Forecasting tourist arrivals to Turkey

Yılmaz, Engin

Universidad Carlos III de Madrid

28 December 2015
Engin Yılmaz

Forecasting tourist arrivals to Turkey

Abstract

Modeling and forecasting techniques of the tourist arrivals are many and diverse. There is no unique model that exactly outperforms the other models in every situation. Actually, a few studies have realized modeling and forecasting the tourist arrivals to Turkey and these studies have not focused on the total tourist arrivals. These studies have focused on the tourist arrivals to Turkey country by country (or OECD countries). In addition to this, structural time series models have not been used in modeling and forecasting the tourist arrivals to Turkey. In this sense, this paper is the first study which uses the seasonal autoregressive integrated moving average model and the structural time series model in order to forecast the total tourist arrivals to Turkey. Two different models are developed to forecast the total tourist arrivals to Turkey using monthly data for the period 2002-2013. The results of the study show that two models provide accurate predictions but the seasonal autoregressive integrated moving average model produces more accurate short-term forecasts than the structural time series model. It is noted that the seasonal autoregressive integrated moving average model shows a very successful performance in the forecasting the total tourist arrivals to Turkey.

Key words: structural time series models; arima; tourist arrivals; tourist demand; Turkey

Introduction

Tourism plays a crucial role in the emerging economies all around the world. The tourism sector is a significant contributor to employment, tax revenues and earnings of foreign exchange. These are only the direct effects of the tourism in these countries. However, we should take into account the externalities of the tourism in the economic system. This sector has created many externalities in the economic system. For instance, it has created new infrastructure facilities, telecommunications opportunities and has augmented interconnection of all the sectors (such as construction, agriculture, entertainment and fishing sector) of the economy. Today, the business volume of tourism equals or even surpasses that of oil exports, food products or automobiles (UNWTO, 2014). Supporting tourism domestically and internationally has been a priority for the emerging countries. Many of these countries grow rapidly thanks to tourism revenues.

Turkey is an emerging country, a candidate country for European Union membership, and one of the attractive touristic places in the south of Europe. Turkey is currently the 6th most popular tourist destination in the world (UNWTO, 2014), attracting more than 30 million tourists each year, and this number grows year by year. The direct contribution of tourism to GDP in 2014 was 41.1bn dollars (WTTC, 2015). It is expected that this contribution will grow in the future years. In spite of the
significance of tourism in Turkey, there are relatively limited studies on modeling and forecasting the total tourist arrivals. The contribution of this paper is to model the total tourist arrivals to Turkey by using the structural time series model (STM) and the seasonal autoregressive integrated moving average model (SARIMA). It compares the structural time series model (STM) with the seasonal autoregressive integrated moving average model (SARIMA) from the point of forecasting accuracy.

Literature review

There are many studies on tourist arrivals. Lim (1997), Li, Song and Witt (2005) and Goh and Law (2011) have realized a detailed review of these studies. There are two main methods in the literature of the modeling and forecasting tourist arrivals: the causal econometric approach and the time series models. The causal econometric approaches are based on the causal relationship between the demand factors and the total tourist arrivals (Song & Li, 2008). Some of these demand factors are countries’ real income, the relative prices, the competitive prices, the exchange rates, the transportation costs, the population and the accommodation costs. Witt and Witt (1995) and Kulendran and King (1997) have concluded that the univariate time series models tend to outperform the causal econometric models. Athanasopoulos, Hyndman, Song and Wu (2011) have researched the time series approaches and the causal econometric models on tourist arrivals; they have indicated that the time series models are better than the causal econometric models. Many studies have used the univariate time series models, such as Preez and Witt (2003), Wong, Song, Witt and Wu (2007), Chu (2008a), Lee, Song and Mjelde (2008), Coshall (2008) and Kulendran and Witt (2001). Chu (2008b) has found that the autoregressive fractionally integrated moving average model (ARFIMA) exhibits the highest forecasting accuracy both in the short-run and in the long-run, but, the SARIMA is the best performing model in the medium-run.

However, univariate time series models do not ensure the analytic comprehensiveness of the dynamic characteristics of these series. Other time series approach in this issue is the structural time series models. Structural time series models center upon the time series components (trend, seasonal, cycle and irregular), Turner and Witt (2001), Kim and Moosa (2001), Greenidge (2001) and Greenidge and Jackman (2010) have shown that the structural time series models are capable of providing reasonably accurate forecasts.

The recent studies about modeling and forecasting of the tourist arrivals have emphasized that ARIMA models have an important superiority on this issue. Torra and Claveria (2014) have compared the forecast accuracy of the different methods for modeling tourist arrivals to Catalonia and have concluded that ARIMA models outperformed self-exciting threshold autoregressions (SETAR) and artificial neural network models (ANN), especially for shorter horizons. Hassani, Silva, Antonakakis, Filis and Gupta (2015) have realized the most comprehensive forecasting comparison among several parametric and non-parametric techniques for modeling European tourist arrivals and have laid emphasis on there is not a single model that its forecasting accuracy consistently outperforms that of all other models. However, Hassani et al (2015) more specifically have indicated that Singular Spectrum Analysis algorithms (SSA), Trigonometric Box-Cox ARMA Trend Seasonal (TBATS) and ARIMA models are viable options for modeling European tourist arrivals.
Akis (1998), Halıcıoğlu (2004), Aslan, Kaplan, Muhittin and Kula (2008), Göçer and Gürmüş (2010), Aktürk and Kıcıközmen (2006) have used the causal econometric approaches for modeling the tourist arrivals to Turkey. Akal (2003) has used the time series model in the forecasting tourist arrivals to Turkey. Akal (2003) has indicated that the autoregressive model (AR) was capable of producing valid modeling of tourist arrivals to Turkey. Akın (2015) has compared SARIMA model, support vector machine model (SVR) and neural network model (NN) in order to forecast the tourist arrivals to Turkey. Akın (2015) has found that the support vector machine model (SVR) is the best approach; SARIMA model is the second best approach and a neural network model (NN) is the third best approach. These studies generally have preferred to focus on only top ten countries’ or OECD countries’ tourist arrivals to Turkey. In this paper, it is preferred to focus on the total tourist arrivals. Tourism dataset includes the total tourist arrivals of 94 countries. This paper is to model the total tourist arrivals to Turkey by using the structural time series model (STM) and the seasonal autoregressive integrated moving average model (SARIMA).

Data

The performance of Turkish tourism industry has been notable in the recent years. Turkey has realized the highest performance in the tourist arrivals and the total tourist arrivals have been increased average %15 in the period of 2002-2013. The data of the total tourist arrivals from 94 countries to Turkey in the period of January 2002-December 2013 are obtained from Turkish Ministry of Tourism and Culture (General Directorate of Investment and Enterprises, 2014). Ministry of Tourism has been providing the data of the number of foreign visitors entering into the country. The original data of the total tourist arrivals was started in 2000 but it is chosen 2002 as a starting date in this paper. It is well known that Turkey had experimented the banking and financial crisis in 2001 and therefore it is chosen 2002 as a first year for the accurate analysis. The total tourist arrivals to Turkey have risen in recent decades. The number of tourist arrivals increased from 13 million in 2002 to 35 million in 2013. The total tourist arrivals to Turkey from 94 countries in the period of 2002-2013 are shown in Figure 1.

Figure 1
Total tourist arrivals to Turkey (monthly)

Source: Republic of Turkey Ministry of Culture and Tourism, General Directorate of Investment and Enterprises, 2014.
It is clear that the data contain clear seasonal structure and this is one of the dynamic characteristics of the total tourist arrivals to Turkey. It is not seen any break (shock) or breaks (shocks) in the data of the total tourist arrivals but it is expected that the models are used in this paper will detect the break (shock) or breaks (shocks).

Model

Firstly the autoregressive integrated moving average model is used in estimating the total tourist arrivals. The seasonal autoregressive integrated moving average model can be written as follows,

\[
\Phi(L^{12}) \varphi(L) \Delta^D \Delta^d y_t = \mu + \Theta(L^{12}) \theta(L) \xi_t
\]  

(1)

and defined by \( (p, d, q) \times (P, D, Q) \) \( 12 \).

\( \Delta^d \) is the first degree nonseasonal differencing operator, \( \Delta^D \) is the first degree seasonal differencing operator, \( \mu \) is the mean, \( L \) is the lag operator, \( \Phi(L^{12}), \varphi(L), \Theta(L^{12}), \theta(L) \) are polynomials of order \( P, p, Q \) and \( q \). The error term \( \xi_t \) is white noise with zero mean and constant variance.

The basic structural model (Harvey, 1989) is

\[
y_t = \mu_t + \gamma_t + \epsilon_t \quad t = 1, ..., T
\]  

(2)

where \( \mu_t, \gamma_t \) and \( \epsilon_t \) are the trend, seasonal and irregular components, respectively. The irregular component, \( \epsilon_t \), is assumed to be random, and the disturbances in all three components are taken to be mutually uncorrelated.

The process generating the trend can be regarded as a local approximation to a linear trend, i.e

\[
\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t
\]  

(3)

\[
\beta_t = \beta_{t-1} + \zeta_t \quad t = 1, ..., T
\]  

(4)

where \( \eta_t \) and \( \zeta_t \) are distributed independently of each other and over time with mean zero and variances \( \sigma^2_\eta \) and \( \sigma^2_\zeta \), \( \mu_t \) and \( \beta_t \) represent the level and slope of the trend (Harvey, 1989). The process generating the seasonal component is

\[
\gamma_t = - \sum_{j=1}^{s-1} \gamma_{t-j} + \omega_t \quad t = 1, ..., T
\]  

(5)

where \( \omega_t \) is an independently distributed disturbance term with mean zero and variance \( \sigma^2_\omega \) and \( s \) is the number of ‘seasons’ in the year (Harvey, 1989). The estimation procedure is done by casting the model in state space form and applying Kalman Filtering (Harvey, 1989).

The state space form of the Basic Structural Model (BSM) is given by the following representation (LaCalle, 2014):
\( y_t = Z \alpha_t + \varepsilon_t \quad \varepsilon_t \sim \text{NID}(0, \sigma^2_\varepsilon) \) \hspace{1cm} (6)

\( \alpha_t = T \alpha_{t-1} + R \eta_t \quad \eta_t \sim \text{NID}(0, V) \) \hspace{1cm} (7)

\[\begin{align*}
\alpha_0 &\sim \text{N}(a_0, P_0) \\
\sigma^2_\eta &\quad 0 \quad 0 \\
V &\quad 0 \quad \sigma^2_\xi \\
&\quad 0 \quad 0 \quad \sigma^2_\omega
\end{align*}\] \hspace{1cm} (8)

for \( t = 1; : : : ; n \).

It is assumed that \( a_0 \) and \( P_0 \) are known and variances are given, Kalman filter can be applied to extract an estimate of the latent components (level, trend and seasonal).

**Results**

It is analyzed firstly SARIMA model in order to forecast the total tourist arrivals to the Turkey. It is used the econometric program TRAMO in the calculation and analysis. TRAMO is a program for estimation and forecasting of regression models with possibly nonstationary ARIMA errors and any sequence of missing values (Gomez & Maravall, 1999). The program interpolates these values, identifies and corrects for several types of outliers, and estimates special effects such as Trading Day and Easter and, in general, intervention-variable type effects (Gomez & Maravall, 1999). This program is prepared by Spanish Central Bank and it estimates the best suited autoregressive moving average model. The basic methodology followed in this program is described in Gomez and Maravall (1994) and Gomez, Maravall and Pena (1999). The program tests for the log/level specification, interpolates missing observations (if any), and performs automatic model identification and outlier detection (Gomez & Maravall, 1999).

<table>
<thead>
<tr>
<th>Interceptor</th>
<th>Trend + Intercept</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>L(Tourist)</td>
<td>-1.44</td>
<td>-2.09</td>
</tr>
<tr>
<td>DL(Tourist)</td>
<td>-4.31</td>
<td>-4.42</td>
</tr>
</tbody>
</table>

In the Table 1, "Tourist" variable represents the data of the total tourist arrivals to Turkey, "L(Tourist)" represents the logarithmic data of the total tourist arrivals to Turkey and "DL(Tourist)" represents the differenced logarithmic data of the total tourist arrivals to Turkey.

Firstly, it is computed the augmented Dickey-Fuller (ADF) test for the unit root hypothesis. Following ADF test (Table 1), it is concluded that the data of the total tourist arrivals is not stationary in level but is stationary at first difference. The automatic procedure selects \((0,1,1) (0,1,1)_{12}\) model for the levels. The model parameters are presented in Table 2.
The estimated SARIMA model is seen as following.

\[ \Delta^{12} y_t = (1-0.54L) (1-0.53L^{12}) \varepsilon_t \]  \hspace{1cm} (9)

TRAMO has found out one outlier (Transitory change) in March of 2003. This probably reflects USA intervention in Iraq in March 2003. USA intervention in Iraq during the first quarter of 2003 had a worse effect on the total tourist arrivals to Turkey.

The Ljung-Box test is used for testing the autocorrelation assumption in residuals. In the Table 3, it is seen clearly that there is not the autocorrelation problem in residuals. Jarque-Berra residual normality test is used for testing the normality assumption in residuals. All residuals are distributed normally. McLeod-Li test is used for verifying the presence of autoregressive conditional heteroscedasticity in residuals. The residuals of the model are not subject to the effect of autoregressive conditional heteroscedasticity. This model passes all the tests and is identified as a good model by the program. A summary of the some important statistics test of the residuals are given in the Table 3.

**Table 3**

<table>
<thead>
<tr>
<th>Statistics tests</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td>3.965</td>
</tr>
<tr>
<td>Q(24)</td>
<td>32.22</td>
</tr>
<tr>
<td>Q2</td>
<td>0.16**</td>
</tr>
</tbody>
</table>

*Normality test is Jarque-Berra residual normality test (Chi-squared value) (Critical values %1 [9.21], %5 [5.99], %10[4.61]).

*Q(24) is Ljung-Box Q Value of order 24 (Chi-Squared Value) (Critical values %1 [42.98], %5 [36.41], %10[33.19]).

*Q2 is McLeod-Li test, for the presence of autoregressive conditional heteroscedasticity ("p-value").

It is analyzed secondly STM for modeling the total tourist arrivals to Turkey. This data is modeled using STAMP econometric package. STAMP is an econometric package for the analysis of both univariate and multivariate state-space models written by Koopman, Harvey and Doornik (2000).

It is started the structural time series analysis with the basic structural model (see equation 2). The model is estimated in logarithm. It is included the fixed seasonal component\(^3\).

The value of the stochastic slope in the model (see equation 4) is found out zero (\( \zeta_t = 0 \)), so this model is re-estimated with fixed slope and this new formulation for trend is described in the following model.

\[ \mu_t = \mu_{t-1} + \beta_t + \eta_t \]  \hspace{1cm} (10)
The statistical results of the fixed slope model are presented in Table 4. It is seen a little bit normality problem in the residuals of this model. It is seen clearly from Ljung-Box test that there is not any autocorrelation problem in residuals. A simple test for heteroscedasticity (H) is obtained by comparing the sum of squares of two exclusive subsets of the sample (Koopman & Ooms, 2006). STAMP do not give the critical values of this test, one can understand whether there is heteroscedasticity or not from the residual graphs. It is concluded from the graphical analysis that there is not heteroscedasticity in the residuals. It is added BIC (Bayesian Information Criteria) for evaluating and comparing model fixed slope with other models.

Table 4
Statistical results of STM (Fixed slope model)

<table>
<thead>
<tr>
<th></th>
<th>Fixed slope model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td>6.61</td>
</tr>
<tr>
<td>H</td>
<td>0.41</td>
</tr>
<tr>
<td>Q(24)</td>
<td>31.96</td>
</tr>
<tr>
<td>BIC</td>
<td>-4.78</td>
</tr>
</tbody>
</table>

*Normality test is Bowman and Shenton residual normality test (Chi-squared value) { Critical values %1 [9.21], %5 [5.99], %10[4.61] }.
*Q(24) is Ljung-Box Q Value of order 24 (Chi-squared value) { Critical values %1 [42.98], %5 [36.41], %10[33.19] }.

It is focused on the possible outliers in the data. STAMP econometric package is able to detect automatically outliers in the data. STAMP found out an outlier in March of 2003. This probably reflects USA intervention in Iraq in March 2003. It is convenient with the SARIMA model results. It is estimated the fixed slope model with this intervention and this new formulation is described in the following model.

\[
y_t = \mu_t + \gamma_t + \text{Intervention(2003.3)} + \epsilon_t \quad t=1,\ldots,T
\] (11)

The statistical results of the fixed slope model are presented in Table 5. The residuals are normally distributed. It is seen clearly from Ljung-Box test that there is not autocorrelation problem in residuals. It is concluded from the graphical analysis that there is not heteroscedasticity problem in the residuals.

Table 5
Statistical results of STM

<table>
<thead>
<tr>
<th></th>
<th>Fixed slope model + Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td>2.22</td>
</tr>
<tr>
<td>H</td>
<td>0.45</td>
</tr>
<tr>
<td>Q(24)</td>
<td>31.11</td>
</tr>
<tr>
<td>BIC</td>
<td>-4.80</td>
</tr>
</tbody>
</table>

*Normality test is Bowman and Shenton residual normality test (Chi-squared value) { Critical values %1 [9.21], %5 [5.99], %10[4.61] }.
*Q(24) is Ljung-Box Q Value of order 24 (Chi-squared value) { Critical values %1 [42.98], %5 [36.41], %10[33.19] }.

This model passes all the diagnostic tests (Normality, Heteroskedasticity and Autocorrelation) and its Bayesian Information Criterion has better than the fixed slope model. The statistics results indicate that the fixed slope model with intervention performs well in terms of diagnostic testing.
Model parameters of the fixed slope model with intervention are presented in Table 6. The stochastic trend plays a dominant role in explaining the total tourist arrivals to Turkey. Other dominant role is taken from the seasonality variable and all of them are significant in this model. Intervention (2003.3) variable and the fixed slope variable are found a significant but they little bit contribute to this model. To examine the forecasting accuracy of the alternative models, one-step-ahead ex post forecasts for each model are generated in 2014. The forecasting result of the SARIMA is indicated in the Figure 2.

Figure 2
SARIMA forecasting values and actual values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>R.M.S.E</th>
<th>T value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>14.75</td>
<td>0.03</td>
<td>385.56</td>
</tr>
<tr>
<td>Slope</td>
<td>0.008</td>
<td>0.003</td>
<td>2.19</td>
</tr>
<tr>
<td>Intervention(2003.3)</td>
<td>-0.17</td>
<td>0.06</td>
<td>-2.76</td>
</tr>
<tr>
<td>Sea_1</td>
<td>-0.84</td>
<td>0.01</td>
<td>-46.362</td>
</tr>
<tr>
<td>Sea_2</td>
<td>-0.73</td>
<td>0.01</td>
<td>-40.12</td>
</tr>
<tr>
<td>Sea_3</td>
<td>-0.37</td>
<td>0.01</td>
<td>-20.08</td>
</tr>
<tr>
<td>Sea_4</td>
<td>-0.09</td>
<td>0.01</td>
<td>-5.43</td>
</tr>
<tr>
<td>Sea_5</td>
<td>0.35</td>
<td>0.01</td>
<td>19.36</td>
</tr>
<tr>
<td>Sea_6</td>
<td>0.48</td>
<td>0.01</td>
<td>26.91</td>
</tr>
<tr>
<td>Sea_7</td>
<td>0.72</td>
<td>0.01</td>
<td>39.96</td>
</tr>
<tr>
<td>Sea_8</td>
<td>0.65</td>
<td>0.01</td>
<td>36.41</td>
</tr>
<tr>
<td>Sea_9</td>
<td>0.50</td>
<td>0.01</td>
<td>27.74</td>
</tr>
<tr>
<td>Sea_10</td>
<td>0.28</td>
<td>0.01</td>
<td>15.64</td>
</tr>
<tr>
<td>Sea_11</td>
<td>-0.38</td>
<td>0.01</td>
<td>-20.93</td>
</tr>
</tbody>
</table>

*Trigonometric seasonal component contains 11 different variables and their results. Author can send this information if reader demands.

*R^2* is a measure of goodness-of-fit in the structural time series, its value is 0.95 for this model.


*R.M.S.E: Root mean square errors.

*Sea_: Seasonal variable.
All the actual values lie within prediction interval of SARIMA. August, September, October, November are less predictive months in this forecasting. January, March, May and July are more predictive months in this forecasting. The forecasting result of STM is indicated in the Figure 3.

Figure 3
STM forecasting values and actual values

All the actual values lie within the prediction interval of STM. February, April, July, October and November are less predictive in this forecasting. June, August and December are more predictive in this forecasting.

These forecasting results are evaluated using the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These errors are calculated as averaging of twelve months. The monthly average MAE and MAPE results are shown in the Table 7.

Table 7
Evaluation of SARIMA forecasting and STM forecasting (average monthly)

<table>
<thead>
<tr>
<th></th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA</td>
<td>82.637</td>
<td>% 2.95</td>
</tr>
<tr>
<td>STM</td>
<td>135.488</td>
<td>% 4.90</td>
</tr>
</tbody>
</table>

The forecasting result of SARIMA is better than STM because the averaging monthly mean absolute percentage error is only % 2.95 in the SARIMA and % 4.90 in the STM. It is found that SARIMA has a high predictive capacity since its mean absolute error and mean absolute percentage error have the smallest forecast errors when compared with STM.
Conclusions

Tourist arrivals are regarded as a significant measure of tourism demand, which have been used frequently in tourism demand modeling and forecasting. The one of the most important problem in tourism demand is the modeling and forecasting tourist arrivals to the country. Modeling and forecasting the total tourist arrivals to Turkey is essential for the tourism policy of Turkey. Actually a few studies have realized modeling and forecasting the tourist arrivals to Turkey and these studies have not focused on the total tourist arrivals. They have focused on the tourist arrivals to Turkey country by country (or OECD countries). This paper is the first study which models and forecasts the total tourist arrivals to Turkey. It is shown that SARIMA model is a powerful tool when modeling and forecasting the total tourist arrivals to Turkey.

Notes

1 One can chose 3 different seasonal component in STAMP; Trigonometric, Fixed and Dummy. These are seen in the Harvey (1989).

References


Submitted: 10/06/2015
Accepted: 19/10/2015