ARIMA and SVM Combination Forecast for Holiday Subway Passenger Traffic

Xialing Cao  
University of Chang’an, School of Highway  
Xi’an, Shaanxi, China  
exling616@163.com

Chaoqun Ma  
University of Chang’an, School of Highway  
Xi’an, Shaanxi, China  
machaoqun314@163.com

Yunpu Jia  
University of Chang’an, School of Highway  
Xi’an, Shaanxi, China  
1547650711@qq.com

ABSTRACT
To explore the distribution regularity of urban passenger traffic during the holiday, this article selected the National Day holiday passenger flow of the subway station on Lines 3 in Xi’an as the basis for data analysis, and adopted the ARIMA (Autoregressive Integrated Moving Average) model, the SVM (support vector machine) model and the mixed model of the two methods to predict the holiday hourly passenger flow of the subway. The research results show that the relative error between the prediction results of the ARIMA & SVM combined model and the actual results is much smaller. So, the combined model (ARIMA & SVM) has strong practicability and generalization and is suitable for the forecast of subway passenger flow, which can provide a foundation of quantification for reasonably carrying out the station passenger transport organization, management, etc during the holiday.

KEYWORDS: urban rail transit, holiday passenger flow, SVM, ARIMA, combined model prediction

1 INTRODUCTION
With the improvement of residents' living standards, the number of leisure-oriented and traveling-oriented holiday trips have been increasing, which has caused tremendous pressure on scenic spots and commercial centers. At the same time, in recent years, urban rail transit network has developed rapidly, and a increasing number of people travel or commute by subway, which result in congestion of subway stations and subway lines. Due to the complexity purposes of travel, the characteristics of passenger flow during the holidays are quite different from working days in terms of passenger flow temporal distribution and spatial distribution, which brings serious challenges to the operation and management of the rail transit system.

The passenger flow forecasting of urban rail transit is the basis of urban rail transit line planning and operation organization (Chaoqun Ma, Yuping Wang et al., 2010). According to the time between forecasting period and now, it can be divided into short-term forecast,
medium-term forecast and long-term forecast. The short-term passenger flow forecasting serves for the operating organization, and the main purpose of short-term forecast is to respond to the fluctuations of short-term passenger flow quickly, formulate the corresponding train capacity allocation plans, and assume effective passenger transport organization measures to ensure the balance of resource allocation and the safety of the operation process.

The commonly method for predicting passenger flow inbound and outbound of the station is divided into linear prediction method and non-linear prediction method. Linear prediction methods include time series model (VOOTM V D, WATSON S. et al., 1996; WILLIAMS B M, HOEL L A. 2003), Kalman filtering model (OKUTANII, STEPHANIDES Y J. 1984; Zhiyong Zhang, Dandan Zhang et al., 2017). Non-linear prediction methods include nonparametric regression (DAVIS G A, NIHAN N L, 1991; Qiao Xie, Binbin Li et al., 2017; Weijia Chen, 2017), neural network model (XIAO J M, WANG X H, 2004; Chonglin Ren, Chengyi Cao, 2011; Shengwei Dong, 2013; Wei Zhou et al., 2014), support vector machine (Jun Yang, Zhongsheng Hou, 2013; Liqin Zhao, 2015). This paper adopts the Autoregressive Integrated Moving Average Model (ARIMA), support vector machine model (SVM) and ARIMA-SVM combined model to predict the inbound and outbound of the subway passenger flow during the holiday. The paper tries to compare the accuracy of prediction of the three models, and finds the prediction method which is closer to the actual situation to achieve a higher prediction accuracy in practical use, and to provide a foundation of quantification for reasonably carrying out the station passenger transport organization, management, etc during the holiday.

2 ANALYSIS OF PASSENGER FLOW CHARACTERISTICS OF ENTER AND EXIT STATION PASSENGERS DURING HOLIDAYS

By analysing the passenger flow data of the stations during the National Day holiday in Xi’an in 2017, it can be see that the total passenger traffic of the metro from 1st to 8th, October was 14,9584 million. This paper selected the data of passengers enter and exit the Station from October 1st to October 8th in 26 Stations of Xi’an Metro Line 3 in 2017. The data interval is 1 hour. Considering that the National Day holiday in 2017 was from 1st to 8th, October. For a total of 8 days (4th, October for the Mid-Autumn Festival), in order to predict the data of passenger flow reasonably, the article adjusted this sample. The data of 1st, 2nd, 3th, October was taken as data of 1st, 2nd, 3th day. The average data of 4th, 5th, October was taken as 4th day. The data of 6th, 7th, 8th, October was taken as data of 5th, 6th, 7th day. Compared with the rigid travel during the working day, the travel time of passengers were flexible, and relatively dispersed during the holiday. In addition, passengers require better public transport service levels, therefore, indicators such as congestion have a greater impact on the choice of passenger travel time. The passenger flow of the rail transit Stations in the commercial tourist attractions area has the characteristics of long peak duration and high peak value during the holiday. During the holiday period, the ratio of the each hour passenger flow entering and exiting of each station to the everyday average hourly entering and exiting are stable. Take the Dayanta Station as an example. As shown in Figure 1~2, the daily average hourly entering and exiting passenger flow rate of the Dayanta Station in the 2017 National Day holiday is relatively stable.
Figure 1: Proportion of each day hourly entering passenger flow and every day average hourly entering passenger flow at Dayanta Station during the National Day holiday.

Figure 2: Proportion of each day hourly exiting passenger flow and every day average hourly exiting passenger flow at Dayanta Station during the National Day holiday.

3 FORECASTING MODEL

3.1 Time Series Forecasting Model

ARIMA (Autoregressive Integrated Moving Average) Model, also known as ARIMA (p,d,q), which is the most common model used for time series forecasting in statistic models. ARIMA model is simple in structure and requires only endogenous variables and no other exogenous variables, but ARIMA requires that the time series data be stable or stable through differentiation.

The ARIMA model has three parameters: p, d, q.

- p -- the lags of the time series data used in the forecasting model, also called AR/Auto-Regressive term
- d -- the data of time series needs to be differencing several steps which is stable, also called Integrated term
- q -- the lags of the forecasting error used in the forecasting model, also known as the MA/Moving Average term.

Assuming p, q, d were known, ARIMA would be expressed mathematically as:
\[ y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \]  \hspace{1cm} (1)

Where \( \gamma \) represents the coefficient of AR and \( \theta \) represents the coefficient of MA.

Main steps of ARIMA modelling:

a. Obtaining time series data of the observed system;

b. Drawing on data, whether the observation is a stationary time series; doing d-order difference operation on the non-stationary time series and then non-stationary time series will transfer into stationary time series;

c. After the second step, a stationary time series has been obtained. The autocorrelation coefficient (ACF) and the partial autocorrelation coefficient (PACF) are obtained for the stationary time series. The best hierarchical p and order q are obtained by analysing the autocorrelation graph and the partial autocorrelation graph;

d. From the d, p, and q obtained above, an ARIMA model would be founded. Then start to model the obtained model.

e. Make Forecasting.

3.2 Support Vector Machine Forecasting Model

SVM (Support Vector Machine) is a method of machine learning. Support vector regression is to use SVM to fit the curve and do regression analysis, i.e., support vector regression (SVR); SVR can pass the kernel function (generally have linear kernel functions, radial basis kernel functions, polynomial kernel functions). Linearly indivisible data into high-dimensional linear data to the basis kernel function and polynomial kernel function, avoiding complex calculations directly in high-dimensional space; balancing between data fitting accuracy and the complexity of the fitting function To minimize the risk of model structure, this can avoid overfitting the fit function.

There is a training set \( \{x_i, y_i\} \), where \( x_i \in \mathbb{R}^D \) (\( x_i \) contains D-class features), \( i = 1, 2, \ldots, n \) they are \( n \) D-dimensional vectors, which are linear regression problems for traffic volume prediction problems. The SVM objective function calculation formula is:

\[
\begin{align*}
\min_{w,b} & = \frac{1}{2} \|w\|^2 \\
\text{s.t.} & \quad y_i(w^T x_i + b) \geq 1, i = 1, 2, \ldots, m.
\end{align*}
\]

(2)  \hspace{1cm} (3)

In the formula, \( w = (w_1, w_2, \ldots, w_d) \) is a normal vector; \( b \) is the amount of difference, which determines the position of the hyperplane. The SVR uses a non-sensitive function \( \xi \), that is, if the error range is within the acceptable range, the objective function can be considered as no loss. If the error value is greater than \( \xi \), the loss is calculated minus the \( \xi \), so the SVR problem is converted into:

\[
\begin{align*}
\min_{w,b} & = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} l_{\xi}(f(x_i) - y_i)
\end{align*}
\]

(4)

Where \( C \) is a regularization parameter, \( l_{\xi} \) is a non-sensitive loss function, and \( l_{\xi} \) is calculated as follows:

\[
l_{\xi} = f(x) \left\{ \begin{array}{ll}
0, & \text{if } |z| \leq \xi \\
|z| - \xi, & \text{if } |z| > \xi
\end{array} \right. \]
\]

(5)
Solving (4) by Lagrangian multiplier method, the Lagrangian multiplier is \( a_i^*, \ a_i^* \geq 0 \) and \( \mu_i, \ \mu_i \geq 0 \), then construct the Lagrangian function and let the Lagrangian function

\[
\omega = \sum_{i=1}^{N} (a_i + a_i^*) x_i \tag{6}
\]

Perform a dual transformation on (6):

\[
\omega = \sum_{i=1}^{N} (a_i + a_i^*) x_i
\]

\[
\cdot \min \left\{ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (a_i - a_i^*)(a_j - a_j^*) \phi(x_i)^T \phi(x_j) + \xi \sum_{i=1}^{N} (a_i + a_i^*) - \sum_{i=1}^{N} (a_i - a_i^*) y_n \right\}
\]

s.t. \( \sum_{i=1}^{N} (a_i^* - a_i) = 0, \ a_i \geq 0, \ a_i^* \leq C \) \tag{7}

Because the Gaussian kernel function in the radial basis kernel function is simple and the performance is better, the Gaussian kernel function (8) is used instead of the inner product in equation (7):

\[
k(x, y) = \exp(-\frac{||x - y||^2}{2\sigma}) \tag{8}
\]

\[
\min \left\{ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (a_i - a_i^*)(a_j - a_j^*) k(x_i - x_j) + l_\xi \sum_{i=1}^{N} (a_i + a_i^*) - \sum_{i=1}^{N} (a_i - a_i^*) y_n \right\}
\]

s.t. \( \sum_{i=1}^{N} (a_i^* - a_i) = 0 \)

\[
a_i \geq 0, \ a_i^* \leq C \tag{9}
\]

According to the KKT optimization conditions, the equation (9) can be obtained:

\[
b = \begin{cases}
  y_i - l_\xi - \sum_{i=1}^{N} \sum_{j=1}^{N} (a_i - a_i^*)(a_j - a_j^*) \cdot k(x_i - x_j) \\
y_i + l_\xi - \sum_{i=1}^{N} \sum_{j=1}^{N} (a_i - a_i^*)(a_j - a_j^*) \cdot k(x_i - x_j)
\end{cases} \tag{10}
\]

Finally, it a nonlinear equation is obtained:

\[
f(x) = \sum_{i=1}^{l} a_i K(x_i, x_j) + b \tag{11}
\]

It can be found that the forecasting of SVM only needs to determine two parameters when while deriving the SVM process.
Main steps of SVM modelling:

a. Prepare the input and output historical data of the model.

b. Normalization processing to the input and output historical data of the model, that is, the original data is changed linearly to the interval \([0, 1]\). The practical forecasting value is obtained by reversal processing to the normalized forecasting result.

c. Select the kernel function of the support vector machine. Since the radial basis function has the widest application range and can directly reflect the distance between two data, this study takes the radial basis function as the kernel function.

d. Use ten-fold cross-validation method to select the regularization parameters and the kernel parameters.

e. The set of sample is trained using the optimal regularization parameters and kernel parameters, and the Lagrange multiplier, \(a_i\), and the offset, \(b\), are obtained to determine the forecasting function (11).

f. According to the prediction function generated in e., the passenger flow of the holiday is forecasted, and the forecasting error is evaluated and analysed; if the error is large, the process returns to the first step, and readjust the model input and the SVM parameters

Wherein, the initial value of the regularization parameter is set to 100; the optimization range is set to \([e^{-1}, e^{10}]\); the initial value of the kernel parameter is set to 0.1; and the optimization range is set to \([e^{-3}, e^{8}]\); the error function selected the mean square error between the estimated value and the true value. When the reduction of the mean square error of the adjacent two optimizations is less than \(10^{-3}\), the optimization is ended. The presupposition error is based on the characteristics of the sample data.

3.3 Autoregressive mobile support vector machine combined forecasting

While ARIMA and SVM models have their own advantages and disadvantages, but they have each advantages in linear or nonlinear model processing and they have a complementary advantages between them. Therefore, the combination of the them for passenger flow forecasting may receive better results. Assume that the time series \(Y_t\), can be regarded as the combination of the linear autocorrelation part \(L_t\) and the nonlinear residual \(N_t\), namely: \(Y_t=L_t+N_t\). This paper adopts following steps to construct a combined forecasting model:

a. Use the ARIMA model the linear part, and the forecasting result is \(L'_t\). The residual of the sequence \(L'_t\) is \(N_t\), and the nonlinear relationship of the sequence \(Y_t\) is included in \(N_t\).

b. Reconstruct the \(N_t\) sequence obtained in the previous step to obtain the SVM sample set, and forecasting the residual by the SVM to obtain the forecasting result \(N'_t\).

c. Combine the \(L'_t\) obtained by the linear forecasting with the \(N'_t\) obtained by the nonlinear forecasting to obtain a forecasting result \(Y'_t=L'_t+N'_t\).
4 ANALYSIS OF RESULTS

4.1 Data Description

The experimental data comes from Xi'an Metro Company. This paper selects the passengers' entering and exiting data from 26 Stations on the 3rd line of Xi'an Metro in 2017 from October 1st to October 8th. The interval of the data is 1h. Considering that the 2017 National Day holiday has a total of eight days unusual, (October 4th is the Mid-Autumn Festival), in order to reasonably forecast the passenger flow data, the data of 1st, 2nd, 3th, October take as data of 1st, 2nd, 3th day. The average data of 4th, 5th, October take as 4th day. The data of 6th, 7th, 8th, October take as data of 5th, 6th, 7th day. The first six days are used as training data, and the seventh day is used as test data, so that it can comparatively analyze with forecast data.

4.2 Experience Circumstance

The model in this paper is based on Python. The operating environment is Windows 10, 64-bit operating system, Intel(R) Core(TM) i5-6200 CPU @2.30GHz 2.40GHz and 8GB of installed memory.

4.3 Model Parameter Selection

Due to space limitations, we can just present the parameter setting process of the ARIMA model and the SVM model here, and the combination model of ARIMA and SVM is omitted.
A. ARIMA Model

First, the article uses "grid search" to iteratively explore different combinations of parameters, and fits a new ARIMA model with the SARIMAX() function of the statsmodels module in Python. Then, the article evaluates its overall quality. Finally, the article explore the entire range of parameters, and produce parameters which has optimal performance. The following code block is iterated by a combination of parameters and uses the SARIMAX function to adapt to the corresponding ARIMA model. After installing each SARIMAX() model, the code prints out respective AIC scores of their own.

```
warnings.filterwarnings("ignore")  # specify to ignore warning messages
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod = sm.tsa.statespace.SARIMAX(data,
                                             order=param,
                                             seasonal_order=param_seasonal,
                                             enforce_stationarity=False,
                                             enforce_invertibility=False)
            results = mod.fit()
            print('ARIMA{}x{}12 - AIC:{}'.format(param, param_seasonal, results.aic))
        except:
            continue
```

Taking the data of Dayanta Station as an example, SARIMAX(1, 1, 1)x(1, 1, 1, 12) is the best choice among all models. After the best parameter values were inserted into the new ARIMAX model, and the model training is carried out with the data of the Dayanta Station from October 1st to October 6th, 2017. We can obtain the following analysis chart:

![Figure 4: Histogram plus estimated density](image-url)
From the figure above, the red line which represents KDE is similar to the N(0,1) line, where N(0,1) is the standard symbol of the normal distribution, indicating that the residue is normally distributed well.

![Correlogram](image)

Figure 5: Autocorrelogram

It can be seen from the figure above that the time series residual has a low correlation with its own lag version.

B. SVM Model

Model parameters were calibrated using data from October 1 to October 6, 2017. In the calculation, the data of the Dayanta Station from October 1st to October 5th, 2017 is taken as historical data. The initial value of the regularization parameter is set to 100, and the initial value of the nuclear parameter is set to 0.1. First, the ten-fold cross-validation Method is adopted to calculate the forecasting value of the passenger flow on October 7th. After that, compare the forecasted value with the actual value, and then correct the regularization parameter and the kernel parameter, finally, process the forecast again. Iteratively perform this steps above until the error of the forecasted value and the actual value less than $10^{-3}$. In addition, considering that applications in practical, there may be cases where the error cannot be satisfied less than $10^{-3}$, so the number of iterations is limited to 100 times. When the error cannot satisfy $10^{-3}$, choose a set of parameters determined by the with the smallest error – time calibration results between forecasted and actual values as the best parameter value. The calibration results of the Dayanta Station are shown in Table 1.

<table>
<thead>
<tr>
<th>Station</th>
<th>Inbound passenger flow model</th>
<th>Outbound passenger flow model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>Dayanta</td>
<td>685.165</td>
<td>999.678</td>
</tr>
</tbody>
</table>

Table 1 Dayanta Station Station Calibration Results
4.4 Forecasting Results

The prediction results of the Dayanta Station are shown in the figure below.

Figure 6: Comparison of the Forecast Results of the Hourly Passenger Flow Enter the Dayanta Station

Figure 7: Comparison of the Forecast Results of the Hourly Passenger Flow Exit the Dayanta Station

It can be seen from the figure above that when using the data of Dayanta Station make forecasting, the combined forecasting model prediction results are more consistent with the actual values, which indicating that the results of the combined forecasting model are better.
The maximum relative error and average relative error of the hourly passenger flow forecast for each station on the 3rd line are counted. The results are shown in the following table.

<table>
<thead>
<tr>
<th>Station</th>
<th>Maximum Relative Error (%)</th>
<th>Average Relative Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enter</td>
<td>Exit</td>
</tr>
<tr>
<td>YU HUA ZHAI</td>
<td>15.21</td>
<td>17.33</td>
</tr>
<tr>
<td>KE JI LU</td>
<td>11.03</td>
<td>15.86</td>
</tr>
<tr>
<td>BEI CHI TOU</td>
<td>17.40</td>
<td>15.91</td>
</tr>
<tr>
<td>XIAN NING LU</td>
<td>13.63</td>
<td>15.70</td>
</tr>
<tr>
<td>TONG HUA MEN</td>
<td>17.32</td>
<td>17.94</td>
</tr>
</tbody>
</table>
The results show that the average relative prediction error of the ARIMA-SVM combined forecasting model is 9.11%(Enter) and 8.81%(Exit), the average relative forecasting error of the ARIMA model is 10.09%(Enter) and 10.25%(Exit), and the average relative forecasting error of the SVM model is 11.65%(Enter) and 12.35%(Exit). Obviously, the ARIMA-SVM combination forecast method is more practical and effective. The forecast results of subway passenger flows are affected by uncertain factors such as traffic planning, weather, holidays, and large-scale activities, which cause large fluctuations in forecast errors on some days, however the forecast error based on ARIMA-SVM combined forecasting model is within an acceptable range.

5 CONCLUSION

A. The forecasting of the passenger flow is a hot spot in transportation forecasting. By analysing the potential law of passenger flow and combining the advantages and disadvantages of each model, a combined forecasting model based on ARIMA-SVM is established to improve the accuracy of passenger flow forecasting and reduce the risk. Through the forecast and analysis of the hourly passenger flow of the subway station in Xi’an Metro Line 3, it is verified that the model has good engineering practicability and popularization, and is suitable for the forecast of subway passenger flow, which can be used for the management of the urban rail transit operation organization, which provides an important basis for decision making.

B. Forecasting of the urban rail transit passenger flow is a process of years of data accumulation, long-term adjustment and application, and gradually improved. Due to the complex factors affecting urban rail transit passenger flow, it is necessary to continuously evaluate the results of holiday passenger flow forecasting, and analyse the accuracy of passenger flow forecasting. The factors distinguish between controllable and uncontrollable.
factors, technical and non-technical factors, and constantly improve and optimize the holiday passenger flow forecasting model established in this paper.

6 ACKNOWLEDGEMENTS

This work is supported by the Shaanxi Provincial Natural Science Basic Research Project, No. 2016JM5063. And Science and Technology Program of the Ministry of Housing and Urban-Rural Development, No. 2016-K2-033.

REFERENCES


