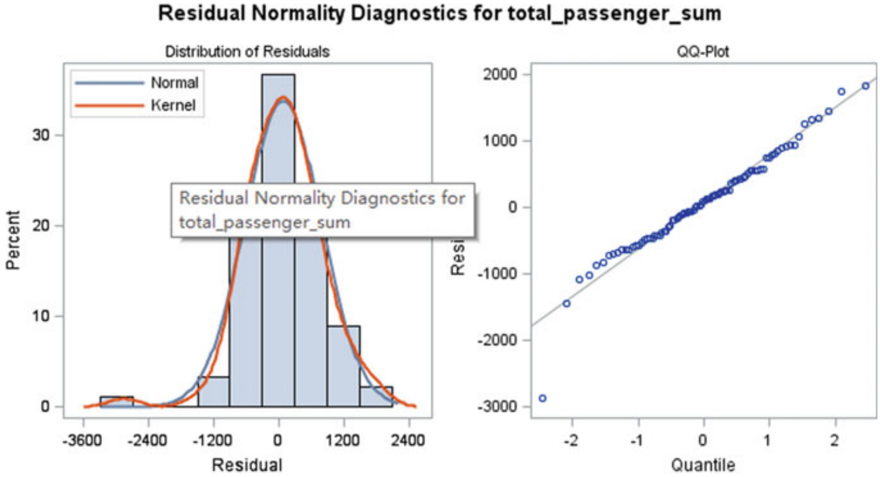

Fig. 36.3 Residual normality diagnostics for AR(3) model

Table 36.8 Prediction based on AR(3) model

Date	Forecast for the daily passenger volume	Std error of forecast	Lower 95% confidence limit	Upper 95% confidence limit
20180401	4910.496318	726.7519712	3486.088629	6334.904007
20180402	5093.005425	869.0218861	3389.753826	6796.257023
20180403	4937.714098	937.5256512	3100.197587	6775.230609
20180404	4904.203801	1043.474524	2859.031316	6949.376287
20180405	4924.457867	1137.024275	2695.931238	7152.984497
20180406	4892.829962	1211.618031	2518.102259	7267.557665
20180407	4863.893423	1284.335326	2346.642439	7381.144406
20180408	4848.827417	1353.560756	2195.897083	7501.75775
20180409	4828.60494	1417.048994	2051.239948	7605.969932
20180410	4806.500485	1476.948875	1911.733884	7701.267087

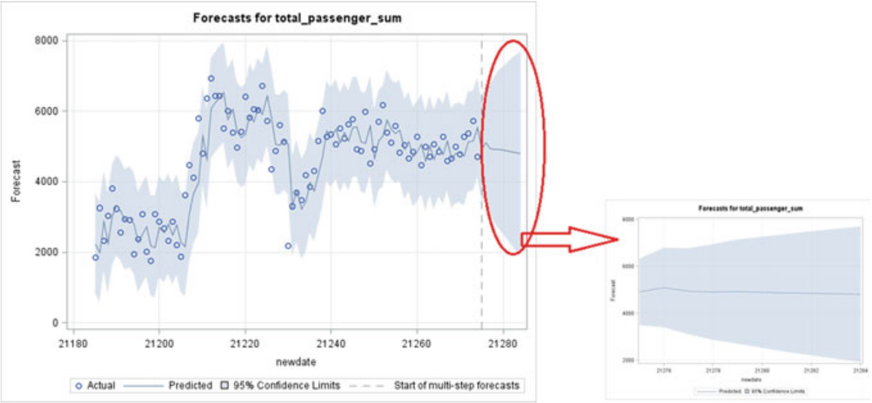

Fig. 36.4 Actual values against prediction based on $AR(3)_{ij}$ model

namely ARIMA(0, 1, 1)and ARIMA(2, 1, 0). The results of the parameter estimation and fitting statistics for each candidate models are summarized in Tables 36.9 and 36.10, respectively.

It can be seen ARIMA(2, 1, 0) has a smaller AIC value indicating a higher goodness of fit, and a smaller MAPE value indicating a higher prediction precision, while its SBC value is slightly higher than ARIMA(0, 1, 1). Therefore, ARIMA(2, 1, 0) model is chosen as the optimal one for the first-order differenced time series. After implementing the parameter estimation by goodness of fit test, all the coefficients are significant. So the optimal model is determined as below.

$$(1 - B)Z_t = -0.33412(1 - B)Z_{t-1} - 0.28956(1 - B)Z_{t-3} + a_t \qquad (36.22)$$

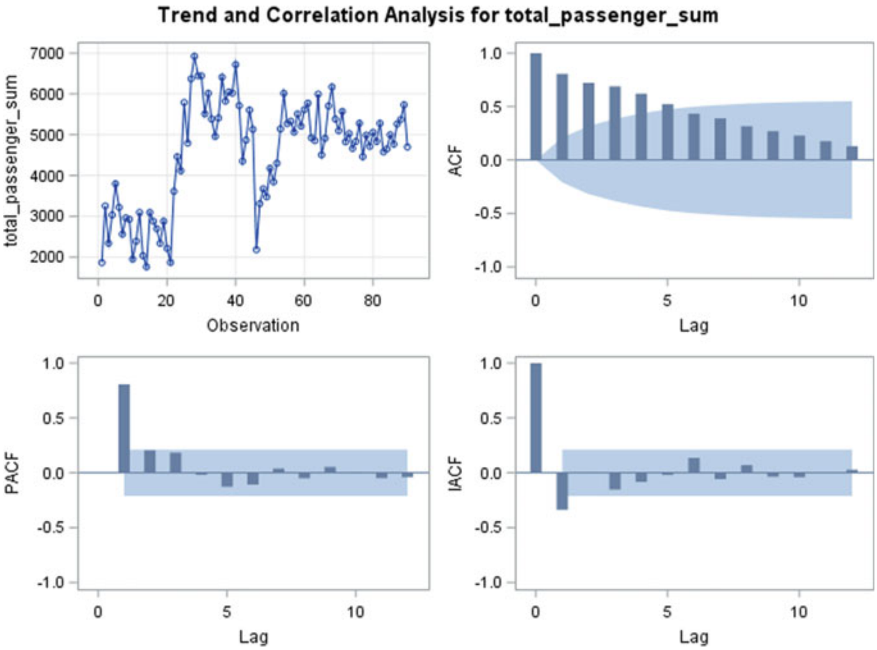

Fig. 36.5 Trend and ACF plots for the original time series

Table 36.9 Parameter estimation results of two candidate models (in Case 2)

(p, d, q)	MU	φ_1	φ_2	φ_3	θ_1	θ_2	θ_3
(0, 1, 1)	32.28610				0.37507		
(2, 1, 0)	32.81746	−0.33412	−0.28956	–		–	–

Table 36.10 Fitting statistics of three candidate models (in Case 2)

(p, d, q)	AIC	SBC	MAPE
(0, 1, 1)	1429.956	1434.934	13.9703
(2, 1, 0)	1428.162	1435.628	13.64128

By performing the residual diagnostics (similar to Case 1), it is observed that there is no dependence between the residuals and the ARIMA(2, 1, 0) model has extracted all the useful information from the time series. Besides, the residual sequence follows normal distribution, indicating the model is adequate. The ten-step ahead prediction of the passenger flow volume by using the model (36.22) and the comparison between actual and prediction are shown, respectively, in Table 36.11; Fig. 36.6.

Although the diagnostic results show that ARIMA(1, 2, 0) is adequate for the sequence fitting, its 95% confidence region of the prediction from the model is still very wide.

Table 36.11 Fitting statistics of three candidate models (in Case 2)

Date	Forecast for the daily passenger volume	Std error of forecast	Lower 95% confidence limit	Upper 95% confidence limit
20180401	5073.267	737.6109	3627.576	6518.958
20180402	5105.553	869.797	3400.783	6810.324
20180403	5137.84	984.39	3208.471	7067.208
20180404	5170.126	1086.969	3039.706	7300.545
20180405	5202.412	1180.668	2888.344	7516.479
20180406	5234.698	1267.46	2750.522	7718.873
20180407	5266.984	1348.678	2623.624	7910.343
20180408	5299.27	1425.275	2505.783	8092.757
20180409	5331.556	1497.96	2395.608	8267.504
20180410	5363.842	1567.279	2292.033	8435.652

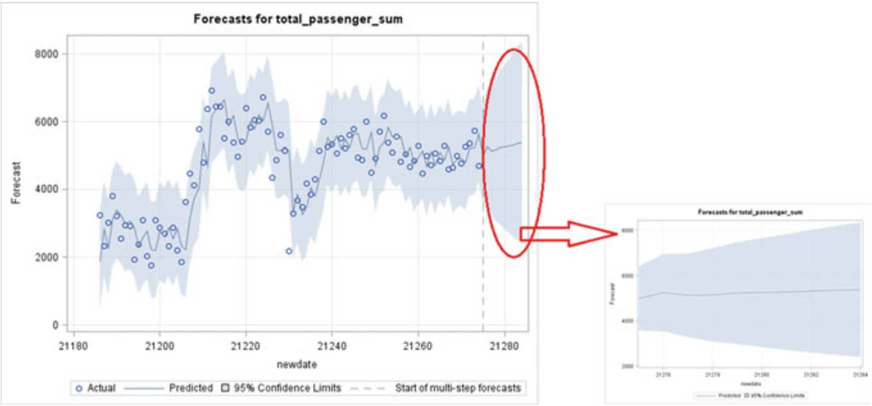

Fig. 36.6 Actual values against forecasts based on ARIMA(2, 1, 0) model

Case 3: ARMA modeling with quadratic function trend. According to the original time series plot in Fig. 36.1, it is observed that the passenger flow time series may have a quadratic trend. The two trend variables, _LINEAR_ and _SQUARE_ (as shown in Table 36.12), representing linear and quadratic relationships, respectively, are pre-generated.

Then, the same steps as in the previous two cases are followed to build a quadratic ARMA model. And three models with minimum BIC value as candidate models are chosen, namely Quadratic + AR(3), Quadratic +ARMA(1, 1)and Quadratic + ARMA(1, 3). The results of the parameter estimation and fitting statistics for each candidate models are summarized, respectively, in Tables 36.13 and 36.14.

Compared with the other two models, the Quadratic + ARMA(1, 3) model has the smallest AIC and MAPE indicating the highest goodness of fit and the highest

Table 36.12 Part of dataset with two trend variables

	Date	total_passenger_sum	_LINEAR_	_SQUARE_
1	180101	1863	1	1
2	180102	3251	2	4
3	180103	2339	3	9
4	180104	3033	4	16
5	180105	3793	5	25
6	180106	3219	6	36
7	180107	2558	7	49
8	180108	2957	8	64
9	180109	2922	9	81
10	180110	1950	10	100

prediction precision. Although Quadratic + ARMA(1,1) has the smallest SBC, its MAPE is the highest indicating the lowest prediction accuracy. Therefore, the Quadratic + ARMA(1, 3) is chosen as the optimal model with quadratic trend.

After implementing the parameter estimation by goodness of fit test, it can be seen that not only the MA and AR coefficients, but also the coefficients of linear and quadratic trend variables are significant. So the optimal model is determined as below.

$$\begin{aligned} Z_t = {} & 113.79672t - 0.92369t^2 + 0.76222Z_{t-1} \\ & + a_t - 0.26487a_{t-1} - 0.05254a_{t-2} + 0.35486a_{t-3} \end{aligned} \tag{36.23}$$

By performing the residual diagnostics (similar to Case 1), it is observed that there is no dependence between the residuals and the Quadratic + ARMA(1, 3) model has extracted all the useful information from the time series. Besides, the histogram and QQ-plot (obtained by using the same method in Case 1) show that the residual sequence follows normal distribution, indicating the model is adequate.

Using the model (36.23), the ten-step ahead prediction and the comparison between actual and prediction are shown, respectively, in Table 36.15; Fig. 36.7.

The 95% confidence region width is significantly narrower, but the prediction does not describe the rapid growth at the end of the sequence, so probably it is caused by some external factors such as weather and holiday policies. If further improvements are needed, the external influences must be included in the model.

36.5 Conclusion

In this work, the prediction on the passenger flow volume in the bus transpiration system is performed, by using three kinds of time series models: AR, ARIMA and

Table 36.13 Parameter estimation results of three candidate models (in Case 3)

Quadratic + (p, q)	MU	φ_1	φ_2	φ_3	θ_1	θ_2	θ_3	a	b
Quadratic + (3,0)	2009.0	0.57459	0.006668	0.20647	–	–	–	107.78168	−0.86291
Quadratic + (1,1)	2075.2	0.82535	–	–	0.26871	–	–	106.02820	−0.85541
Quadratic + (1,3)	1895.5	0.76222	–	–	0.26487	0.05254	−0.35486	113.79672	−0.92369

Table 36.14 Fitting statistics of three candidate models (in Case 3)

Quadratic + (p, q)	AIC	SBC	MAPE
Quadratic + (3, 0)	1442.351	1457.35	13.73232
Quadratic + (1, 1)	1443.143	1455.642	14.00114
Quadratic + (1, 3)	1439.571	1457.07	13.43795

Table 36.15 Forecasts based on Quadratic + ARMA(1, 3) model

Date	Forecast for the daily passenger volume	Std error of forecast	Lower 95% confidence limit	Upper 95% confidence limit
20180401	4934.188	693.2455	3575.452	6292.924
20180402	5074.388	774.2533	3556.879	6591.896
20180403	4629.014	806.6707	3047.968	6210.059
20180404	4536.912	908.7964	2755.704	6318.121
20180405	4450.85	963.1695	2563.073	6338.628
20180406	4368.952	993.3936	2421.936	6315.967
20180407	4289.787	1010.538	2309.169	6270.406
20180408	4212.268	1020.367	2212.386	6212.15
20180409	4135.564	1026.034	2124.575	6146.553
20180410	4059.041	1029.312	2041.627	6076.455

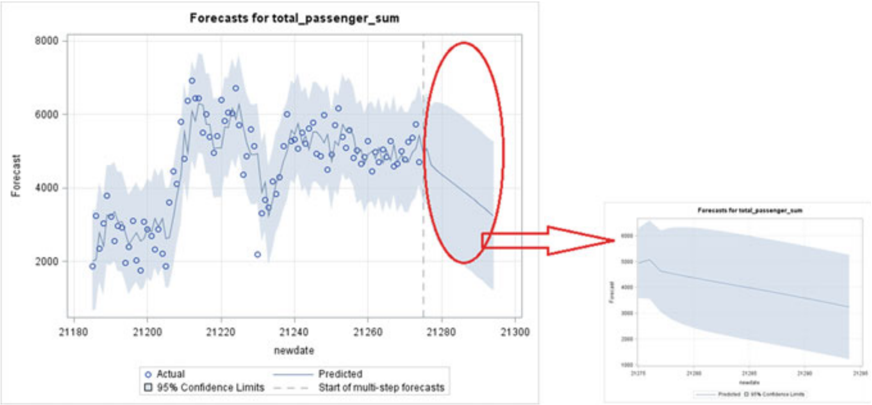

Fig. 36.7 Actual values against forecasts based on Quadratic + ARMA(1, 3) model

Table 36.16 Fitting statistics of three optimal models

Model	AIC	SBC	MAPE	MSE
AR(3)	1445.263	1455.262	13.57413	504694.3
ARIMA (2,1,0)	1428.162	1435.628	13.64128	531843.6
Quadratic + ARMA (1,3)	1439.571	1457.07	13.43795	443210.1

quadratic ARMA. At first, the bus IC card payment records were transformed into a time series, which represents the daily passenger volume in line No. 18. Then, the time series analysis was used and two optimal models, AR(3) and ARIMA(2, 1, 0), were found. Both models performed well in terms of goodness of fit but failed to attain accurate predictions. In order to achieve a higher prediction accuracy, the ARMA model with the quadratic trend was further explored and combined, and a Quadratic + ARMA(1, 3)model was established for the time series, which achieves a better balance between fitting and forecasting. The fitting statistics of those models are shown in Table 36.7.

Each model has its own advantages and disadvantages. They are discussed as follows one by one.

- **AR**(3): The AR(3) model has no obvious advantages and disadvantages, because its performance is not outstanding either in goodness of fit or prediction accuracy. Its SBC, MAPE and MSE value are all in the middle level, except its AIC value is slightly higher than the other two models. The only advantage worth mentioning is that since differencing the process is not needed, this model is the simplest and the most straightforward one, and the cost is thus the lowest for the application.
- **ARIMA**(2, 1, 0): In terms of fitting effect, the ARMA(2, 1, 0) model has the lowest AIC and SBC values indicating the highest goodness of fit. However, it owns the highest MAPE and MSE among the three models, indicating the greatest deviation between the predicted value and the true value. Moreover, its prediction confidence region widens over time, so it may perform poorly in the long-term prediction. But since it has the best fitting effect, it can accurately describe the surge trend at the end of the original time series, so the prediction result will be reliable when the model is used to predict the most recent value.
- **Quadratic + ARMA**(1, 3): Compared with the other two models, Quadratic + ARMA(1, 3) model has the smallest MAPE and MSE value, so it achieves the highest prediction accuracy. Most importantly, this model has a unique advantage over the others, and it has a narrower prediction confidence interval of a constant width over time, so it will perform more effectively with high prediction accuracy.

The initial objective of this project and the main demand from the traffic management is to improve the forecast accuracy. Due to this, the accuracy of the prediction is the most important factor for the solution performance evaluation. So it can be concluded that the Quadratic + ARMA(1, 3) model is the most appropriate, compared to the other two models. Although ARIMA(2, 1, 0) model fits the current data the best and its short-term prediction shows relatively higher volatility, it may be more useful for short-term prediction.

36.6 Open Questions and Potential Improvements

Although the ARMA model with quadratic function trend performs best in our case, its application range is limited, because the time sequence must show a quadratic trend. In the reality, only the short-term change of passenger flow may show such a trend. For the long-term daily passenger flow, if the data span is more than one year, it usually fluctuates within a limited range near a fixed value. So the stationary time series model may be more suitable for such kind of data. In addition, in view of the change of daily passenger flow in certain city, a seasonal factor with week cycle may be considered because of the difference of the commuting time between weekdays and weekends. In this case, a seasonal ARIMA model may be built to fit the series.

As mentioned in the end of Sect. 36.4, the time series method has limitations. When the prediction time span is long, only a rough future trend line can be obtained, but not the specific volatility. In order to accurately describe the future fluctuations, more external factors, such as weather, temperature, holidays and events, might be introduced into the model. When the historical data is updated continuously and the sample size is increasing, the algorithm should be updated and adjusted accordingly.

Acknowledgements The authors would like to acknowledge the support of Xiongdi Shenzhen Emperor Technology Company for the research fund (RDS10120190006) and the data used in this work.

References

1. Chen, Y., Wang, D.: Intelligent Traffic Information Collection, Analysis and Application. China Communication Press (2011)
2. Zhou, C., Zhang, Z., Tang, W.: System and methods of passenger demand prediction on bus network. Comput. Sci. **45**, 527–535 (2018)
3. Delwiche, L.D., Slaughter, S.J.: The little SAS book: A Primer, 5th edn. SAS Institute (2012)
4. Yule, G.U.: On a method of investigating periodicities in disturbed series, with Special Reference to Wolfer's sunspot numbers. Philos. Trans. R. Soc. London, Ser. A **226**, 267–298 (1927)
5. Walker, G.: On periodicity in series of related terms. Proc. R. Soc. London, Ser. A. **131**(818) (1931)
6. Slutzky, E.: The summation of random causes as the source of cyclic processes. Econometr. Soc. **5**(2), 105–146 (1937)
7. Wold, H., Kendall, M.: A study in the analysis of stationary time series. J. Roy. Stat. Soc. **102**(2), 295–298 (1939)
8. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, 5th edn. Wiley, London (2015)
9. Xia, K.: Deep Analysis of SAS: Data Processing, Analytical Optimization and Business Applications. China Machine Press (2015)
10. Zhu, H.: N-day average volume based time-series analysis for passenger flow of metro. In: 2010 International Conference on Multimedia Information Networking and Security (2010)
11. Li, Z., Bi, J., Li, Z.: Passenger flow forecasting research for airport terminal based on SARIMA time series model. IOP Conf. Ser. Earth Environ. Sci. **100**(1), 1–7 (2017)
12. Xu, X., Dou, Y., Zhou, Z., Liao, T., Lu, Y., Tan, Y.: Railway passenger flow forecasting based on time series analysis with big data. In: Chinese Control and Decision Conference, pp. 3584–3590 (2018)
13. Jamil, M.S., Akbar, S.: Taxi passenger hotspot prediction using automatic ARIMA model. In: 2017 3rd International Conference on Science in Information Technology (2017)

14. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods, 2nd edn. Springer, New York (2006)
15. Harris, R., Sollis, R.: Applied Time Series Modelling and Forecasting. Wiley, London (2003)
16. Dickey, D.A., Fuller, W.A.: Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. **74**(36), 427–431 (1979)
17. Wei, W.S.W.: Time Series Analysis: Univariate and Multivariate Methods, 2nd edn. Pearson Addison Wesley (2006)
18. Tsay, R.S., Tiao, G.C.: Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. J. Am. Stat. Assoc. **79**(385), 84–96 (1984)
19. Tsay, R.S., Tiao, G.C.: Use of canonical analysis in time series model identification. Oxford Univ. Press **72**(2), 299–315 (1985)
20. Choi, B.: ARMA Model Identification. Springer, New York (1992)
21. Hannan, E.J., Rissanen, J.: Recursive estimation of mixed autoregressive moving-average order. Oxford Univ. Press **69**(1), 81–94 (1982)
22. Guidolin, M., Pedio, M.: Essentials of Time Series for Financial Applications. Academic Press, London (2018). (an imprint of Elsevier)
23. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
24. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
25. Xu, G.: Statistical Forecasting and Decision-making. Shanghai University of Finance & Economics Press (2016)