Passenger Flow Forecasting Research for Airport Terminal Based on SARIMA Time Series Model

Ziyu Li1,a, Jun Bi1,b and Zhiyin Li1,c
1 MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China.
a15120764@bjtu.edu.cn, b bilinghc@163.com, cZhangjun@163.com

Abstract. Based on the data of practical operating of Kunming Changshui International Airport during 2016, this paper proposes Seasonal Autoregressive Integrated Moving Average (SARIMA) model to predict the passenger flow. This article not only considers the non-stationary and autocorrelation of the sequence, but also considers the daily periodicity of the sequence. The prediction results can accurately describe the change trend of airport passenger flow and provide scientific decision support for the optimal allocation of airport resources and optimization of departure process. The result shows that this model is applicable to the short-term prediction of airport terminal departure passenger traffic and the average error ranges from 1% to 3%. The difference between the predicted and the true values of passenger traffic flow is quite small, which indicates that the model has fairly good passenger traffic flow prediction ability.

1. Introduction
The airport is currently carrying scale increased year by year, the traditional method of airport resource allocation has been unable to adapt to the requirements of the operation of the airport. Dynamic allocation and scheduling of airport terminal passenger service resources is one of the effective ways to improve passenger service levels and operational efficiency within the terminal, while the relatively accurate passenger traffic forecasting is the prerequisite for dynamic allocation and scheduling.

In the field of civil aviation, the growth of traffic pressure will inevitably lead to the phenomenon of terminal congestion. Therefore, to achieve high efficiency terminal operation, the airport put forward higher requirements on the arrival terminal passenger traffic forecasts. The domestic research about prediction of travelers arrive in a short time is still in its infancy, because of the passenger traffic in a short time volatility. Wonkyu Kim [1] adopts the prediction method based on probability density function, according to the travel time of passengers arriving at the airport terminal, the number of passengers arriving at the terminal in different periods is estimated. Wang Long [2] used the time series method to study the passenger flow deviation rate under different statistical time series in year and month. Guo Yuanyuan [3] obtains data by means of field investigation in the terminal, and uses chaotic time series theory and RBF neural network to predict the passenger flow. Tian Yuan [4] uses data mining method to build a passenger flow combination forecasting model based on clustering algorithm, decision tree model and K-nearest neighbor algorithm.

At present, most of the passenger flow forecasting algorithms choose the time granularity of the month, quarter or year, and the granularity of time is coarse, so it cannot observe the change of
passenger flow in a short period of one day. In addition, most passenger flow forecasting studies do not take seasonal effects into account, which results in greater prediction error.

2. SARIMA model

The seasonal time series model, also known as SARIMA model, is derived from the autoregressive single moving average model (ARIMA). The model can employ Box-Jenkins model identification, estimation and prediction procedures. With the increase of historical data, the model can be adjusted in real time, which can not only improve the prediction accuracy, but also be applied to practical engineering. SARIMA model is a short-term prediction model. The key point is to process the data, at the same time, the error generated by the value fit is taken as an analysis factor. The prominent advantage of the model is that the short-term prediction results are more accurate.

Before constructing the ARIMA model, the stationary test must be carried out first. If the time series is stable, the ARMA model can be established directly; if the time series is non-stationary, the difference sequence can be used to smooth the sequence. A d order integral time series ARIMA model can be established as follows:

\[ x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q} + \sum_{j=1}^{7} \theta_j \varepsilon_{t-j} \]

Where: \( p \) is the autoregressive order; \( q \) is the moving average order; \( d \) is the difference number; \( \phi \) and \( \theta \) denote model parameters.

SARIMA model is a combination of stochastic seasonal model and ARIMA model. If the considered time series has periodic features, then the SARIMA model can be constructed. The general form of the SARIMA model is represented as:

\[ \varphi_p(L)A_p(L^s)(\Delta^d \Delta_s y_t) = \theta_q(L)B_Q(L^s) \varepsilon_t \]

Where: \( s \) represents the seasonal cycle of change in the sequence; \( L \) stands for hysteresis operators; \( \varphi_p(L) \) and \( A_p(L^s) \) represent non-seasonal and seasonal autoregressive polynomials respectively; \( \theta_q(L) \) and \( B_Q(L^s) \) represent non-seasonal and seasonal moving average polynomials respectively; \( P, Q, p, q \) represent the maximum lag order of the seasonal and non-seasonal autoregressive and moving average operators respectively; \( d, D \) indicate non-seasonal and seasonal difference times respectively. In practice, if the original sequence contains both the trend and seasonal, can be expressed as SARIMA\((p, d, q)(P, D, Q)\) model.

3. Case analysis

This paper need to analyze the process of arrival of passenger terminal flow at T2 terminal of Kunming Changshui International Airport, mine temporal and spatial distribution characteristics of passenger flow, and predict the passenger flow in the future for a period of time by using the SARIMA model.

The empirical study takes the airport terminal passenger flow data from each 10 min of per day between August 1, 2016 and August 25, 2016 as the historical database, and uses the passenger flow data from the airport terminal each 10 min of per day between August 26, 2016 and August 31, 2016 as the validation data for the prediction effect. The empirical study uses Eviews to build SARIMA model. The prediction process is divided into dynamic prediction and static prediction. The prediction process we choose is dynamic prediction which is based on historical data and real data updated in real time.

3.1. Original sequence analysis

Make daily statistics on arrival passenger flow, as shown in Figure 1. As can be seen from the Figure 1, the airport terminal passenger flow is relatively stable, essentially fluctuates around a value, and there is no obvious trend of increasing or decreasing. The sequences exhibit seasonal fluctuations with a 24 hour cycle. The daily 6:00-8:00 is the peak time of airport terminal passenger flow, and the peak passenger flow can reach 800 to 1200 passengers every 10 minutes. Passenger traffic tends to be stable between 8:00 am and 10:00 pm, the passenger traffic of each 10 minutes is range from 600 person
time to 800 person time. The passenger traffic is very low between 10 p.m. and 5 a.m., because the airport will be closed at night. Through the above analysis, we can see that the seasonal time series model is one of the ideal models to describe the original sequence.

The original sequence is tested by stationary, and Table 1 is the ADF test result of the original sequence. From the result, we can see that the p value of the t statistic is 0.7069 (greater than 0.05), which indicates that the time series of passenger flow has unit root, which is a non-stationary process.

![Figure 1. Airport Terminal Passenger Flow Sequence Diagram](image1)

Table 1. ADF Test Results of Raw Data

| Inspection method                      | t-Statistic | Prob.* |
|----------------------------------------|-------------|--------|
| Augmented Dickey-Fuller test statistic | -1.127974   | 0.7069 |

3.2. Model building

3.2.1. Stabilization of time series. The paragraph text follows on from the subsubsection heading but should not be in italic. According to the above analysis, seasonal differential processing of the original sequence is required. The original passenger flow data are counted by 10min at intervals and it takes one day (that is, 144 periods) as a cycle, so Therefore, the parameter s=144 of the ARIMA model can be determined. Figure 2 is a time series diagram after a seasonal differential operation. It can be seen that the data basically fluctuates around 0, and the fluctuation range is basically symmetrical. The unit root test result of the sequence after seasonal difference is shown in Table 2. From the result, we can see that the p value of the t statistic is 0.7069(greater than 0.05), which indicates that the time series of passenger flow has unit root, which is a nonstationary process. The results show that the p value of the t statistic is 0.0000 (less than 0.05), indicating that the new sequence does not have a unit root, which is a stationary process.

![Figure 2. Time Series Diagram After the Seasonal Differential Operation](image2)

Table 2. ADF Test Results of the Sequence After Seasonal Difference

| Inspection method                      | t-Statistic | Prob.* |
|----------------------------------------|-------------|--------|
| Augmented Dickey-Fuller test statistic | -40.85855   | 0.0000 |

Figure 3 is auto correlation graphs and partial correlation graphs of the sequence after seasonal difference. As you can see from the diagram, the autocorrelation coefficient of the sequence decays rapidly to 0 after a very short period of delay. Combining the results of the ADF test and the
autocorrelation map, it can be considered that the sequence after the seasonal difference is a stationary sequence.

![Image](image.jpg)

**Figure 3. Diagram of the Autocorrelation Function and Partial Autocorrelation Function of Stable Sequence**

### 3.2.2. Model order determination and parameter estimation.

The SARIMA model is established by using the processed time series, and the model order can be determined by the diagram of the autocorrelation function and the partial autocorrelation function. It can be seen from Figure 3: PAC is truncated in the 6 order, so the parameter $p$ of the autoregressive process AR is 6; The AC function assumes a fast decay property, and the parameter $q$ of the moving average model MA is temporarily determined to be 5. Thus, the SARIMA model of time series is basically determined. According to the basic principle of model selection, the model of passenger flow forecasting can be preliminarily determined as $SARIMA(6,0,5)(1,1,0)^{144}$.

The parameter estimation methods of the model include moment estimation, maximum likelihood estimation and least squares estimation. This project uses the most widely used least squares estimation method to estimate parameters. Model estimation values are shown in Table 3.

| Parameter   | Estimation value | standard deviation |
|-------------|------------------|--------------------|
| AR(1)       | 0.539902         | 0.016316           |
| AR(2)       | 0.941326         | 0.016307           |
| AR(3)       | 0.222245         | 0.019289           |
| AR(4)       | -0.738893        | 0.021187           |
| AR(5)       | -0.531023        | 0.017095           |
| AR(6)       | 0.473437         | 0.018990           |
| SAR(144)    | -0.165346        | 0.023826           |
| MA(1)       | 0.457232         | 0.004859           |
| MA(2)       | -0.787077        | 0.006708           |
| MA(3)       | -0.823167        | 0.004540           |
| MA(4)       | 0.390056         | 0.006645           |
| MA(5)       | 0.966380         | 0.004818           |

### 3.2.3. Model test.

Before using the model to predict, we need model test. The model test is divided into two parts: parameter significance test and model significance test.

(1) Parameter significance test
For the passenger flow forecasting model, the parameter significance test is carried out, and the results are shown in Table 4. As shown in Table 4, the P values of the regression coefficients of AR, SAR and MA are less than 0.05, and the model has basically passed the significance test, which can be considered to be in line with the requirements.

| Parameter | t-Statistic | Prob.* |
|-----------|-------------|--------|
| AR(1)     | 33.08980    | 0.0000 |
| AR(2)     | 57.72571    | 0.0000 |
| AR(3)     | 11.52199    | 0.0000 |
| AR(4)     | -34.87455   | 0.0000 |
| AR(5)     | -31.06287   | 0.0000 |
| AR(6)     | 24.93038    | 0.0000 |
| SAR(144)  | -6.939801   | 0.0000 |
| MA(1)     | 94.10262    | 0.0000 |
| MA(2)     | -117.3341   | 0.0000 |
| MA(3)     | -181.3000   | 0.0000 |
| MA(4)     | 58.69516    | 0.0000 |
| MA(5)     | 200.5837    | 0.0000 |

(2) The model test of significance
Model significance test is used to test the validity of the model. A good model can extract the relevant information of sequence values. That is to say, the residual sequence should be a sequence of white noise. Therefore, the significance test of the model can be regarded as the test of its residual sequence. The residual sequence of the model SARIMA(6,0,5)(1,1,0)_{144} is tested by white noise, as shown in Table 5. As can be seen from the table, the P value of the LB statistic is greater than the significance level which is 0.05, and the residual sequence belongs to the white noise sequence. Therefore, it can be considered that the model fits the data better.

| Delay order | Q-Statistic | Prob.* |
|-------------|-------------|--------|
| 6           | 112.77      | 0.226  |
| 12          | 512.54      | 0.356  |
| 18          | 645.53      | 0.434  |

3.3. Forecast result and analysis
Based on the model of seasonal time series fitted above, we forecasted the airport terminal passenger traffic at different time periods (time interval: 10min) from August 26, 2016 to August 31, 2016. The prediction contrast chart is shown in Figure 4. According to formula (1), we calculated the relative error of prediction in different time periods, and plotted the relative error map of different time periods. And the prediction errors of airport terminal passenger flow are statistically analyzed, as shown in Table 6.

\[
\text{error} = \frac{\hat{y}_i - y_i}{y_i}
\]

Where: \(y_i\) is the true value; \(\hat{y}_i\) is the predicted value. Under normal circumstances, the airport from 10 in the evening to 5 the next morning in the closed state, the airport terminal passenger flow is basically 0, so the statistical relative error is of little significance. In this empirical analysis, when analyzing the relative error of the forecast results, the relative error value of the forecast results in different time periods between 5 a.m. and 10 p.m. is analyzed.
Figure 4. (a)-(f) is the contrast diagram respectively of predicted results at different time intervals (time intervals: 10min) from August 26, 2016 to August 31, 2016.

Figure 5. (a)-(f) is the relative error diagram respectively at different time intervals (time intervals: 10min) from August 26, 2016 to August 31, 2016.

Table 6. Error analysis

| Date     | Mean of relative error |
|----------|------------------------|
| 2016/8/26| 1.24%                  |
| 2016/8/27| 1.68%                  |
| 2016/8/28| 1.07%                  |
| 2016/8/29| 2.02%                  |
| 2016/8/30| 2.49%                  |
| 2016/8/31| 1.74%                  |

As can be seen from the above forecast results and the error analysis, the change trend of the real value is basically consistent with the prediction curve. The average daily forecast relative error of airport terminal is between 1% and 3%. The difference between the predicted and the true values of passenger traffic flow is quite small, which indicates that the model has fairly good passenger traffic flow prediction ability.

4. Conclusion
The passenger flow of the airport terminal presents a periodic fluctuation, and the SARIMA model can better describe the fluctuation of passenger flow at the airport terminal. In this paper, SARIMA model is used to predict short-term passenger flow, which can get more accurate prediction value. The prediction results can lay the foundation for the dynamic optimization of airport departure resources. In this paper, the impact of flight schedule on passenger flow is not considered when forecasting passenger flow. In the future forecast research, flight information is taken as the influencing factor to improve the prediction accuracy.

References
[1] Kim W, Park Y, Jong Kim B. Estimating hourly variations in passenger volume at airports using dwelling time distributions[J]. Journal of Air Transportation Research Record 1423, 34-39.
[2] Wang Long, Wang Min. Intervention Research on Deviation Rate of Civil Aviation Traveler Flow in Time Order: Taking Airport in Shenzhen and Kunming as Examples[J]. Journal of Chengdu University of Technology(Social Sciences), 2008, 16(3): 74-77.
[3] Guo Yuanyuan. Terminal departures passenger traffic prediction based on chaotic time series[D]. Harbin Institute of Technology, 2013.
[4] Tian Yuan. Passenger Flow Forecasting Research For Airport Terminal Based on Data Mining[D]. Harbin Institute of Technology, 2014.
[5] Boxgep, Jenkinsgm, Reinselgc. Time series analysis ——Forecasting and control[M]. Beijing: People's Posts and Telecommunications Press, 2009.

[6] Van D V, Dougherty M, Watson S. Combining kohonen maps with ARIMA time series models to forecast traffic flow[J]. Transportation Research Part C, 1996, 4(5):307-318.