Tourism Management 48 (2015) 64–72

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

A novel approach to model selection in tourism demand modeling

Melda Akın*

Istanbul University, Department of Econometrics, Beyazıt, Istanbul, Turkey

HIGHLIGHTS

- We compare Seasonal ARIMA, Support Vector Regression, and Neural Network models.
- We use monthly tourist arrival data to Turkey from different countries.
- We identify the components of the time series using structural time series modeling.
- We obtain a rule set for model selection using identified components.

ARTICLE INFO

Article history:
Received 7 February 2013
Accepted 5 November 2014
Available online

Keywords:
Time series
Neural networks
SARIMA
Support Vector Regression
Particle swarm optimization
Structural time series modeling
C5.0 algorithm
Tourism data

ABSTRACT

In many studies on tourism demand modeling, the main conclusion is that none of the considered modeling approaches consistently outperforms the others. We consider Seasonal AutoRegressive Integrated Moving Average, ν-Support Vector Regression, and multi-layer perceptron type Neural Network models and optimize their parameters using different techniques for each and compare their performances on monthly tourist arrival data to Turkey from different countries. Based on these results, this study proposes a novel approach to model selection for a given tourism time series. Our approach is based on identifying the components of the given time series using structural time series modeling. Using the identified components we construct a decision tree and obtain a rule set for model selection.

1. Introduction

Modeling tourism demand is important in countries where the income from tourism constitutes a considerable percentage of their gross domestic product. In 2011 Turkey moved to sixth place in the World in terms of tourist arrivals. However, it is twelfth in terms of tourism receipts (UNWTO, 2012). Understanding the nature of such discrepancies and establishing measures to improve them is essential.

In the literature there are many studies on tourism demand modeling. Lim (1997) gives a detailed review of 100 studies. Song and Li (2008) have reviewed 121 studies on tourism demand modeling and forecasting between 2000 and 2007. Goh and Law (2011) consider 155 papers published between 1995 and 2009. Among the techniques they examined were econometric methods, time-series models, and other emerging methods including artificial intelligence (AI) (i.e. machine learning) techniques. They noted that none of the considered methods consistently outperform the others. Recently Athanasopoulos, Hyndman, Song, and Wu (2011) have compared univariate and multivariate time series approaches, and econometric models on a large number of tourism time series. They have found that pure time series based approaches are better than the models with explanatory variables. Ahmed, Atiya, El Gayar, and El-Shishiny (2010) have compared several machine learning methods on the M3 competition data set that consists of 3003 time series with different characteristics. They have identified multilayer perceptron type neural network (NN) and the Gaussian process regression (GPR) as the top two methods. Hong, Dong, Chen, and Wei (2011) considered the Support Vector Regression (SVR) model with chaotic genetic algorithm (CGA) (SVRCGA) where the parameters the SVR are determined using CGA. They compared SVRCGA with SVMG models and ARIMA models on the Barbados annual tourist arrivals data and found SVRCGA to be superior. In a recent work by Wu, Law, and Xu (2012)

* Tel.: +90 212 440 00 00x11669.
E-mail address: akin01@istanbul.edu.tr.

http://dx.doi.org/10.1016/j.tourman.2014.11.004

GPR was compared with ARMA and two SVR methods, namely ν-SVM, and g-SVM in tourism demand forecasting in Hong Kong. Based on their findings they recommend using GPR. In an interesting work, Chen (2011) proposed combining linear and nonlinear models to improve forecasting accuracy. Among the models combined with ARMA models are SVR and NN models. He shows that such combined models can achieve considerably better predictive performances.

In all of the above-mentioned studies none of the approaches emerges as the best one in all cases. The reason for this observation was attributed to variations of conditions and data type (Li, Song, & Witt, 2005). There is, however, a more fundamental reason for this observation. According to the **No Free Lunch Theorems** (Wolpert, 2001), given any two learning algorithms, there can be any number of cases where one algorithm may be better than the other and vice versa. For this reason, instead of trying to identify “the best” algorithm, it may be more appropriate to try to identify the factors that may affect the performance of the modeling approaches under consideration, and then make a selection for a given time series.

This study proposes a novel approach for model selection for tourism time series. We consider monthly tourist arrival data between January 2001 and December 2011 published by the Turkish Ministry of Tourism (Republic of Turkey Ministry of Culture and Tourism, General Directorate of Investment and Enterprises, 2012). Using time series representing tourist arrivals to Turkey from different countries, namely, Germany, Russian Federation, United Kingdom, Bulgaria, Netherlands, Iran, France, Georgia, U.S.A., and Italy, we compare the forecasting accuracies of one classical, namely SARIMA, and ν-SVR and NN, two machine learning algorithms with universal approximation property, i.e. they can represent any arbitrary nonlinear function provided that they have appropriate parameters (Hammer & Gersmann, 2003; Hornik, 1991). Based on this comparison, we determine the set of heuristic rules for selecting the most appropriate approach given a tourism time series. This is the main contribution of this study. We optimize the parameters of ν-SVR using particle swarm optimization.

The organization of the rest of the paper is as follows. In Section 2, we give a brief introduction to autoregressive integrated moving average modeling, structural time series modeling, and support vector machine modeling, and particle swarm optimization. The proposed approach is given in Section 3. The data sets used in this study, and the identified BSM, SARIMA, SVR, and NN models are given in Section 4. The generated decision tree and the constructed rule set for model selection are given in Section 5. We present our conclusions in Section 6.

2. Background

2.1. Autoregressive integrated moving average modeling

The autoregressive integrated moving average (ARIMA) model is a generalization of the autoregressive moving average (ARMA) model (Whittle, 1951) and can be written as

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t, \quad (1)$$

where y'_t is the differenced series. The predictors on the right hand side include both the lagged values of y_t and the lagged errors e_t. This called an $ARIMA(p,d,q)$ model, where p is the order of the autoregressive part, d is the degree of first differencing involved, and q is the order of the moving average part.

The seasonal ARIMA (SARIMA) (Hyndman & Athanasopoulos, 2012) model is denoted as $ARIMA(p,d,q)(P,D,Q)_m$, where p and P represent the orders of the non-seasonal and seasonal autoregressive parameters, respectively; q and Q represent the orders of the non-seasonal and seasonal moving average parameters, respectively, and d and D represent the numbers of regular and seasonal differences required, respectively and m is the number of periods per season. The terms of the seasonal part of the model are similar to the non-seasonal components of the model as in Equation (1), but they contain backshifts of the seasonal period. The additional seasonal terms are multiplied with the non-seasonal terms to obtain the SARIMA model as follows

$$\Psi_p(B)\Phi_P(B^m)(1-B)^d(1-B^m)^D y_t = \theta_q(B)\Theta_Q(B^m)e_t \quad (2)$$

where B is the backward shift operator.

2.2. Structural time series models

Structural time series models (Harvey, 1989) are linear Gaussian state-space models for time series based on a decomposition of the series into a number of components. They are specified by a set of error variances, some of which may be zero.

The models considered are the following:

- The **local level model** is the simplest model with an underlying level m_t which evolves by

$$m_{t+1} = m_t + \xi_t, \quad \xi_t \sim N\left(0, \sigma_\xi^2\right) \quad (3)$$

The observations are given by

$$y_t = m_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right) \quad (4)$$

There are two parameters, σ_ξ^2 and σ_ε^2. It is an $ARIMA$(0,1,1) model, but with restrictions on the parameter set.

- The **local linear trend model** has the same measurement equation, but additionally it has a time-varying slope in the dynamics for m_t, given by

$$\begin{aligned} m_{t+1} &= m_t + n_t + \xi_t, \quad \xi_t \sim N\left(0, \sigma_\xi^2\right) \\ n_{t+1} &= n_t + \zeta_t, \quad \zeta_t \sim N\left(0, \sigma_\zeta^2\right) \end{aligned} \quad (5)$$

with three variance parameters. When $\sigma_\zeta^2 = 0$ Equation (5) reduces to the local level model, and $\sigma_\xi^2 = 0$ implies a smooth trend. This is a restricted $ARIMA$(0,2,2) model.

- The **basic structural model** (BSM) is a local trend model with an additional seasonal component. The measurement equation is

$$y_t = m_t + s_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right) \quad (6)$$

where s_t is a seasonal component with dynamics

$$s_{t+1} = -s_t - \ldots - s_{t-s+2} + w_t, \quad w_t \sim N\left(0, \sigma_w^2\right) \quad (7)$$

The boundary case $\sigma_w^2 = 0$ corresponds to an arbitrary deterministic seasonal pattern, sometimes known as the “dummy variable” version of the BSM.

Typical components of a BSM model can be seen in Fig. 1.

2.3. Support vector machine models

Support Vector Machines (SVM) were introduced initially for addressing classification problems (Vapnik, 1995). Support Vector