GA-ARIMA Model-Based Analysis of Arrival time at Bus Stop

Guo Xiang¹,a, Guan Deyong¹,b*, Xue Yuxuan¹,c
¹Shandong University of Science and Technology School of Transportation Qingdao, Shandong
²Author’s e-mail: 15054213381@163.com
³*e-mail: gx449044865@163.com

Abstract—Background With the ever-changing urban development and a growing number of vehicles, the road congestion problem has become a hotspot issue at present. As an important constituent part of road traffic system, public traffic system has become an effective force relieving the traffic congestion problem. Objective To explore how to better reduce the prediction error of arrival time at bus stop on basis of actual GPS bus data. Method The arrival time at bus stop was analyzed based on GPS data, a time series optimization model based on genetic algorithm was established, followed by the prediction error analysis of arrival time at bus stop. Result The prediction result of GA-ARIMA model is superior to that of the traditional time series model, which indicates the effectiveness of this model.

1. INTRODUCTION
Vigorously developing the public transportation has become one of effective means to solve the problem of traffic congestions. That bus arrival time is not controlled by passengers is an important factor influencing the attractiveness of bus, which, however, can be practically strengthened by improving the prediction accuracy of bus arrival time. In this way, passengers can make effective travel plans according to bus arrival time. However, how to realize practical and effective prediction of bus arrival time is a problem needing an urgent solution.

Many foreign scholars have started investigated the prediction of bus arrival time very early. For instance, Marko Celan proved that both data content of public traffic system and bus operating time would impact the prediction accuracy of bus arrival time at bus stop[1]; Tianli Tang studied the prediction problem of travel time within uncertain distance intervals between bus stop based on neural network and support vector machine (SVM) technology[2]; Mansur As proposed a gradient boosted decision tree (GBDT) algorithm-based machine learning model, and then predicted the arrival time at bus stop and the traffic flow by combining weather and travel factors[3]. Although domestic (China) researches regarding the prediction problem of bus arrival time have a late start, the research effect is evident with fast progress. To be specific, Mei Deng analyzed the variation trend of arrival time in different time periods based on a generalized arrival time model at bus stop[4]; Based on bus positioning data, Yishao Huang used the IPSO-ELM model to analyze factors such as bus stop spacing and arrival time and predict passenger flow volume at each bus stop[5]; With data samples of different types and time granularities, Mei Li conducted a comparative analysis of the results obtained through three time prediction models[6]; Cheng Wang analyzed the prediction accuracy of SLMBP model for
passenger flow volume at bus stop based on massive bus data[7]; Panlong Lv used machine learning and decision-making tree model to identify and analyze the bus commuting model based on the integration of various data types[8]; Yulong Shan predicted GDP data according to the traditional ARIMA model and genetic algorithm-based neural network optimization model, and then compared the models after optimization[9].

This study aims at optimizing the prediction model for arrival time at bus stop. In consideration of low accuracy of parameter estimates obtained through the traditional time series model, a genetic algorithm-based ARIMA optimization model was built in this study, the parameter estimates obtained through the time series were optimized by virtue of the global optimization advantage of genetic algorithm, and then, according to the analysis of practical cases, it was verified that the prediction result of the improved model was better than that of the traditional model.

2. DATA PROCESSING AND ARRIVAL TIME ANALYSIS OF BUS STOP

2.1. Data preprocessing

The bus data were collected at acquisition terminal and then saved in a database. The collected data in original format included over 30 data fields such as bus ID and data ID. The accuracy and speed of data analysis were improved through data preprocessing—data cleaning and valid data extraction, concretely as follows:

1) Reserve useful data fields: A total of 11 useful data fields were extracted from the 30 data fields and saved as csv files.

2) Delete repeated data: The original bus data repeatedly uploaded were deleted.

3) Process abnormal data: The original data not completely identical were cleaned.

4) Process data drift: As for multiple continuous data which shared the same speed (0) but had not exactly the same longitude and latitude, it would be judged that all data except for the first one were drift data, and the longitude and latitude of all subsequent data with speed of 0 were substituted by the longitude and latitude of the first data.

5) Data compensation: Some data fields were incomplete or missing in the uploaded original bus data, so the data should be compensated according to continuous data.

6) Process data positioning error: A threshold value was set for the bus speed. If this threshold value was exceeded, the involved data would be identified as positioning error data and replaced according to the longitude and latitude positions of the two data before and after it, and a certain weighting compensation was made for the wrong data positioning direction.

2.2. Arrival time analysis of bus stop

Generally speaking, resident travel is of certain regularity, which will have a certain bearing on bus arrival time at bus stop. This is the most obvious data feature used to investigate the differences between public traffic vehicles and social vehicles.

![Fig. 1 Scatter diagram of arrival time at Jinggangshan Road Fuchunjiang Road Bus Stop](image)

As shown in Fig. 1, the clustering property of arrival time is stronger in rush hours, during which the peak value of arrival time can easily appear; The discreteness of arrival time is stronger in hollow period.
Although it takes additional time for buses to stop at bus stop in comparison with cars when running on road, their travel time distribution characteristics are similar to a certain extent. Therefore, two fitting models for car travel time distribution, that is Normal Distribution and Lognormal Distribution, were used to probe into the distribution characteristics of bus travel time. The arrival time histograms in rush hours and hollow period and the distribution fitting graph are shown in Fig. 2.

![Fig. 2 Arrival time histograms in rush hours and hollow period and distribution fitting graph](image)

The fitting effect of arrival time distribution at bus stop is presented in Table 1, where blue line represents normal distribution curve and red line denotes lognormal distribution curve. The arrival time data conform to lognormal distribution, and the lognormal distribution fitting is superior to normal distribution fitting.

| data phase | distribution model | Test method of fitting effect | P value  | Decision making level (5%) |
|------------|--------------------|------------------------------|----------|---------------------------|
| peak       | Normal distribution| K-S                          | 0.00068  | refuse                    |
|            | Lognormal distribution| K-S                       | 0.31109  | Can't refuse               |
| off peak   | Normal distribution| K-S                          | 0.00041  | refuse                    |
|            | Lognormal distribution| K-S                       | 0.48321  | Can't refuse               |

3. OVERVIEW OF TIME SERIES MODEL

3.1. Time series model

The time series analysis model is a relatively accurate data prediction model established using the principle and method of probability theory and statistics under fixed and small sample data size. This model is applied to parameter assumption and further to prediction, adaptive control, etc.

ARIMA (p,d,q) model, namely autoregressive integrated moving average model, is one kind of time series analysis method. This model is mainly applied to the time series prediction under unsteady data circumstance. This model is seen in Formula 1:

$$\left(1-\sum_{i=1}^{p}\theta_i B^i\right)\left(1-B^d\right)x_t = \left(1-\sum_{i=1}^{q}\theta_i B^i\right)\varepsilon_t$$

(1)

Where: $p$ —— autoregression order;

$d$ —— difference order;

$q$ —— moving average order.

The difference operation is performed for the data using difference method at order d so that the data become steadier. By combining autoregression model AR (p) and moving average model MA (q), the post-difference autoregressive integrated moving average model ARIMA (p, d, q) is obtained as seen in Formula 2.
\[
\begin{align*}
\phi(B)V^d x_t &= \theta(B)e_t \\
E(e_t) &= 0, Var(e_t) = \sigma^2, E(e_t e_s) = 0, s \neq t \\
E(e_t e_s) &= 0, \forall s < t 
\end{align*}
\]  

Where \( V^d = (I-B)^d \) and \( V^d x_t \) are series after the difference operation of \( x_t \) for \( d \) times. At the time, the series \( V^d x_t \) is steady.

In the solving process of practical problems, the time series closely related to season is called seasonal time series. The time series presents periodic change, so the data levels at special time within each period are basically the same. If the data observed at some time is deducted by that observed at the corresponding time in the next period, the data periodicity can be reduced and eliminated as much as possible, and then a new steady time series is obtained.

\[
V_s x_t = (I-B)^s y_t = x_t - x_{t-s}
\]

is set and called season difference operator, and then the following formula is obtained:

\[
V_s x_t = (I-B)^s y_t = x_t - x_{t-s} \]

Due to the possibility of long-time trend, D-order season difference should also be taken into consideration, and the calculation formula is as below:

\[
V_s^D x_t = (I-B)^D y_t = V_s^{-1} x_t - V_s^{-1} x_{t-s} 
\]

If \( V_s^D x_t = (I-B)^D y_t = V_s^{-1} x_t - V_s^{-1} x_{t-s} \) is steady, the model satisfying the formula is called ARIMA model:

\[
\phi(B)^D V_s^D x_t = \theta(B)^D e_t
\]

Where:

\[
\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \\
\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q
\]

\( \{z_t, t \in T\} \) is ARIMA \((p,d,q)\) series, which satisfies:

\[
\phi(B) V^d z_t = \theta(B) e_t
\]

\( \phi(B) \) and \( \theta(B) \) meet the autoregression and moving average model conditions, and \( e_t \) is zero-mean white noise series. Formula 6 is substituted into formula 5 to obtain general product seasonal model as follow:

\[
\phi_p(B) \phi_q(B)^D V_s^D x_t = \theta_p(B) \theta_q(B)^D e_t
\]

Here subscripts \( p,q \) are used. By labeling the order of each operator, the seasonal model is simplified into \( ARIMA(p,d,q) \times (P,D,Q) \) model.

3.2. Theoretical approach of time series model

1. Characteristic function of time series model

The autoregression function and partial correlation function were used to describe the correlation between sequential values in the ARIMA model.

(1) Autoregression function

The autoregression function is a measurement of interaction degree of the same event in different periods. The calculation formula for the autoregression coefficient with k-order delay is as below:

\[
\hat{\rho}_k = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(x_{i+k} - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}
\]

Where \( \overline{x} \) —— sample mean.
Assume that $\hat{\rho}_k$ can never be zero as the time passes by, this is called trailing, and on the contrary, it is called truncation when it is constantly zero after one $k$ value is reached. Because of sample randomness and error, $\hat{\rho}_k$ will not completely realize the truncation with constant zero value under theoretical circumstance, but this can be judged by testing whether it is significantly zero according to hypothesis.

②Partial correlation function

When the $k$-order delay autocorrelation coefficient is solved, the partial correlation function is introduced merely for testing the correlation between $x_i$ and $x_{i-k}$. Its principle is a $k$-order delay autocorrelation function which eliminates the error disturbance due to the middle $k-1$ random variables. The $k$-order delay partial correlation is as below:

$$D_k = \begin{bmatrix} 1 & \hat{\rho}_1 & \ldots & \hat{\rho}_k \\ \hat{\rho}_1 & 1 & \ldots & \hat{\rho}_k \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{k-2} & \hat{\rho}_{k-1} & \ldots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \hat{\rho}_1 \\ \hat{\rho}_2 \\ \vdots \\ \hat{\rho}_k \end{bmatrix}$$

The truncation/trailing characteristic of $\theta_{kk}$ can be defined in a similar way to $\hat{\rho}_k$.

2. Basic principle of model order determination

If a time series is steady without white noise after preprocessing, it can be used to construct a correlation model. To construct a time series model, model recognition should be conducted first, and then a proper model will be chosen according to the autocorrelation coefficient and partial correlation coefficient of the sampled data. Tab. 2 presents the basic order determination principle of ARMA(p,q) model.

| $\hat{\rho}_k$ | $\theta_{kk}$ | Model       |
|----------------|---------------|-------------|
| Trailing       | Truncation of order p | AR(p)Model |
| Truncation of order q | Trailing       | MA(q)Model  |
| Trailing       | Trailing       | ARMA(p,q)Model |

In practical data processing, the samples can be of very strong randomness, so their autocorrelation and partial correlation coefficients will not be identical with theoretical truncation, but instead, in reality, small-amplitude oscillation of zero value often occurs at truncation part. In this case, the oscillation value can serve as the basis for truncation processing, and how to determine the corresponding order remains to be investigated. There is no absolute standard, and the processing relies upon the researcher’s experience. Jenkins, Watts and Quenouille obtained a conclusion through related researches, namely, the range of doubled standard deviation can be used to assist in the judgment.

The significant level is taken as $\alpha = 0.05$. When the two correlation coefficients of test data remarkably exceed the doubled standard deviation within 0-k orders, over 95% of the correlation coefficients thereafter are located within the specified range, and the coefficient fluctuation is turned into small-value fluctuation, it can be usually considered as k-order truncation; If over 5% of the correlation coefficients go beyond this range, or the mutation process is slowed down, it is generally regarded as trailing.

3. Estimation method of model parameters
Different estimation methods of model parameters cause errors to different degrees, which may directly lead to the degradation of model accuracy. The commonly used estimation methods include: ① moment estimation method ② least square method ③ maximum likelihood method.

4. Model test and optimization
   ① Stationarity and reversibility tests
   After the parameters are estimated, the reasonability of the obtained series model should be judged through tests, generally stationarity and reversibility tests.

   \[ \phi(B)x_t = \theta(B)e_t \]  \hspace{1cm} (11)

   The necessary and sufficient condition for stationarity and reversibility of the ARMA model is that the roots of both \( \phi(B) = 0 \) and \( \theta(B) = 0 \) are outside the unit circle.

   ② Applicability test
   The applicability test of a model generally consists of significance test of the model and that of model parameters.

   If the sample data information is completely extracted from series values, this is a white noise time series. If not, the fitted model is inapplicable, and it is necessary to select other models for sake of practicality test. Hence, the model significance test is namely white noise test of the sample difference time series.

   The significance test of parameters means testing whether an unknown parameter is significantly non-zero, and it aims to simplify the model to the greatest extent. When the test statistic satisfies \( Q_{\alpha} \) or \( Q_{\alpha - 1} \), the original hypothesis will be rejected, and this parameter is considered significant; or otherwise this parameter is insignificant, and then the model should be re-fitted after excluding the independent variables corresponding to insignificant parameters, so as to construct a more refined fitted model.

   ③ Model optimization
   The quality of a fitted model can be verified from two aspects: likelihood function value, as a general rule, the greater the likelihood function value, the better the fitting effect; number of unknown parameters in the model, naturally without significant precision change, the smaller the number of unknown parameters, the better. The common criteria include AIC, BIC and SBS, all of which can be called minimum information test.

   According to AIC criterion, the model which makes weighting function of fitting precision and parameter number reach the minimum value is considered relatively optimal, and AIG function is as below:

   \[ AIC = -2ln + 2n \]  \hspace{1cm} (12)

   Where: \(-2ln\) — maximum likelihood function value of the model 
   \(2n\) — number of unknown parameters in the model

   However, when the sample size tends to be infinitely great, the model selected in accordance with AIC criterion will not be converged to real model. To compensate for the shortcoming of AIC criterion, BIC and SBC criteria are proposed successively, where SBC criterion is obtained according to the Bayes theory and defined as follow:

   \[ SBC = -2ln + ln(n) \]  \hspace{1cm} (13)

   Where: \(ln(n)\) — number of unknown parameters in the model

   As it is impossible to compare AIC and SBC function values of all models, SIC and SBC function values of only a finite number of models can be investigated within a range as comprehensive as possible. It is deemed that the model making AIC and SBC function reach minimum values is relatively optimal and can be the final fitted model.
4. GA-ARIMA MODELING

4.1. Modeling

The precision of the traditional time series analysis method depends on the precision of model parameters to a great extent. Genetic algorithm, which is used to simulate biological evolutionary process, has been extensively applied to numerous fields in recent years as an effective means of solving complicated optimization problem. As a global optimization algorithm, genetic algorithm does not rely upon gradient information in the calculation process or require that the objective function should be derivable, and moreover, it has no concrete requirements for search space. Nevertheless, it has certain defects in practical application, which are mainly manifested by algorithm premature, poor local optimization ability, low convergence rate, etc. In view of complexity of time series model solving and superior characteristics of genetic algorithm, this research attempted to construct an improved time series model based on genetic algorithm, namely GA-ARIMA model. Genetic algorithm was applied to parameter estimation of the time series model, and the optimal time series model was finally obtained through global search in the solution space of parameters according to valid fitness function. Under given error indexes, the proposed model was compared with the traditional time series analysis method in the aspect of precision. The basic flow of GA-ARIMA model is as follows:

Step 1: Acquire time series;
Step 2: Preprocess time series, and do difference operation for unsteady data to obtain steady data;
Step 3: Model recognition, namely select a proper known model which accords with the actual time series process;
Step 4: Model order determination, namely apply BIC criterion to model order determination;
Step 5: Estimation of model parameters;
Step 6: Test the model fitting effect;
Step 7: If the accuracy and robustness requirements are not satisfied, the initial estimated parameters will be input into genetic algorithm;
Step 8: Encode space solution and generate an initial population;
Step 9: Calculate individual fitness;
Step 10: Select chromosomes with high fitness for duplication, crossover and mutation, and generate new population;
Step 11: If the new population does not meet termination condition, return to and repeat Step 9 until the condition is satisfied;
Step 12: Decode, substitute the trained parameter estimates into Step 6 to test the model fitting effect;
Step 13: If the accuracy and robustness requirements are not satisfied, repeat steps following Step 6 until the accuracy requirement is satisfied;
Step 14: Obtain the optimal model result.

4.2. Model evaluation indexes

The time series can be evaluated through various error evaluation methods. The three following evaluation indexes are selected:

(1) Robustness index

\[ R_i = \max \left[ \frac{|M_i - F_i|}{M_i} \right] \times 100\% \]  

(14)

Where,  

- \( M_i \) —the i (th) observation data;
- \( F_i \) —the i (th) predicted data;
- \( n \) —sample size.

(2) Mean absolute percentage error

\[ \text{MAPE} = \frac{\sum_{i=1}^{n} |\frac{M_i - F_i}{M_i}|}{n} \times 100\% \]  

(15)

Where MAPE—mean absolute percentage error.
(3) Root mean square

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (M_i - F_i)^2} \times 100\%
\]  

Where: RMSE—root mean square error.

5. Instance Analysis

5.1. Data needed in instance analysis

“Jinggangshan Road Fuchunjiang Road Bus Stop” in No. 26 Bus Line in Huangdao District Development Zone, Qingdao, was selected to carry out a prediction research on arrival time at the bus stop. The research lasted 21 days from August 1 to August 21, the research data of 14 hours (06:00-20:00) every day were selected, the sample size was 196, and some instance data are seen in Tab. 3.

| BUSRDID  | ROUTEID | PRODUCTID | STATIONSEQNUM | UTC         | STOPPINGTIME |
|----------|---------|-----------|----------------|-------------|--------------|
| 1856162796 | 6026    | 10399875  | 7              | 1502599822  | 38           |
| 1856683630 | 6026    | 10399875  | 7              | 1502609010  | 50           |
| 1857362329 | 6026    | 10399875  | 7              | 1502618983  | 50           |
| 1855464972 | 6026    | 10399886  | 7              | 1502587898  | 38           |
| 1857382706 | 6026    | 10399886  | 7              | 1502619289  | 40           |
| 1857974313 | 6026    | 10399886  | 7              | 1502632685  | 39           |

5.2. Predictive calculation of arrival time at bus stop based on traditional ARIMA model

The observation data values of arrival time at Jinggangshan Road Fuchunjiang Road Bus Stop were taken as samples to draw a time series chart as seen in Fig. 3.

![Fig. 3 Observation data of arrival time at bus stop](image)

It can be known from the observation data of bus arrival time shown in Fig. 3 that no evident linear trend is displayed in the series with a certain fixed cycle but not white noise. Therefore, linear information should be obtained through first-order difference operation, followed by 14-step periodic difference operation to explore whether seasonal fluctuation information exists. After the difference operation, the stationarity of the series is judged through the autocorrelation function and partial correlation function.

![Fig. 4 Post-difference partial correlation coefficient](image)
The post-difference partial correlation coefficient is shown in Fig. 5.2, indicating that the post-difference series contains very marked seasonal effect. Fig. 5, which shows the post-difference autocorrelation coefficient, manifests a certain short-time correlation in the post-difference series.

Second, the optimal model is selected using AIC or SBC criterion. As we cannot acquire AIC and SBC values of all models, a comprehensive consideration should be taken under limited conditions, and the model making AIC and SBC values reach the minimum is taken as the final fitted model. Following a comparison, it can be known that the two function values reach the minimum under p=(1) and q=(2)(14), so it is taken as the fitted model. In other words, the post-difference series complies with formula 17:

\[ (1 - \phi_1 B)x_{t-14} = [1 - \theta_1 B^2] [1 - \theta_{14} B^{14}] e_t \]  

(17)

The parameter estimation is conducted with the maximum likelihood method. The estimated values of parameters \( \phi_1 \), \( \theta_1 \), \( \theta_{14} \) and \( \sigma_e^2 \) are obtained through the approximate calculation:

\[ \hat{\phi}_1 = -0.53486, \quad \hat{\theta}_1 = 0.53157, \quad \hat{\theta}_{14} = 0.63753, \quad \sigma_e^2 = 6.42 \times 10^6 \]

According to the parameter estimates, the time series model of travel time at the bus stop is obtained:

\[ (1 + 0.53486B) [1 - B^{14}] x_t = [1 - 0.53157B^2] [1 - 0.63753B^{14}] e_t \]  

(18)

The fitted model is verified at significance level of \( \alpha = 0.05 \), the test results of the fitted model are seen in Tab. 4, and in the residual white noise test, p value is always greater than 0.05, so \( e_t \) can be considered as white noise, namely the residual series does not contain unextracted information. P value is always smaller than 0.05 in the parameter significance test, in other words, all parameters are significantly non-zero, and the model holds.

| evaluating indicator | ARIMA Model | GA-ARIMA Model |
|----------------------|-------------|----------------|
| Robustness           | 19.71%      | 17.44%         |
| MAPE                 | 14.37%      | 12.33%         |
| RMSE                 | 12.93%      | 10.67%         |

The stationarity and reversibility tests are carried out for the fitted model. Both roots are outside the unit circle when \( \phi(B) = 0 \) and \( \theta(B) = 0 \), proving that the model is steady and reversible. The arrival time at the bus stop is predicted using the fitted time series model and the fitting precision is calculated for the convenience of a comparative analysis of different models.

5.3. Predictive calculation of arrival time at bus stop based on GA-ARIMA model

Based on the model recognition based on the traditional time series model as described in section 18, the following can be obtained by arranging formula 18:

\[ x_t - (1 + \phi_1)x_{t-1} + \phi_1 x_{t-2} - \theta_1 x_{t-14} + (1 + \phi_1)x_{t-15} - \phi_1 x_{t-16} = \]

\[ e_t - \theta_1 e_{t-1} - \theta_{14} e_{t-14} + \theta_1 \theta_{14} e_{t-16} \]  

(19)

The objective function Q is taken as the square sum of absolute errors of predicted and observed arrival time data every day:
\[ Q = \sum_{t=1}^{100} (x_t - \hat{x}_t)^2 \]  

(20)

Where daily residual error can be expressed through the recursive method according to parameters as well as the observed values and residual errors of historical arrival time, and it is a highly nonlinear function of parameters \( \phi_1, \ldots, \phi_p \) and \( \theta_1, \ldots, \theta_q \).

First, the four parameters of genetic algorithm are set, that is, initial number of individuals: 50, crossover probability: 0.8, mutation probability: 0.01 and number of iterations under which the evolution is terminated: 500.

The parameter estimation is implemented via genetic algorithm and the fitted model is tested. The residual white noise test and parameter significance test demonstrate that this model is effective, and a fitted model is formed as below:

\[
(1 + 0.4056 B)(1 - B^4)(1 - B)\epsilon_t = (1 - 0.4568 B^2)(1 - 0.4135 B^4)\epsilon_t
\]

(21)

Where \( \sigma^2 = 6.78 \times 10^8 \).

In the end, the model passes the stationarity and reversibility tests, thus indicating the model applicability.

5.4. Analytical result comparison of two prediction models

19 groups of actual data are taken as observation data for prediction, and then analytically compared with two groups of measured data in the aspect of fitting precision.

| Residual white noise test | Parameter significance test |
|---------------------------|-----------------------------|
| Delay order | \( \chi^2 \) statistic | P value | Parameter estimates | \( T \) statistic | P value |
| 4 | 5.91 | 0.2157 | \( \phi_1 \) | 5.73 | <0.001 |
| 8 | 10.57 | 0.3473 | \( \phi_4 \) | 4.62 | <0.001 |
| 12 | 20.13 | 0.2694 | \( \phi_1 \) | -6.11 | <0.001 |

The fitting precisions are displayed in Tab.5. Faced with irregular fluctuation existing in the bus arrival time at the bus stop, both prediction models can still contribute to high prediction precision and complete prediction of bus arrival time very well. Based on a comparison of the precisions of the time series models—ARIMA model and GA-ARIMA model—established through the two methods, it can be seen from their robustness and error indexes that the time series model has improved the fitting precision after its parameters are estimated using genetic algorithm. This further proves that the precision of model prediction result can be improved through parameter optimization, and this can be applied in practice.

6. CONCLUSION

Precision prediction of bus arrival time at bus stops helps passengers to reasonably plan their travel time and improves the attractiveness of bus travel. The genetic algorithm was employed in this research to improve and optimize the time series model, and an instance analysis was carried out based on actual bus data. The data results manifest that in comparison with the traditional model, the improved time series model reduces the prediction error to a certain degree. Therefore, the improved GA-ARIMA model, which is of certain practicability, can further improve the accuracy of bus arrival time mastered by passengers.

REFERENCES

[1] Marko Čelan, Marjan Lep. Bus-arrival time prediction using bus network data model and time periods[J]. Future Generation Computer Systems, 2020, 110.
[2] Tianli Tang, Ronghui Liu, Charisma Choudhury. Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data [J]. Sustainable Cities and Society, 2020, 53.

[3] Mansur As, Tsunenori Mine, Tsubasa Yamaguchi. Prediction of Bus Travel Time Over Unstable Intervals between Two Adjacent Bus Stops [J]. International Journal of Intelligent Transportation Systems Research, 2020, 18(3).

[4] Deng Mei, Li Chunda. Dynamic prediction model of bus stop time under different platform conditions [J]. Journal of Chongqing Jiaotong University (NATURAL SCIENCE EDITION), 2019, 38 (12): 105-109.

[5] Huang Yishao, Han Lei. Short term passenger flow prediction method for bus stops based on improved extreme learning machine [J]. Transportation system engineering and information, 2019, 19 (04): 115-123.

[6] Li Mei. Short term bus passenger flow prediction based on deep learning [D]. Beijing Jiaotong University, 2019.

[7] Wang Cheng. Research on short-term bus passenger flow prediction model based on big data [D]. Dalian Maritime University, 2019.

[8] LV Panlong. Mining and analysis of bus commuter travel characteristics based on multi-source data [D]. South China University of technology, 2019.

[9] Shan Yulong. Application of ARIMA model and genetic algorithm optimized neural network in GDP prediction [D]. Lanzhou University, 2014.