SARIMA MODELLING APPROACH FOR RAILWAY PASSENGER FLOW FORECASTING

Miloš MILENKOVIĆ1*, Libor ŠVADLENKA2, Vlastimil MELICHAR3, Nebojša BOJOVIĆ4, Zoran AVRAMOVIĆ5

1, 4, 5Division for Management in Railway, Rolling stock and Traction, The Faculty of Transport and Traffic Engineering, University of Belgrade, Serbia
2, 3Dept of Transport Management, Marketing and Logistics, Jan Perner Transport Faculty, University of Pardubice, Czech Republic

Received 25 November 2014; revised 3 February 2015, 6 August 2015; accepted 29 October 2015; first published online 7 March 2016

Abstract. In this paper, railway passenger flows are analyzed and a suitable modeling method proposed. Based on historical data composed from monthly passenger counts realized on Serbian railway network it is concluded that the time series has a strong autocorrelation of seasonal characteristics. In order to deal with seasonal periodicity, Seasonal AutoRegressive Integrated Moving Average (SARIMA) method is applied for fitting and forecasting the time series that spans over the January 2004 – June 2014 periods. Experimental results show good prediction performances. Therefore, developed SARIMA model can be considered for forecasting of monthly passenger flows on Serbian railways.

Keywords: railway, passenger service, time series, forecasting, SARIMA.

Introduction

Forecasting of future demand for transport service represent important element of success for a transport company. Forecasting also provides basic input for planning and control of functional areas including transport operations planning, marketing and finance.

There are two broad categories of forecasting approaches in transportation field: parametric and non-parametric (Smith et al. 2002; Vlahogianni et al. 2004; Wei, Chen 2012). The main difference between these two categories lies in functional dependence between independent variables and dependent variable. Within the non-parametric methods there are applications of neural networks (Dougherty 1995; Vlahogianni et al. 2004; Xie, Zhang 2006; Zhang, He 2007; Zhang, Ye 2008; Tsai et al. 2009; Alekseev, Seixas 2009), non-parametric regression (Smith et al. 2002; Williams et al. 1998; Clark 2003), Kalman filters (Whittaker et al. 1997; Stathopoulos, Karlaftis 2003; Saab, Zouein 2001; Wang et al. 2007) and Gaussian maximum likelihood (Tang et al. 2003). For example, Wei and Chen (2012) developed a hybrid forecasting approach based on a combination of Empirical Mode Decomposition (EMD) and Back-Propagation Neural (BPN) networks to predict short term passenger flow in metro systems. Xu et al. (2004) applied spatio-temporal data mining to forecast the railway passenger flow based on spatio. Approach is decomposed into two phases. In first phase, it forecasts the time sequence of the target object using statistical principles. In second phase, it figures out the spatial influence of neighbor objects using a neural network, and finally combines the two forecasting results using linear regression. Li et al. (2012) used Grey model for forecast annual outgoing passenger flows of a railway station. Authors analyzed monthly fluctuations in passenger flows and applied proportional coefficient method for prediction of passenger flows for each month.

In the traditional parametric techniques, historical average (Smith, Demetsky 1997; Stephanedes et al. 1981), smoothing techniques (Williams et al. 1998), and AutoRegressive Integrated Moving Average (ARIMA) (Hansen et al. 1999; Williams et al. 1998; Lee, Fambro 1999; Williams 2001; Kamarianakis, Prastacos 2005) have been applied to forecast transportation demand. For example, Grubb and Mason (2001) estimated a very long time series about air passenger traffic using Holt–Winters forecasting methods. They estimated the histor-
ich growth using Holt–Winters decomposition and produced long lead-time forecasts. On the base of long data series, they evaluated the out-of-sample forecasting performance of this and other forecasting procedures and concluded that presented modification of Holt–Winter method greatly improves forecasting performance for long lead-times. Presented approach also considers an assessment of uncertainty in the predictions. By modified Holt–Winters procedure they varied the trend used for predictions and estimated the sensitivity of forecasts to assumptions about the future trend. Grose et al. (2007) presented two gravity models for the estimation of air passenger flows between pairs of cities. Models are based on variables describing economic activity and geographical characteristics of city pairs. This means that developed models can be applied in case when new service is considered or historical data are not available.

Within the class of parametric methods, AutoRegressive Integrated Moving Average (ARIMA) has become one of the common parametric forecasting approaches since the 1970s (Wei, Chen 2012). In addition, with the characteristics of seasonality and trends in traffic data, some researches use seasonal ARIMA to predict traffic flow (Williams, Hoel 2003; Tan et al. 2009) and international air passenger flow (Faraway, Chatfield 1993; Lim, McAleer 2002; Chen et al. 2009).

As already mentioned, many railway planning processes including financial, asset, capital investment and service plan are predicted on the estimate of passenger demand to be handled within a given time period. Forecasts of traffic volume represent one of the most important outputs of a marketing plan and also the important input to the corporate plan. Current planning process in JSC ‘Serbian Railways’ lacks of any planning tool, so having that in mind, as well as the actual market share and increasing competition there is a need for better planning instruments in the company. Therefore, in this paper we present the first detailed application of Seasonal ARIMA (SARIMA) models in order to generate passenger traffic forecasts for the case of JSC ‘Serbian Railways’. We use the data representing monthly passenger flows on all lines of Serbian railway network provided by Statistical Office of the Republic of Serbia (http://webrzr.stat.gov.rs). At the time of our analysis a time series of monthly passenger flows from January 2004 to June 2014 (126 months) was available.

Paper is organized as follows. Section 1 introduces ARIMA and SARIMA models and describes Box–Jenkins methodology for their selection. In Section 2, SARIMA modeling process has been applied for deriving an efficient model to assess rail passenger demand for a case of Serbian railways. Last section concludes this paper.

1. ARIMA and SARIMA models

Box et al. (2013) introduced the ARIMA method. This method now represents one of the most frequently used univariate time series modeling tools. ARIMA models are based on AutoRegressive Model (AR), the Moving Average Model (MA) and the combination of the AR and MA, the ARMA models (Suhartono 2011). AR model includes lagged terms on the time series itself, and MA model includes lagged terms on the noise or residuals. The first requirement for ARIMA modeling is that the time series data to be modeled are either stationary or can be transformed into stationary. Therefore, the letter ‘I’ (Integrated) means that the first order differencing is applied in order to stationarize given time series. First order differencing requires that there is a need to find between observations in two successive months. In case that there is a time series with trends, seasonal pattern and short time correlations a Seasonal ARIMA (SARIMA) model can be used. Besides three main components present in ARIMA model, there is a need for seasonal differencing in order to make a seasonal time series stationary. Therefore, a SARIMA model has four components:

- the non-seasonal and seasonal AR polynomial term of order $p$ and $P$, respectively:
  \[ \phi_p(B) = 1 - \phi_1B - \phi_2B^2 - \cdots - \phi_pB^p; \]
  \[ \Phi_P(Bs) = 1 - \Phi_1Bs - \Phi_2Bs^2 - \cdots - \Phi_PBs^P; \]
- the non-seasonal and seasonal Moving Average (MA) part of order $q$ and $Q$, respectively:
  \[ \theta_q(B) = 1 + \theta_1B + \theta_2B^2 + \cdots + \theta_qB^q; \]
  \[ \Theta_Q(B^s) = 1 + \Theta_1B^s + \Theta_2B^{2s} + \cdots + \Theta_QB^{qs}; \]
- non-seasonal differencing operator, \((1 - B)^d\) is the of order $d$ used to eliminate polynomial trends;
- seasonal differencing operator, \((1 - B^s)^D\), of order $D$ used to eliminate seasonal patterns.

Parameters $\phi$ and $\theta$ are the ordinary ARMA coefficients, $\Phi$ and $\Theta$ are the seasonal ARMA coefficients, $B$ is the backshift operator, whose effect on a time series $Y_t$ can be summarized as $B^sY_t = Y_{t-d}$.

Therefore, the generalized form of SARIMA\((p, d, q) \times (P, D, Q)\) model for a series $Y_t$ can be written as (Box et al. 2013; Cryer, Chan 2008):

\[ \phi_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)e_t, \]

where: $s$ is the length of the periodicity (seasonality) and $e_t$ is a white noise sequence.

Box–Jenkins method in essence involve three steps for fitting the ARIMA and SARIMA models. These are identification, estimation and validation of the model (Box et al. 2013; Prista et al. 2011). First step is generally based on an analysis of AutoCorrelation Function (ACF) and Partial ACF (PACF) and their comparison with theoretical profiles of these functions in AR, MA and ARMA processes. The output of the identification step is determined appropriate \((p,d,q) \times (P,D,Q)\) model structure. The model structure selected in previous step has to be fitted to the time series and its parameters to be estimated. This is the essence of the second step and it is done by using the conditional sum of squares or maximum likelihood method. Validation of selected model is
performed within the diagnostic checking by analysis of stationarity, invertibility as well as the presence of redundancy in model parameters. If a selected model fails in diagnostic check it is necessary to repeat the whole procedure again. After an appropriate model is found it can be used for forecasting purpose (Box et al. 2013; Prista et al. 2011).

Subjectivity is mostly present in model identification step. In the fact that this step is based essentially on graphical interpretations of ACF/PACF estimates. To cope with this subjectivity and to improve the determination the final orders of the ARMA processes there are a lot of model selection criteria proposed in literature (De Gooijer et al. 1985).

The most frequently used are information criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and normalized version of BIC. These information criteria are designed to deal with the fit of the nonlinear models and to account for the number of the parameters in the model as well. They consist of the natural log of the Mean Squared Error (MSE) and a penalty for the number of parameters being estimated (Yaffee, McGee 2000):

\[
\text{AIC} = T \ln(\text{MSE}) + 2k; \quad (6)
\]

\[
\text{BIC} = T \ln(\text{MSE}) + k \ln(T); \quad (7)
\]

\[
\text{Normalized BIC} = \ln(\text{MSE}) + k \frac{\ln(T)}{T}, \quad (8)
\]

where: \( T \) is number of observations, \( k \) is the number of parameters in the model \( k = p + q + P + Q + 1 \). Common procedure often involves estimating the largest model for which it is assumed to correctly capture the dynamics of a time series and then decreasing its size (dropping the lags) until the minimum value of AIC, BIC or Normalized BIC is reached. Presented methodology will be used in next chapter for analysis of railway passenger flows on Serbian railways.

2. SARIMA model for rail passenger flow forecasting: a case study of Serbian railways

In this chapter, we outline the practical steps, which need to be undertaken to use SARIMA time series models for passenger flows forecasting on Serbian railways. To test alternative SARIMA models we obtained a time series of a total monthly number of passengers travelled by the railway from the Statistical Office of the Republic of Serbia (http://webrzs.stat.gov.rs). Data set covers the period from January 2004 to June 2014 (126 monthly values). The time series is presented on Figure 1. First 114 monthly observations are used for fitting purpose, whereas the remaining 12 observations served for verification the forecasting capability of selected SARIMA model.

According to the Box–Jenkins methodology, in SARIMA models there is an explicit consideration of trend and seasonal non-stationarities by the model structure. However, it is needed to analyze non-stationarities of variance before the model fitting. As it can be seen from the Figure 1, time series data \( (x_t, t = 1, \ldots, 126) \) for rail passenger traffic in Serbia have strong seasonality pattern with a trend, which is not constant throughout the whole study period, rather slightly decreasing in period from 2004 to 2010 and increasing in period from 2011 to 2014, an increase in variance with the serious mean is identified from variance–mean plots (Figure 2).

Therefore, we log-transformed the data in order to stabilize the variance and then used these log transformed data set as input to the SARIMA analysis (Figure 3). Different SARIMA models were applied to find the best fitting model. Among all candidate models, the most suitable predictive model for our case was found on the base of Box–Jenkins model and normalized BIC. This approach involved selection of the candidate model set, estimation of the model and determination of normalized BIC and diagnostic check.
The set of alternative models is determined by analyzing the sample estimates of ACF and PACF, and defining the three major orders of the SARIMA models: \(d\), \(D\), and \(S\). Lags 1, 2, 3, 4, 9, 10, 11, 12, 13, 14, 22, 23, 24, 25, 26, 29 and 30 had large autocorrelations with values 0.80, 0.64, 0.45, 0.28, 0.25, 0.36, 0.46, 0.51, 0.38, 0.27, 0.22, 0.31, 0.39, 0.29, 0.19, –0.17 and –0.26, respectively (Figure 4). According to these values, namely, the very sharp decrease of autocorrelation values in first three lags; it is obvious that there exists a long-term trend. In order to eliminate it, the first lag difference term has to be included in the structure of the SARIMA model \(d = 1\). Annual lags and its multiplies also reported large autocorrelation values. Consequently, seasonal difference terms are included in models \(S = 12, D = 1\). Further justification of these conclusions is obvious from ACF and PACF plots of differenced time series (Figure 4). Based on this conclusions, we selected SARIMA \((p,1,q) \times (P,1,Q)_s\) as the basic structure of the alternative SARIMA models.

Among the statistical models, SARIMA \((0,1,0) \times (0,1,1)_12\) was the most appropriate with the lowest normalized BIC of 7.056 and a Mean Absolute Percent Error (MAPE) of 4.176 (Table 1). 70.8% of the variance of the time series was covered by this model (stationary \(R^2\)). According to the Ljung–Box statistics it is proven that the selected model is correctly specified. Namely, the value of 0.542 shown here is not significant, so we can be confident in correct specification of the model. This model has the following equation:

\[
\log Z_t = \log Z_{t-1} + \log Z_{t-12} - \log Z_{t-13} + \varepsilon_t + 0.966\varepsilon_{t-12}.
\]

(9)
Table 1. Normalized BIC, MAPE and stationary R-squared of SARIMA models

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE</th>
<th>Normalized BIC</th>
<th>Stationary R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA(0,1,1)×(0,1,1)_{12}</td>
<td>4.324</td>
<td>7.199</td>
<td>0.675</td>
</tr>
<tr>
<td>SARIMA(0,1,1)×(0,1,0)_{12}</td>
<td>5.452</td>
<td>7.443</td>
<td>0.545</td>
</tr>
<tr>
<td>SARIMA(0,1,0)×(0,1,1)_{12}</td>
<td>4.176</td>
<td>7.056</td>
<td>0.708</td>
</tr>
<tr>
<td>SARIMA(1,1,1)×(0,1,1)_{12}</td>
<td>4.303</td>
<td>7.233</td>
<td>0.679</td>
</tr>
<tr>
<td>SARIMA(1,1,1)×(1,1,1)_{12}</td>
<td>4.304</td>
<td>7.288</td>
<td>0.679</td>
</tr>
</tbody>
</table>

According to diagnostic checks SARIMA model appears as stationary and invertible without redundant parameters. According to the Ljung–Box Q-test and p-value (Ljung–Box Q = 15.747, p-value > 0.05) residuals were white noise and therefore there is no significant auto-correlation between residuals at different lag times (Figure 5). Significance level of $\alpha = 0.05$ was used for all tests.

Twelve months of model forecasts are evaluated where the first month (July 2013) after the fitting part represent the origin of the forecast. Forecasts were obtained in the data original scale $(\hat{x}_l, l = 1, \ldots, 12)$. Performance of selected SARIMA model was evaluated by comparison of $l$-step forecasts and monthly number of passengers travelled in period July 2013 – Jun 2014. In order to perform this comparison monthly forecast errors are computed and following accuracy criteria applied: Root Mean Square Error (RMSE), Mean Error (ME), Absolute Percent Error (APE) and MAPE. Performance of SARIMA model are compared also against the Seasonal Exponential Smoothing (SES) method (Table 2). For this purpose trial version of IBM SPSS (version 19) software is used.

![Residual ACF and Residual PACF](image)

Figure 5. ACF and PACF of residuals

Table 2. Forecasts of railway passenger flows in thousands (July 2013 – June 2014): observed passenger counts $(x_l)$, forecasted passenger counts $(\hat{x}_l)$, monthly forecast errors $(e_l)$ and monthly absolute percent error (APE) are displayed for Seasonal Exponential Smoothing (SES) model and SARIMA (SAR) model

<table>
<thead>
<tr>
<th>Month, year</th>
<th>Step $(l)$</th>
<th>Obs$(x_l)$</th>
<th>Forecasts $(\hat{x}_l)$</th>
<th>Forecast errors $(e_l)$</th>
<th>APE$_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SES</td>
<td>SAR</td>
<td>SES</td>
</tr>
<tr>
<td>July 2013</td>
<td>1</td>
<td>664</td>
<td>660.12</td>
<td>663.03</td>
<td>-3.88</td>
</tr>
<tr>
<td>August 2013</td>
<td>2</td>
<td>623</td>
<td>655.84</td>
<td>641.02</td>
<td>32.84</td>
</tr>
<tr>
<td>September 2013</td>
<td>3</td>
<td>628</td>
<td>583.49</td>
<td>574.15</td>
<td>-44.51</td>
</tr>
<tr>
<td>October 2013</td>
<td>4</td>
<td>614</td>
<td>598.45</td>
<td>598.03</td>
<td>-15.55</td>
</tr>
<tr>
<td>November 2013</td>
<td>5</td>
<td>574</td>
<td>547.04</td>
<td>546.12</td>
<td>-26.96</td>
</tr>
<tr>
<td>December 2013</td>
<td>6</td>
<td>566</td>
<td>540.73</td>
<td>539.27</td>
<td>-25.27</td>
</tr>
<tr>
<td>January 2014</td>
<td>7</td>
<td>531</td>
<td>526.29</td>
<td>516.53</td>
<td>-4.71</td>
</tr>
<tr>
<td>February 2014</td>
<td>8</td>
<td>507</td>
<td>485.6</td>
<td>488.51</td>
<td>-21.40</td>
</tr>
<tr>
<td>March 2014</td>
<td>9</td>
<td>610</td>
<td>567.81</td>
<td>557.16</td>
<td>-42.19</td>
</tr>
<tr>
<td>April 2014</td>
<td>10</td>
<td>600</td>
<td>613.22</td>
<td>601.82</td>
<td>13.22</td>
</tr>
<tr>
<td>May 2014</td>
<td>11</td>
<td>537</td>
<td>658.05</td>
<td>645.82</td>
<td>121.05</td>
</tr>
<tr>
<td>June 2014</td>
<td>12</td>
<td>515</td>
<td>649.67</td>
<td>649.48</td>
<td>134.67</td>
</tr>
<tr>
<td>Mean</td>
<td>1:12</td>
<td>580.75</td>
<td>590.52</td>
<td>585.08</td>
<td>9.77</td>
</tr>
<tr>
<td>Sum</td>
<td>1:12</td>
<td>6969</td>
<td>7086</td>
<td>7021</td>
<td>117</td>
</tr>
</tbody>
</table>
Figure 6 displays observed values and SARIMA model fit and forecast values. Model gives slightly lower forecasts than observed values (September 2013 – March 2014), but pattern of model forecasts almost matched the one in observed passenger counts except for period from May 2014 to June 2014. The main reason for this is a slightly increasing trend for the last two years of the study period. RMSE during the prediction period (RMSE = 56.45) was 1.9 times the RMSE of the fitting period (RMSE = 29.69). Negative errors and low ME (ME = 4.32) registered in eight of twelve forecasts indicate minority of underestimation in global terms. However, it has to be noted that this conclusion is valid for the case of random error, but this will not be the case if the error is systematic (system changes – transportation, activity or flow patterns). MAPE was 7.13% reflecting the slightly lower forecasts during the July 2013 and September 2013 – March 2014. The SARIMA model forecasts also registered better performances with respect to SES, resulting in 9.8% reduction in RMSE, 44.2% reduction in ME, and 9.7% reduction in MAPE.

Outliers are known to cause trouble in time series model identification, estimation and forecasts (Prista et al. 2011). Rail passenger flow data set presented four apparent outliers, which are incorporated into the model (Figure 7). They are of different kinds. These outliers are mainly additive that affects a single observation (August 2007, September 2009 and February 2012), but there is also one transient outlier whose impact exponentially decays (June 2010). The main reason for existence of outliers at the beginning of year (like those in February 2012) lies in winter vacations and non-existence of daily and weekly school migrations. During the August 2007 significant flows appeared especially within the international passenger services. Higher volumes of passenger flows in September 2009 resulted by ending of summer vacations and beginning the activities at schools and faculties. June of 2010 in Serbia was very hot and this reflected on lower volumes of passenger flows on Serbian railways.

Conclusions

In this paper, we proposed a sophisticated mathematical model for generating rail passenger traffic forecasts for the case of JSC ‘Serbian Railways’. Different SARIMA models were tested to select an appropriate. The results of diagnostic check indicate that the SARIMA(0,1,0) × (0,1,1)12 is the most appropriate for modeling the rail passenger demand on Serbian railways. This method can be a valuable decision support tool for company’s marketing sector considering from one side, the current situation of rail passenger sector in Serbia (market share, revenues generated, number of passengers) and lack of any sophisticated planning tools, and from the other side, the accuracy of forecasting approach.
Future research will be concentrated on developing more efficient forecasting models, which will also use additional information from market surveys and demographic studies.

References


