International Journal of Statistics and Applied Mathematics 2018; 3(2): 338-342

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2018; 3(2): 338-342

www.mathsjournal.com
Received: 24-01-2018
Accepted: 26-02-2018

Shalini Chandra
Department of Mathematics & Statistics, Banasthali University, Rajasthan, 304022, India

Kriti Kumari
Department of Mathematics & Statistics, Banasthali University, Rajasthan, 304022, India

Correspondence
Shalini Chandra
Department of Mathematics & Statistics, Banasthali University, Rajasthan, 304022, India

Forecasting foreign tourist arrivals in India using time series models

Shalini Chandra and Kriti Kumari

Abstract
This study aims to compare various time series models to forecast monthly foreign tourist arrivals to India. The models which are considered here include Naive I & Naive II, seasonal autoregressive integrated moving average (SARIMA) and Grey models. The forecasting performance of these models has been compared under mean absolute percentage error (MAPE), U-statistic and turning point analysis (TPA) criteria. Empirical findings show that Naive I gives better forecast of foreign tourist arrivals to India relative to other time series models under MAPE and U-statistic criteria. In addition, SARIMA is found to be better model as compared to other models according to TPA criterion.

Keywords: Foreign tourist arrivals, Naive, SARIMA, Grey

1. Introduction

Since accurate forecasting of tourism demand is very important for the government's policy-making and development of infrastructure. Hence, there are a number of time series models such as Naive I & Naive II, autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) for forecasting foreign tourist arrivals (see Martin and Witt (1989) and Chu (1998)) [15, 3]. Over the past three decades, SARIMA is one of the most popular time series model in the field of tourism demand forecasting. For example, Lim and McAleer (1999, 2002) [13] studied Box-Jenkins ARIMA model to forecast Malaysian tourist arrivals and foreign tourist arrivals in Australia, respectively, while Goh and Law (2002) [7] applied the SARIMA model in forecasting tourism using 10 country wise foreign tourist arrivals in Hong Kong. Huang and Min (2002) [9] fitted SARIMA model to forecast the visitor arrivals in Taiwan. Also, Kim and Moosa (2005) [10] used SARIMA with other competing models to compare the performance of direct and indirect forecasting of international tourist flows to Australia. Chen *et al*. (2009) [4] compared the accuracy of SARIMA model with Holt-Winters and Grey model to forecast inbound air travel arrivals for Taiwan. Moreover, Nguyen *et al*. (2013) [16] employed SARIMA and Grey model to study the forecasting of foreign tourist arrivals in Vietnam. The results show that the performance of SARIMA model was effectively superior to other time series models.

Further, SARIMA model is more popular for tourism demand forecasting. But it has some limitations that it requires certain statistical properties. However, there exist some quantitative models to forecast time series data which are free from any statistical assumptions such as Grey model developed by Deng (1982) [5]. Grey model has been successfully applied in tourist arrivals forecasting. Yu and Schwartz (2006) [20] compared the forecasting performance of the simple models with Grey model to forecast annual U.S. tourist arrivals data. They found that SARIMA model is the best one for forecasting tourism demand as compared to other forecasting models. Recently, Xu *et al*. (2016) [19] studied international tourism market of China using Markov chain Grey model and compared it with the conventional Grey model under various comparison criteria. The experimental results indicate that the Markov chain Grey model is considerably more accurate than the conventional Grey models.

To the best of our knowledge, researchers in India have not paid much attention to study the forecasting performance of FTAs to India using various quantitative models, though tourism industry in India has shown a significant growth in past few years. In this paper, an attempt has

been made to forecast monthly foreign tourist arrivals in India using time series models. The data were obtained from the site (*http://www.indiastat.com*), Govt of India for the period of Jan. 2003 to Dec. 2016. In addition, various time series models will be fitted including Naive I & Naive II, SARIMA and Grey models. The forecasting performances of various models have also been compared under mean absolute percentage error (MAPE), U-statistic and turning point analysis (TPA) criteria.
The plan of the paper is as follows. Section 2 discusses the time series models related to the study of foreign tourist arrivals to India. Section 3 discusses the different comparison criteria. Section 4 describes the data. Section 5 explains results of time series models obtained from empirical study. Section 6 presents the comparison of forecasting performance of various models under the MAPE, U-statistic and TPA criteria. The concluding remarks are given in the last section.

2. Time series models
Time series models have been commonly used to forecast international tourist arrivals for non-stationary time series data. It includes Naive I & Naive II models, SARIMA and Grey models. A brief overview of these models is given in the following subsections.

2.1 Naive model
Two different Naive methods are very easy and simple for forecasting time series data. The first method is Naive I and second method is Naive II. In Naive I method, each forecast is generated by using the previous value of time series data. In Naive II method, forecast of current observation is equal to the previous value and multiplied by the growth rate of current observation over previous observation.
Naive I and Naive II models for observed time series, say X_t and forecasted time series, X_t at time t are given below.

Naïve I: $\widehat{X_t} = X_t$

Naïve II: $\widehat{X_t} = X_{t-1}\left[1 + \frac{X_{t-1} - X_{t-2}}{X_{t-2}}\right]$

2.2 SARIMA model
SARIMA model is the combination of non-seasonal and seasonal autoregressive (AR) and moving average (MA) models. SARIMA model is usually represented in multiplicative form such as SARIMA (p,d,q)(P,D,Q)$_s$, where (p, P) and (q, Q) indicate the seasonal and non-seasonal order of AR and MA, respectively, with seasonal length s, non-seasonal difference (d) and seasonal difference (D).
SARIMA model for a time series, say X_t(t = 1, 2....T), is given by

$$\varphi_p(B)\emptyset_P(B^s)\Delta^d\Delta_s^D X^t = \theta_q(B)\vartheta_Q(B^s)a_t$$

where, $\varphi_p(B) = 1 - \varphi_1(B)$ - … - $\varphi_p(B^p)$, $\emptyset_P(B^s)$ = 1-$\emptyset_1(B^s)$-…-$\emptyset_P(B^{Ps})$, $\theta_q(B)$ = 1+ $\theta_1(B) + \cdots + \theta_q(B^q)$ and $\vartheta_Q(B^s)$ = 1+ $\vartheta_1(B^s)$ +…+ $\vartheta_Q(B^{Qs})$ are non-seasonal and seasonal AR and MA models, respectively, B is the backshift operator, $\Delta^d\Delta_s^D X_t = (1-B)(1-B^s)X_t$, $\varphi_p < 1, \emptyset_P < 1, \theta_q < 1, \vartheta_Q < 1$ and a_t. ~ white noise (0,σ^2).

2.3 Grey model
So far we noticed that the SARIMA model require certain assumption on distribution owing to the fact that they have certain statistical limitations. However, there exist some quantitative models to forecast time series data which are free from any distributional assumption of data such as Grey model. Grey model is one of the commonly used time series model. This model is a forecasting model consists of a first order differential equation, which can be evaluated as follows:

Suppose $X^{(0)} = \{X^{(0)}(1), X^{(0)}(2), X^{(0)}(3), \ldots\ldots, X^{(0)}(n)\}$ is original data series. From the original data sequence, $X^{(0)}$ the first-order accumulated generating operation (AGO) sequence, say $X^{(1)}$ is generated which is given as .

$$X^{(1)} = \{X^{(1)}(1), X^{(1)}(2), X^{(1)}(3), \ldots, X^{(1)}(n)\}$$

The Grey model of first order linear differential equation is written as

$$\frac{dX^{(1)}}{dt} + a * X^{(1)} = b.$$

Where, the parameters a and b can be estimated using the ordinary least squares (OLS) method, as follows:

$$\hat{a} = \binom{a}{b}$$

$$= (B^T B)^{-1} B^T Y$$

Where,

$$Y = \begin{pmatrix} X^{(0)}(2) \\ X^{(0)}(3) \\ . \\ . \\ . \\ X^{(0)}(n) \end{pmatrix} \text{ and } B = \begin{pmatrix} -\frac{1}{2}[X^{(1)}(1) + X(2)]\ 1 \\ -\frac{1}{2}[X(2) + X^{(1)}(3)]\ 1 \\ . \\ . \\ . \\ . \\ -\frac{1}{2}\left[X^{(1)}(n-1) + X^{(1)}(n)\right]\ 1 \end{pmatrix}$$

Finally, the forecasting equation of Grey model for k periods is given by

$$\widehat{X^0(k)} = X^{(1)}(k) - X^{(1)}(k-1)X^{(1)}$$

$$\widehat{X^0(k)} = (1 - e^a)\left(X^{(0)}(1) - \frac{b}{a}\right)e^{-a(k-1)}, \text{ k=1,2,…}$$

3. Measurements of forecasting accuracy
The performance of all such forecasting models for monthly time series data has been evaluated using mean absolute percentage error (MAPE), U-statistic and turning point analysis (TPA) as the criteria of measuring forecasting accuracy, which are discussed below.

3.1 Mean absolute percentage error (MAPE)
MAPE is a relative measure of errors in prediction. It usually expresses accuracy as a percentage, and is defined by the following formula

$$\text{MAPE} = \frac{\sum_{t=1}^{n}(|X_t - \widehat{X_t}|)/X_t}{n} \times 100$$

where n is sample size, X_t is actual value of the time series and is the forecast in the t_{th} month.
Lewis (1982) demonstrates that the value of MAPE being less than 10% indicates high accuracy of forecasting. Moreover,

when it lies between 10-20% forecasting is good, 20-50% is reasonable and more than 50% denotes inaccuracy in forecasting.

3.2 U-statistic

It is also a relative measure given by Theil (1961). U statistic is calculated by using the following formula

$$U = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t - \widehat{X_t})^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t)^2} + \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\widehat{X_t})^2}}$$

In addition, the value of U lies in the range 0 to 1. U = 0 indicates that the forecasting is good, and, if U is close to 1, it reflects the inaccuracy of forecasting technique. It has many advantages except one that it sometimes becomes more difficult to interpret than other relative measures (see Chu (1998)) [3].

3.3 Turning point analysis (TPA)

Turning point analysis is a graphical way of examining forecasting performance of the models (see Bails and Peppers (1993) for details). TPA predicts a change in the trend of actual and forecasted time series. Changes in trend can be divided into upturns and downturns. If both the changes, actual change ($\Delta X_t = X_t - X_{t-1}$) and forecasted change ($\Delta\widehat{X_t} = \widehat{X_t} - \widehat{X_{t-1}}$) are in different directions, it is said that turning point error (TPE) occurs. The number of turning point errors (TPE) being large imply inaccuracy of forecasting model. Moreover, if the changes in actual trend and forecasted trend are in same direction, it is said that corrected direction occurred. Also, TPA represents the accuracy of time series models in terms of the percentage of corrected direction instances by calculating D-statistic (D_{stat}). D_{stat} is evaluated by the ratio of corrected direction instances and the total number of classified instances (Xie *et al*. 2014), which is as follows:

$$D_{stat} = \sum_{t=1}^{n}\frac{d_t}{n} \times 100$$

where,

$$dt = \begin{cases} 1, (X_t - X_{t-1})(\widehat{X_t} - \widehat{X_{t-1}}) > 0 \\ 0, \text{otherwise} \end{cases}$$

where d_t is the corrected direction in the t_{th} month. Note that MAPE and U-statistic are the measures of the deviation between actual and forecasted values. Therefore, the forecasting performance is better when the values of these measures are lesser. Moreover, D_{stat} is the measures of corrected direction between actual and forecasted values. Therefore, the higher D_{stat} denote high accuracy of time series models.

3.4 Posterior variance ratio (PVR)

The posterior-variance test is the ratio between root mean square of the variance for residuals and root mean square of the variance for the actual data, defined by

$$C = \frac{S_2}{S_1}$$

where,

S_1 = standard deviation for actual series

S_2 = standard deviation for residuals series.

According to Deng (1986) [6], the value of posterior variance ratio (C) <0.35 indicates high accuracy of Grey model. Moreover, when it is <0.5 model is called qualified, and when <0.65 model is said to be just qualified. Also, model is said to be unqualified when C≥ 0.65.

4. Data

This study applied the monthly time series data of foreign tourist arrivals (FTAs) (in numbers) from January 2003 to December 2016 in India for empirical study. Time series data have been downloaded from the official website, http://www.indiastat.com. Also, the series is seasonally adjusted by using ratio to moving average method. The whole time series is divided into two periods (1) 2003:1-2012:12, consists of 120 observations, to estimate the individual models; (2) data of 2013:1-2016:12 are used to generate the out-of-sample forecasts for different models. The R-3.0.3 software is used for the overall empirical analysis.

5. Empirical results

5.1 Naive models

In Naive I method, each forecast is generated by using the previous value of time series data, for instance, the forecast value of FTAs for January 2013 is the arrival figure of December 2012 likewise the forecast for February 2013 is the arrivals figure of January 2013. Similarly, the forecast values of rest of the time points can be obtained for foreign tourist arrivals using Naive I method. In Naive II method, forecast of January, 2013 is equal to the figure obtained in December 2012 multiplied by the growth rate of January 2013 over December 2012. The actual and forecasted results obtained from Naive I & Naive II methods are given in Table 1 from January 2013 to December 2016.

5.2 Estimated SARIMA models

Since, seasonal ARIMA is a more suitable model for non-stationary time series data. Thus, Box-Jenkins method (Box and Jenkins (1976)) based SARIMA model is applied for modelling and forecasting FTAs data. Firstly, the standard HEGY test (Hylleberg *et al*. (1990)) is applied for the testing of seasonal and non-seasonal unit root in seasonally adjusted data at level and first difference values. The results of the HEGY test along with the p-values are reported in Table 2 at level and first difference values. It shows that the null hypothesis of a unit root in FTAs, cannot be rejected at zero frequency 5% level of significance indicating that the FTAs series is having trend at level, however, there is no significant seasonal component. Further, the non-stationary FTAs series is transformed to stationary series by using first difference (d=1). The results of HEGY test at first difference reveal that the null hypothesis of a unit root in FTAs is to be rejected at 5% level of significance. The orders of seasonal and non-seasonal autoregressive (AR) and moving average (MA) models have been identified by autocorrelation and partial autocorrelation functions. The best model fitted to the series is selected corresponding to minimum Akaike information criterion (AIC). From Table 3 it is noticed that (1, 1, 2) $(1, 0, 2)_{12}$ is the best fitted SARIMA model for seasonally FTAs series according to the minimum AIC. Statistical independence of residuals is examined via Box-Pierce Q-statistic (Ljung and Box (1978)) for all lags and it is found that the null hypothesis of independence of residuals is not rejected. The results of Box-Pierce Q-statistic are presented in Table 4. Hence, SARIMA (1, 1, 2) $(1, 0, 2)_{12}$ could be considered an adequate model for out-of-sample forecasts of FTAs.

Table 1: Forecasting results of Naive models

Months/Year	Actual	Naïve I	Naïve II
Jan-13	579649	562271.4	545559.6212
Feb-13	561234.4	579649	597563.6733
Mar-13	579372.1	561234.4	543404.8049
Apr-13	545787.3	579372.1	598095.9654
May-13	587866.7	545787.3	514149.3297
Jun-13	581359.4	587866.7	633190.36
Jul-13	542986	581359.4	574924.1316
Aug-13	574632.3	542986	507145.4873
Sep-13	592841.4	574632.3	608123.0091
Oct-13	585484.3	592841.4	611627.5148
Nov-13	606551.5	585484.3	578218.5008
Dec-13	613504.2	606551.5	628376.7509
…	…	…	…
Jan-16	679174.6	681563.2	688863.4649
Feb-16	690367.4	679174.6	676794.3711
Mar-16	740148.2	690367.4	701744.6574
Apr-16	725568.4	740148.2	793518.5786
May-16	744949.7	725568.4	711275.8
Jun-16	708624.4	744949.7	764848.7111
Jul-16	789131.9	708624.4	674070.3973
Aug-16	791638	789131.9	878785.9345
Sep-16	803855.4	791638	794152.0588
Oct-16	738102.1	803855.4	816261.3519
Nov-16	736368	738102.1	677727.2505
Dec-16	774365.3	736368	734637.9741

5.3 Estimated results of Grey model

The Grey model is initiated by converting the original FTAs time series data to a Grey differential equation using accumulated generation operation (AGO). The complete process of the Grey model has been conducted using inverse accumulated generation operation (IAGO). Also, the least squares method is used to estimate the parameters for Grey differential equation. The estimated values of the parameters a and b for Grey model are -0.006628205 and 266972.1, respectively. Also, the accuracy of Grey model is evaluated by using PVR criterion. The calculated value of PVR, 0.2202 is <0.35. It indicates that the accuracy of Grey model is high for FTAs data.

6. Forecasting performance of different models

In this section, we present comparisons of forecasting performance of different models using MAPE, U-statistic and TPA criteria. Firstly, TPA has been explored to determine the D-statistic value. TPA reveals that the actual and forecasted changes are in same direction or not. If the actual changes (AC) and forecasted changes (FC) has different direction then TPE occurs. Graphical representation of AC and FC is reported in Figure 1-4 for Naive I & Naive II, SARIMA and Grey models. Figure 1-4 shows the closeness of actual and forecasted foreign tourist arrivals. If the forecast points are very close to the straight line, it is said that the forecast model is the best. Further, MAPE, U-statistic, D-statistic and TPE have been calculated for the comparative analysis of different models. The results of MAPE, U-statistic, D-statistic and TPE are reported in Table 5. Overall, the empirical analysis reveal that time series models have good forecasting performance since MAPE values are less than 10% (Lewis 1982). For Naive I model the coefficient of U is very close to 0. Furthermore, the experimental results reported in Table 5 also support that the forecasting accuracy of Naive I model is better than other time series models in terms of lesser MAPE and U-statistic values. On the other hand, SARIMA model is better than Naive I & Naive II and Grey models according to the higher D-statistic value, 58.33. There are 19 turning point errors. Therefore, SARIMA model forecasting performance is more accurate than other models based on TPA criterion.

Table 2: Results of HEGY test

Roots	Level		First difference	
	Test statistic	p-values	Test statistic	p-values
π_1	-2.329	0.1	-3.65	0.01
π_2	-3.378	0.01	-3.2	0.01
$\pi_{3:4}$	7.893	0.01	6.677	0.01
$\pi 5:6$	8.403	0.01	5.81	0.01
$\pi 7:8$	8.582	0.01	7.505	0.01
$\pi 9:10$	9.016	0.01	7.465	0.01
$\pi 11:12$	6.305	0.01	5.626	0.01

Table 3: Fitted SARIMA models

Models	AIC criterion
(1, 1, 2) (1, 0, 2)$_{12}$	2660.34
(2, 1, 1) (2, 0, 1)$_{12}$	2664.98
(1, 1, 2) (2, 0, 1)$_{12}$	2660.41
(0, 1, 2) (2, 0, 1)$_{12}$	2661.28

Table 4: The results of Box-Pierce statistic for SARIMA model

Lags	1	2	3	4	5	6	7	8	9
Statistic	0.1	0.34	0.71	0.94	5.2	7.91	7.91	8.02	9.5
P-values	0.75	0.84	0.87	0.91	0.38	0.24	0.34	0.43	0.38

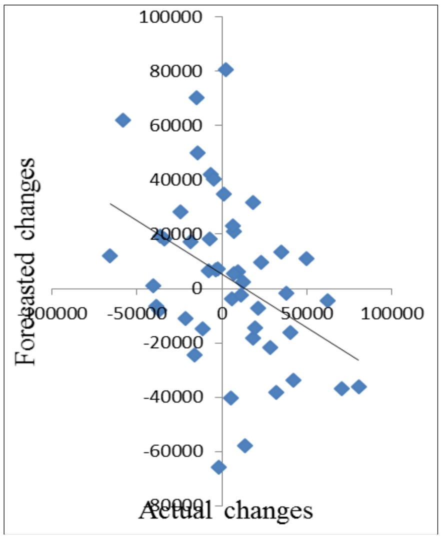

Fig 1: AC and FC for Naive I model

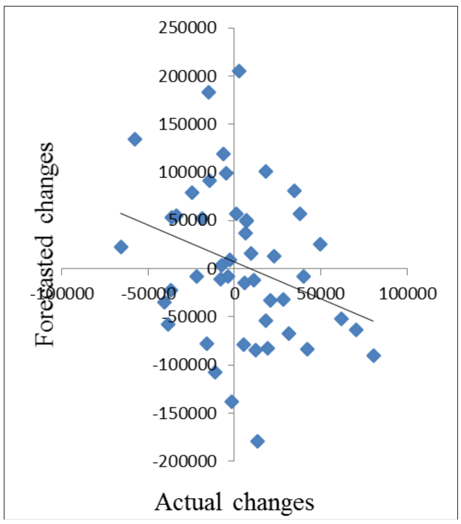

Fig 2: AC and FC for Naive II model.

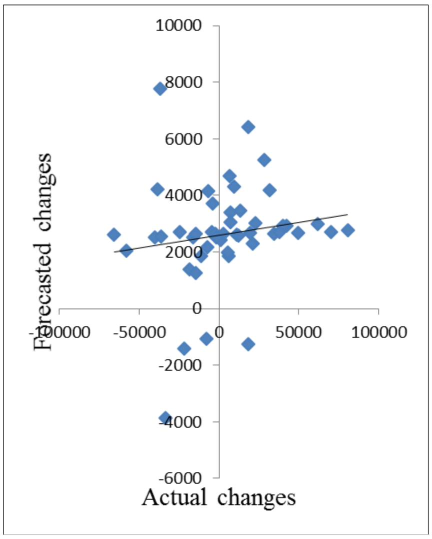

Fig 3: AC and FC for SARIMA model

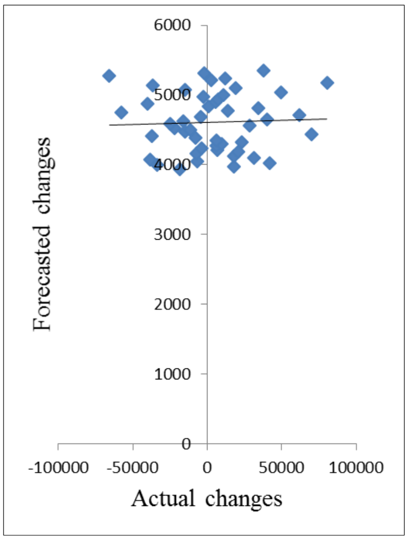

Fig 4: AC and FC for Grey model.

Table 5: Forecast comparison under different criteria

Models/Criteria	Naïve I	Naïve II	SARIMA	Grey
MAPE	3.59	6.27	5.00	5.81
U_{stat}	0.023	0.038	0.035	0.032
D_{stat}	35.41	41.66	58.33	54.16
TPE	27	26	19	21

7. Conclusions

This study evaluated the performance of Naive I & Naive II, seasonal autoregressive integrated moving average (SARIMA) and Grey models for forecasting foreign tourist arrivals (FTAs) in India from January 2003 to December 2016 using mean absolute percentage error (MAPE), U-statistic and turning point analysis (TPA) criteria. Also, the posterior variance ratio criterion is used to check the accuracy of Grey model. Empirical findings show that the Naive I model outperform the other models when compared under the MAPE and U-statistic criteria. Additionally, SARIMA is found to be better model as compare to Naive I & Naive II and Grey models according to TPA criterion.

8. References

1. Box GEP, Jenkins GM. Time series analysis. Forecasting and Control, 1976.
2. Bails DG, Peppers LC. Business fluctuations-forecasting techniques and applications. New Jersey: Prentice-Hall International, 1993.
3. Chu FL. Forecasting tourism demand in Asian Pacific countries. Annual of Tourism Research. 1998; 25(3):597-615.
4. Chen CF, Chang YH, Chang YW. Seasonal ARIMA forecasting of inbound air travel arrivals to Taiwan, Transportmetrica, 2009; 5(2):125-140.
5. Deng J. Control problems of Grey systems. Systems and Control Letters, 1982; 1(1):288-294.
6. Deng J. Grey prediction and Grey decision making. Huazhong University of Science and Technology, Wuhan, China (in Chinese), 1986.
7. Goh C, Law R. Modeling and forecasting tourism demand for arrivals with stochastic non-stationary seasonality and intervention. Tourism Management. 2002; 23(5):499-510.
8. Hylleberg S, Engle RF, Granger CWJ, Yoo BS. Seasonal integration and cointegration. Journal of Econometrics. 1990; 44:215-238.
9. Huang JH, Min JCH. Earthquake devastation and recovery in tourism: the Taiwan case. Tourism Management. 2002; 23(2):145-154.
10. Kim JH, Moosa IA. Forecasting international tourist flows to Australia: a comparison between the direct and indirect methods. Tourism Management. 2005; 26(1):69-78.
11. Ljung GM, Box GEP. On a measure of the lack of fit in time series models. Biometrica. 1978; 65:297-303.
12. Lewis CD. International and business forecasting methods. London: Butterworths, 1982.
13. Lim C, McAleer M. A seasonal analysis of Malaysian tourist arrivals to Australia. Mathematics and Computers in Simulation. 1999; 48(6):573-583.
14. Lim C, McAleer M. Time series forecasts of international travel demand for Australia. Tourism Management. 2002; 23(4):389-396.
15. Martin A, Witt F. Forecasting tourism demand: A comparison of the accuracy of several quantitative methods. International Journal of Forecasting. 1989; 5:1-13.
16. Nguyen LT, Shu HM, Huang FY, Hsu MB. Accurate forecasting models in predicting the inbound tourism demand in Vietnam. Journal of Statistics and Management Systems. 2013; 16(1):25-43.
17. Theil H. Economic forecasts and policy (2nd ed.). Amsterdam: North Holland, 1961.
18. Xie G, Wang S, Lai KK. Short-term forecasting of air passenger by using hybrid seasonal decomposition and least squares support vector regression approaches. Journal of Air Transport Management. 2014; 37:20-26.
19. Xu S, Wangshu S, Jianzhou W, Yixin Z, Yining G. Using a Grey- Markov model optimized by Cuckoo search algorithm to forecast the annual foreign tourist arrivals to China. Tourism Management. 2016; 52:369-379.
20. Yu G, Schwartz Z. Forecasting short time series tourism demand with artificial intelligence models. Journal of Travel Research. 2006; 45:194-203.