Computers, Materials & Continua
DOI:10.32604/cmc.2022.022309
Article

Tech Science Press

A New Hybrid SARFIMA-ANN Model for Tourism Forecasting

Tanzila Saba[1], Mirza Naveed Shahzad[2,*], Sonia Iqbal[2,3], Amjad Rehman[1] and Ibrahim Abunadi[1]

[1]Artificial Intelligence & Data Analytics Research Lab (AIDA), CCIS Prince Sultan University, Riyadh, 11586, Saudi Arabia
[2]Department of Statistics, University of Gujrat, Gujrat, 50700, Pakistan
[3]Department of Statistics, University of Sialkot, Sialkot, 51310, Pakistan
*Corresponding Author: Mirza Naveed Shahzad. Email: nvd.shzd@uog.edu.pk

Received: 03 August 2021; Accepted: 29 October 2021

Abstract: Many countries developed and increased greenery in their country sights to attract international tourists. This planning is now significantly contributing to their economy. The next task is to facilitate the tourists by sufficient arrangements and providing a green and clean environment; it is only possible if an upcoming number of tourists' arrivals are accurately predicted. But accurate prediction is not easy as empirical evidence shows that the tourists' arrival data often contains linear, nonlinear, and seasonal patterns. The traditional model, like the seasonal autoregressive fractional integrated moving average (SARFIMA), handles seasonal trends with seasonality. In contrast, the artificial neural network (ANN) model deals better with nonlinear time series. To get a better forecasting result, this study combines the merits of the SARFIMA and the ANN models and the purpose of the hybrid SARFIMA-ANN model. Then, we have used the proposed model to predict the tourists' arrival in New Zealand, Australia, and London. Empirical results showed that the proposed hybrid model outperforms in predicting tourists' arrival compared to the traditional SARFIMA and ANN models. Moreover, these results can be generalized to predict tourists' arrival in any country or region with a complicated data pattern.

Keywords: SARFIMA; hybrid model; tourists' arrival forecasting; ANN

1 Introduction

The rapid development in the tourism industry in the last 30 years has contributed to many countries' economies. According to the World Travel & Tourism Council (WTTC), 2019, tremendous evolution was observed in the tourism industry. It has created 313 million jobs and created prosperity in industries related to tourism and has increased taxations. As a result, the tourism industry had generated 10.3% of the global GDP. Consequently, every country is paying much attention to growing tourists' arrival in their territory. Therefore, tourism policy makers and business practitioners are interested in knowing an accurate forecast of tourist volume for properly distributing resources and formulating pricing strategies [1].

One well-known approach to model tourists' arrival data is to use time series models. For this purpose, models based on the Box-Jenkins methodology have been extensively used in the past several decades. For example, the autoregressive integrated moving average (ARIMA) and its sub-models are well-recognized to predict future observations based on linearly correlated past observation and a white noise error term. However, these models work efficiently only for short memory time series processes [2]. On the other hand, most time-series data, including financial data, stock-exchange data, and numbers of tourists' arrivals, have long memory characteristics. For such types of time series data, Granger [3] introduced the autoregressive fractional integrated moving average (ARFIMA) model that speculates the ARIMA by allowing non-integer values of the differencing parameter (d). Recall, Robinson [4], that the long memory in time series is termed autocorrelation at long lags, and the strength of this long memory is measured by d of the ARFIMA model.

The ARFIMA model has been applied in several studies for accurate forecasting. For example, Doornik et al. [5] forecasted US and UK inflation rates using the ARFIMA model and showed its superiority over the ARIMA model. Chu [6] has found better forecasting about tourists' demand in Hong Kong, Japan, and Korea by ARFIMA model than ARIMA. Moreover, Cheung [7] used the ARFIMA model for predicting the foreign exchange rates. Along with the ARFIMA model, Prass et al. [8] also applied the seasonal autoregressive fractional integrated moving average (SARFIMA) model to predict the mean monthly water level Paraguay River in Brazil. The SARFIMA model has also been used by Mostafaei et al. [9] to predict Iran's oil supply. In recent years, Peng et al. [10] developed the new hybrid random forest-LSTM model to forecast tourist arrival data and justified by the Beijing city and Jiuzhaigou valley data that this hybrid approach outperforms. Waciko et al. [11] used the Thief-MLP hybrid approach to forecast short-term tourists' arrival to Bali-Indonesia. Wu et al. [12] forecasted daily tourist arrival to Macau SAR, China with a hybrid SARIMS–LSTM approach and obtained the good fitted results.

It is important to note that all Box-Jenkins models require the linearity of the process under study (see, e.g., [13]). This means that future observations should have a linear relationship with present and past observations. Therefore, ARFIMA and SARFIMA models become inappropriate when time series are generated from nonlinear processes. Nonlinear models such as the Artificial Neural network (ANN) have been used in several studies to overcome this problem. For instance, Sarvareddy et al. [14] studied the characteristics of ANN and showed its forecasting power over traditional time series models. Zhang et al. [15] studied ANN forecasting for seasonal and trend time series and showed that the ANN has nonlinear architecture and performs well for linear time series. In several studies, ANN has been constantly compared with the traditional Box-Jenkins model and long memory models, and its better performance is justified. Prybutok et al. [16] observed that the ANN analysis performs better for time series datasets, specifically in the presence of nonlinearity.

Note that there is no definitive proper measure to check whether the process generated from time series is linear or nonlinear. Therefore, it would be difficult to decide in advance which model should be used. Furthermore, there is no standard model that is appropriate for both linear and nonlinear time series data. Therefore, by combining the characteristics of two or more models, the accuracy can be increased. Different hybrid models have been developed following this idea, particularly by combining linear autoregressive models and ANN models. The beauty of these types of hybrid models is that they can capture the characteristics of both linearity and non-linearity in the time series [17]. This idea was first used by Zhang [18] to introduce a hybrid methodology that combines both ARIMA and ANN models. Following [18], a hybrid of ARIMA models and support vector machines (SVMs) were used by Pai et al. [19] to forecast stock prices. Chen et al. [20] examined the forecasting accuracy of the hybrid of seasonal ARIMA (SARIMA) and SVM models for Taiwan's machinery industry production.

Chaâbane [21] found a hybrid of ARFIMA and ANN model more efficient over individual ARFIMA and ANN to predict electricity price.

Moreover, Chaâbane [22] found a better performance of hybrid ARFIMA–least-squares SVM than the competing models while predicting electricity spot prices. In general, practitioners rely on two methods to forecast tourist volume. One is time series analysis methods [23] and the other is artificial intelligence methods [24–26]. However, many recent studies have proved that the combination of the aforementioned methods leads to better forecasts [27–29].

In the present study, our main interest is to introduce a new hybrid model for better and accurate prediction using the complicated time series data, the new hybrid model is proposed by hybridizing the SARFIMA and ANN model using the Zhang [18] approach. The second goal is to evaluate the predictability of the developed model on the different natured tourists' arrival datasets, for this, the datasets of the three countries were retrieved and better results have been obtained. Finally, the performance of the suggested model is verified using out-of-sample data as real-time analysis, and the results predicted by the model matched with the real tourists' arrival values.

The rest of the paper is organized as follows. An overview of ARFIMA, SARFIMA, and ANN models is given in Sections 2, 3, and 4, respectively. The proposed hybrid SARFIMA-ANN model is presented in Section 5. Section 6 reports on the proposed hybrid model's empirical results using three real datasets on tourists' arrival and discussion. The concluding remarks are given in Section 7.

2 ARFIMA (p, d, q) Model

When there is a long memory presence in time series data, the frequently used model is the ARFIMA model, introduced by Granger et al. [30] and its properties were further investigated by Baillie [31]. Let us assume that $\{y_t\}_{t\in W}$ is a stationary process with the white noise error term ε_t with zero mean and variance $\sigma^2 > 0$. Let B be the backshift operator defined by $B^k(y_t) := y_{t-k}$. Further, assume that $\{y_t\}_{t\in W}$ is a linear process satisfying the following relationship

$$\psi(B)(1-B)^d y_t = \vartheta(B)\varepsilon_t,\ t \in W, \tag{1}$$

where $d \in (-0.5, 0.5)$, $\psi(\cdot)$ and $\vartheta(\cdot)$ are the following polynomials of degree p and q respectively, that is

$$\psi(B) := 1 - \psi_1(B) - \psi_2(B^2)\cdots\psi_p(B^p),$$

$$\vartheta(B) := 1 - \vartheta_1(B) - \vartheta_2(B^2)\cdots - \vartheta_q(B^q),$$

where $\psi_k, 1 \le k \le p$ and $\vartheta_l,\ 1 \le l \le q$ are real constant terms. Then the series $\{y_t\}_{t\in W}$ follows a fractional differencing model ARFIMA (p, d, q). Note that the ARFIMA (p, d, q) model is an improved form of the traditional ARIMA (p, d, q) model due to considering fractional differencing parameter d. If $d \in (-0.5, 0.5)$, the series $\{y_t\}_{t\in W}$ satisfies the property of stationary and inevitability and indicates a long memory process. On the other hand, $d = 0$ indicates a short memory, whereas $d \in (-0.5, 0)$ tells that the process has an intermediate memory level.

3 SARFIMA $(p, d, q)(P, D, Q)s$ Model

There are several situations in which time-series data have a long memory and exhibit periodic patterns. The appropriate model for such time-series data is the SARFIMA $(p, d, q)(P, D, Q)s$, an extended form of the ARFIMA process [32]. The following are propositions that need to be satisfied to consider SARFIMA as an appropriate model.

Proposition 1: Let $\{y_t\}_{t\in W}$ be the stochastic stationary process following spectral density function $f_y(.)$. Suppose that there is a real quantity $t \in (0, 1)$, a constant C_f and $K \in [0, \pi]$ (with one or a finite number of occurrences) such that $f_y(\omega) \sim C_f|\omega - K|^{-t}$ When $\omega \to$ G. Then $\{y_t\}_{t\in W}$ follow a long memory process.

When t $\in (-1, 0)$, then the process y_t is said to be an intermediate memory process [33].

Proposition 2: Let $\{y_t\}_{t\in W}$ be a seasonal stationary process with

$$\psi(B)\zeta(B^s)(1 - B)^d y_t = \vartheta(B)\varrho(B^s)\varepsilon_t \tag{2}$$

where ε_t is a white noise process with zero mean and variance $\sigma^2 > 0$, s is the seasonal period, $\Delta_s^D = (1 - \mathrm{B}^S)^D$ is the seasonal difference operator, $\psi(.)$, $\zeta(.)$, $\vartheta(.)$ and $\varrho(.)$ are the polynomials of degrees p, q, P, and Q respectively defined by

$$\psi(B) = \sum_{k=0}^{p}(-\psi_k)\mathrm{B}^k \qquad \vartheta(B) = \sum_{h=0}^{Q}(-\vartheta_h)\mathrm{B}^h$$

$$\zeta(B) = \sum_{l=0}^{P}(-\zeta_l)\mathrm{B}^l \qquad \varrho(B) = \sum_{g=0}^{q}(-\varrho_g)\mathrm{B}^g$$

where $\psi_k, 1 \le k \le p, \vartheta_h, 1 \le l \le Q, \zeta_l, 1 \le l \le P,$ and $\varrho_g, 1 \le g \le Q$ are constant integers with $\psi_0 = \zeta_0 = \vartheta_0 = \varrho_0 = -1$.

Proposition 3: Assume that $\{y_t\}_{t\in W}$ is the SARFIMA $(p, d, q)(P, D, Q)s$ process, with mean zero and seasonal period s $\in N$. Suppose $\psi(W)\zeta(W^s) = 0$ and $\vartheta(W)\varrho(W^s) = 0$, have no zero in common. Then, the following axioms are factual:

i. The process $\{y_t\}_{t\in W}$ follows the stationary process when $D < 0.5$, $(d + D) < 0.5$, and $\psi(W)\zeta(W^s) \neq 0$ for $|\mathrm{W}| \le 1$.
ii. The stationary process $\{y_t\}_{t\in W}$ follows long memory process when $(d + D) \in (0, 0.5)$, $D \in (0, 0.5)$ and $\vartheta(W)\varrho(W^s) \neq 0$ for $|\mathrm{W}| \le 1$.
iii. The stationary process $\{y_t\}_{t\in W}$ follows long memory process when $(d + D) \in (-0.5, 0.5)$, $D \in (-0.5, 0)$ and $\psi(W)\zeta(W^s) \neq 0$ for $|\mathrm{W}| \le 1$.

Based on Katayama's previous work [34], model estimation using SARFIMA requires a few steps. Firstly, identifying the long memory process and finding fractional difference parameter d. Secondly, identifying the model and estimating parameters, and finally, applying some diagnostic checks. After obtaining a suitable model, it can be used for predictive purposes.

4 Artificial Neural Network Based Forecasting

When the restriction of linearity on time series data is relaxed, enormous nonlinear models have been developed for obtaining better forecasts. An artificial neural network (ANN) is one of them. The critical characteristic of ANN over other nonlinear models is its ability to deal with a large class of functions. Moreover, ANN does not require any prior assumption for the estimation process. Instead, its architecture is entirely determined from the characteristic of data. For more details on ANN, we refer the interested readers to [35].

The ANN architecture consists of an input layer, an output layer, and multiple hidden layers depending upon the complexity of data. Information passes through each layer in terms of neurons. For forecasting time series data (y_t), the simplest neural network architecture takes lagged values

$(y_{t-1}, y_{t-2}, y_{t-3}, \ldots., y_{t-p})$ as input. We used the Rabie et al. [36] heuristic/Ad-hoc input selection criteria to select the input nodes for ANN. Then the nonlinear relationship between output y_t and input values $(y_{t-1}, y_{t-2}, y_{t-3}, \ldots., y_{t-p})$ is defined by

$$y_t = \alpha_0 + \sum_{j=1}^{u} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{v} \beta_{ij} y_{t-i}\right) + \varepsilon_t, \tag{3}$$

where ε_t is an error term with zero mean and standard deviation σ_t, α_j and β_{ij} for $i = 0,\ 1,\ 2,\ \ldots,\ v$ and $j = 0,\ 1,\ 2,\ \ldots,\ u$ are the model parameters to be estimated. They are also known as connection weights. The function g is known as the activation function used in hidden layers to transfer the input if a certain threshold is met. One can use different activation functions like hyperbolic, sigmoid, etc. However, the most popular is the sigmoid, which is defined by

$$g(y) = \frac{e^y}{1 + e^y}.$$

The ANN model in (3) can be termed as univariate nonlinear autoregressive (NAR) model, that is:

$$y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, \ldots., y_{t-p}, v) + \varepsilon_t \tag{4}$$

Here $f(.)$ is a function estimated by the network structure and connection weights, and v is a vector of all parameters. The ANN model (3) is quite powerful in terms of its architecture as it can estimate the arbitrary function by increasing the number of hidden nodes v.

Since there is no standard mechanism to determine the appropriate ANN architecture for the given data, multiple experiments can be conducted to choose suitable values for p and q. After selecting u and v, the model is ready for training and to estimate the network parameters.

5 Hybrid Methodology and the Proposed Hybrid Method: SARFIMA-ANN

The tourists' arrival time series data may consist of many components such as linearity, nonlinearity, seasonality, heteroscedasticity, or a non-normal error. One standard approach for forecasting such time-series data dealing with all components does not exist. One may think SARFIMA is a better option for this case. However, SARFIMA cannot deal with complex nonlinear structures. The second choice might be the ANN, which deals well with nonlinear structure but may also provide unsatisfactory results when modeling the linear data [37]. In other words, both SARFIMA and ANN models are successful only in their domains. Zhang [18] introduced hybrid models that can model both linear and nonlinear structures of time series data to overcome this problem.

Following Zhang [18], we propose a hybrid of the SARFIMA and ANN model in this study. In the proposed model, time-series data is composed as a function of linear and nonlinear components. To be more precise, $y_t := f(N_t, L_t)$, where N_t is termed as nonlinear part and L_t is termed as linear part. It is important to note that the hybrid approach is one of the efficient approaches that provide high accuracy rate in forecasts by establishing an additive relationship between a linear and nonlinear component of data, that is,

$$y_t = N_t + L_t. \tag{5}$$

The different methods can estimate the linear and nonlinear parts of (5) to develop the model. The defined methodology used in this work has three steps. In the first step, the linear portion of the time series data is modeled by SARFIMA, considering it follows a long memory process. From the fitted SARFIMA model, the forecasted values $\widehat{L_t}$ are obtained. In the second step, residuals from the

SARFIMA model are generated and, under the assumption that these residuals exhibit nonlinear patterns, an ANN model is trained. To be more precise, the input-output relationship is ANN is estimated by the following relationship,

$$e_t = g(e_{t-1}, e_{t-2}, e_{t-3}, \ldots., e_{t-p}) + \varepsilon_t, \tag{6}$$

where $g(.)$ is a nonlinear regression function determined by the ANN model. This provides us the prediction of the nonlinear part $\widehat{N}_t$. In the last step, predictions of the linear and nonlinear component L_t and N_t, are combined to generate the cumulative prediction, that is,

$$\widehat{y}_t = \widehat{N}_t + \widehat{L}_t$$

The pictorial representation of our proposed hybrid methodology is given in Fig. 1. In addition, the algorithm of hybrid SARFIMA-ANN is presented in Fig. 1.

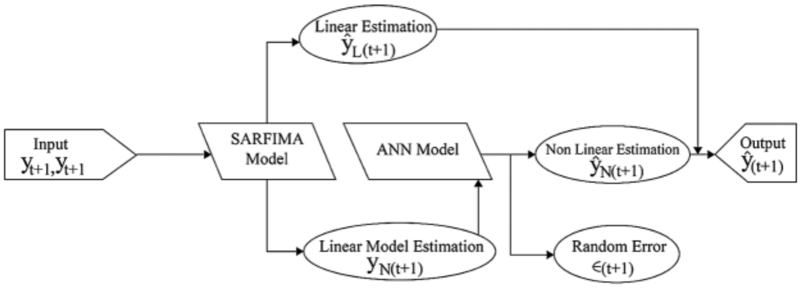


Figure 1: Hybrid SARFIMA-ANN model

6 Application and Empirical Results

To check the performance of the proposed hybrid SARFIMA-ANN model on forecasting tourists' arrival, we consider the three real data sets. As, over the past three decades, tourism has become one of the world's most flourishing industries. International tourists' arrival has conventionally been used as a benchmark to assess any country's security condition and economic development. It significantly impacts GDP, employment rate, import and export, and many public and private sectors. This significant impact attracts the researcher to study the flow of tourists' arrival in a particular country. The number of tourists' arrivals can be considered a time series process due to the consistent change over time and therefore, the prediction model may be applied. Tourists' arrival data get more attention in several studies (see, [38–40]). In this study, the following three tourists' arrival datasets are considered to implement and justify the proposed hybrid model's competency over the other models.

Dataset 1. This dataset is related to the tourism industry, growing gradually in New Zealand due to its amazing natural attraction sites. To forecast tourists' arrival in this country, monthly data from January 2000 to September 2018 is retrieved from www.stats.govt.nz, a sample of 225 observations. The plot of this dataset in Fig. 2 depicts that the considered data is stationary in the mean but has seasonal variation. The autocorrelation function and partial autocorrelation function showed the presence of seasonality at $s = 6$.

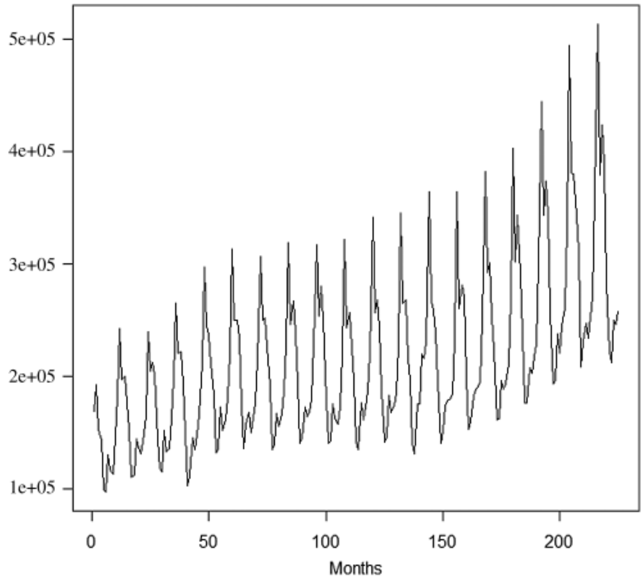


Figure 2: Monthly data of tourists' arrivals in New Zealand from January 2000 to September 2018

Dataset 2. The second dataset contains the monthly number of tourists' arrival in Australia from January 2000 to August 2018, giving 224 observations, which are retrieved from www.abs.gov.au. The data series is regarded as nonlinear and non-Gaussian and suitable to evaluate for analysis. This time-series data has been plotted in Fig. 3, showing seasonality at $s = 6$ with the observed trend. The data is non-stationary and the first difference is taken of the data for further analysis.

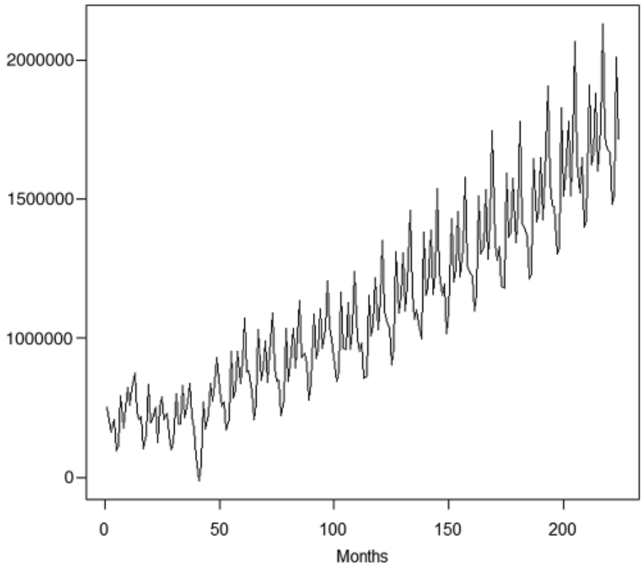


Figure 3: Monthly data of tourists' arrivals in Australia from January 1976 to August 2018

Dataset 3. The proposed model is also applied to the number of tourists' arrival in London, UK. The quarterly data set has 66 observations, corresponding to 2002Q1–2018Q2, taken from www.data.london.gov.uk and plotted in Fig. 4. There exist seasonal fluctuations at $s = 64$ in the series. For modeling and forecasting from this series, the first difference of the data is taken.

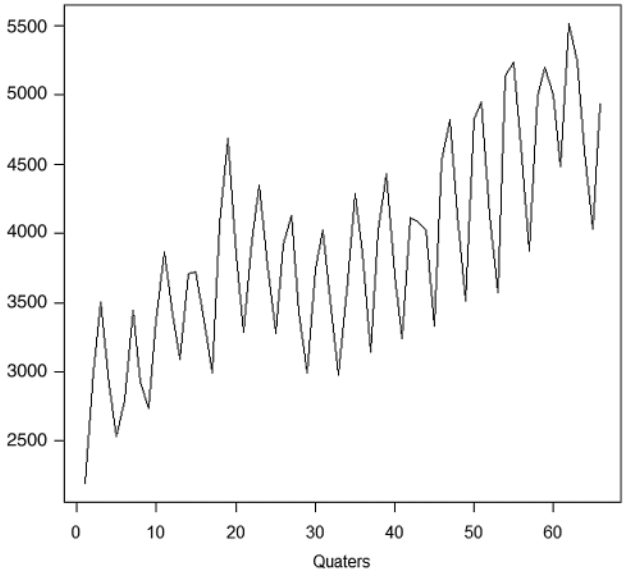


Figure 4: Quarterly data of tourists' arrivals (000s) in London from 2002Q1 to 2018Q2

These three datasets are used in the present study to demonstrate the effectiveness of the proposed hybrid method. Note that these tourists' arrival datasets have seasonal fluctuations due to seasonal changes and this situation requires explaining such fluctuations by some suitable seasonal models. The summary of the datasets is given in Tab. 1. The considered datasets are far from normality as indicated by skewness and kurtosis values and further confirmed by the Jarque-Bera test. Furthermore, Augmented Dickey-Fuller, Philips–Perron, and Kwiatkowski, Phillips, Schmidt, and Shin tests indicate that dataset 1 is stationary whereas the other two datasets are non-stationary. To make the datasets suitable for analysis, datasets are made stationary by taking first differences. Then, long memory parameter d and Hurst parameter H are estimated to ensure that the considered data sets follow long memory processes. For tourists' arrival datasets of New Zealand, Australia, and London, the estimated values for d are 0.4812, 0.3704, and 0.2351, and for H are 0.9812, 0.8715, and 0.7338, respectively. Since $0 < d < 0.5$ and $H > 0.5$, the criteria explained in Proposition 1 is satisfied. This ensures that the considered datasets came from long memory processes.

Table 1: Descriptive statistics for the tourists' arrival datasets in three selected stations

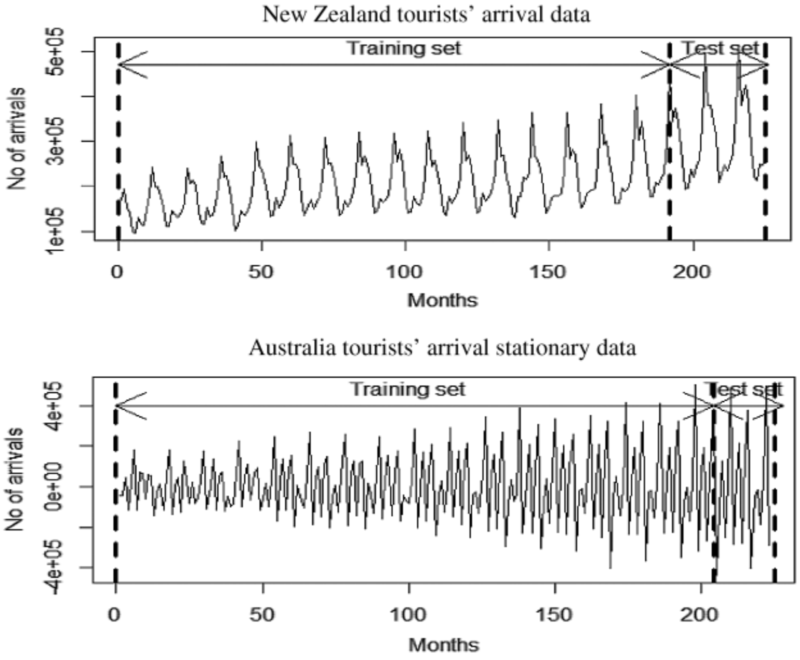
Measurement	Tourists' arrival		
	New Zealand	Australia	London
No. of observations	225	224	66
Mean	216360.7	1109242	42230.77
Standard deviation	74147.21	347781.3	650.1225
Kurtosis	4.566376	0.572944	−1.118681

(Continued)

Table 1: Continued

Measurement	Tourists' arrival		
	New Zealand	Australia	London
Skewness	1.150351	−0.42002	0.418667
Jarque-Bera test	15.82900	13.89100	5.029100
Minimum	97516.00	487900.0	2195.000
Maximum	513349.0	2130700.0	5514.000

In order to apply and explain the performance of SARFIMA, ANN, and hybrid SARFIMA-ANN models, the datasets are partitioned into the training and testing part. To be more precise, New Zealand tourists' arrival data from January 2000 to December 2015 (85.71%) is considered to train models and the set from January 2016 to August 2018 (14.29%) is considered for model testing. Similarly, Australian tourists' arrival data from January 2000 to December 2015 (87.71%) is used for model training, and the rest from January 2016 to August 2018 (14.29%) is considered for model testing. Similarly, in London tourists' arrival data, set from first quarter 2002 to fourth quarter of 2013 (72.73%) is considered a training set and from first quarter 2014 to the second quarter, 2018 (27.27%) is considered a testing set. This partition of datasets is also sketched in Fig. 5.


New Zealand tourists' arrival data

Australia tourists' arrival stationary data

Figure 5: Continued

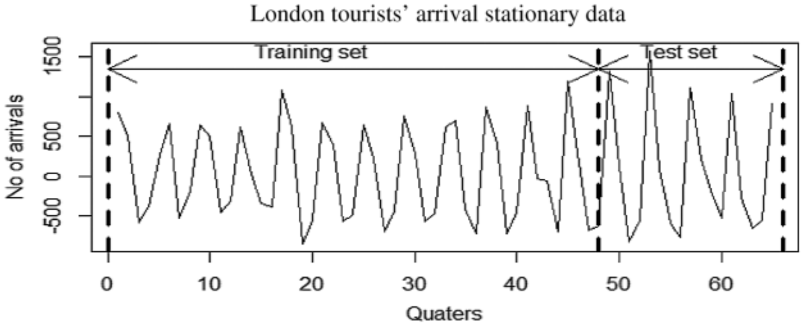


Figure 5: Partition of tourists' arrival into training and testing part

6.1 SARFIMA Model for Tourists' Arrival Datasets

The fitting of SARFIMA $(p, d, q)(P, D, Q)s$ model on training part of tourists' arrival datasets requires suitable values for p, q, P and Q. In the case of New Zealand and Australian tourists' arrival datasets, we take $p = 1, 2, 3, 4; P = 10, 11, 13; q = 1, 2$ and $Q = 7, 8, 9, 13, 14$ on the basis of autocorrelation functions (ACF) and partial autocorrelation function (PACF) values and pattern. Then we take all possible combinations of p, q, P and Q , and fitted 120 SARFIMA models. Analogously, based on ACF and PAF, the suitable values for the London tourist arrival dataset are $p = 1$, $q = 1, 2$, $P = 4, 8$, $Q = 4$, and $s = 4$. For this dataset, we further estimate SARFIMA models with all possible combinations of p, q, P and Q. After fitting all possible models on each dataset, we have selected models for each dataset based on the minimum values of Akaike Information criteria (AIC) and Bayesian Information Criteria (BIC), keeping in view the values of d and D satisfy the stationary and long memory process condition. Tab. 2 presents these four best-fitted models along with their ranks. The first ranked models have been observed as the most parsimonious models among all SARFIMA models for each dataset.

Table 2: SARFIMA model selection for tourists' arrival datasets

Order	σ^2	d	D	AIC	BIC
New Zealand tourists' arrival dataset					
SARFIMA (3, 0, 1)(10, 0, 7)$_6$	1302570	−0.2793	0.22200	4299.06(1)	4386.9
SARFIMA (1, 0, 2)(10, 0, 8)$_6$	1285160	0.1480	−0.04525	4301.52(2)	4386.9
SARFIMA (1, 0, 1)(10, 0, 8)$_6$	1332220	0.4677	0.01888	4306.49(3)	4388.4
SARFIMA (1, 0, 2)(10, 0, 7)$_6$	1355550	0.0913	0.48095	4307.16 (4)	4389.1
Australia tourists' arrival dataset					
SARFIMA (1, 1, 3)(5, 0, 6)$_3$	1134430	0.376	−0.143	4700.07(1)	4764.8
SARFIMA (1, 1, 2)(5, 0, 6)$_3$	1498450	0.158	0.196	4731.28(2)	4792.6
SARFIMA (1, 1, 2)(5, 0, 12)$_3$	1499060	0.296	−0.109	4731.77(3)	4813.5
SARFIMA (1, 1, 1)(5, 0, 6)$_3$	1101370	0.459	−0.157	4742.36(4)	4820.2

(Continued)

Table 2: Continued

Order	σ^2	d	D	AIC	BIC
		London tourists' arrival dataset			
SARFIMA $(2, 1, 1)(4, 0, 4)_4$	78394.8	−0.7579	0.4990	756.428(1)	791.219
SARFIMA $(1, 1, 2)(4, 0, 4)_4$	79214.5	0.01898	−0.1098	758.870(2)	798.009
SARFIMA $(2, 1, 1)(8, 0, 4)_4$	83298.3	−0.6033	−0.2129	760.423(3)	797.387
SARFIMA $(1, 1, 2)(8, 0, 4)_4$	83298.3	−0.6033	−0.2129	760.423(3)	797.387

6.2 *Artificial Neural Network Model*

To obtain the most accurate ANN model, numerous ANN models were established for the considered datasets using two hidden layers with varying 2 to 30 nodes in the first hidden layer, and 2 to 7 nodes in the second hidden layer. By varying the number of nodes in the first and second hidden layers, 145 models are developed, whereas each model is trained 50 times. Due to the space limitation, the detailed results are not presented here. However, from these models, the best models are selected based on minimum mean squared error (MSE) and root mean squared error (RMSE). Consequently, we obtain ANN(6 × 2 × 1), ANN(4 × 2 × 1) and ANN(10 × 2 × 1) for New Zealand, Australia, and London tourists' arrival datasets, respectively. The prediction from the best ANN model for each dataset is presented in Fig. 6 for comparison.

6.3 *Hybrid SARFIMA-ANN Model*

The hybrid algorithm mainly consisted of two steps as discussed earlier. In the first step, a SARFIMA model is fitted to analyze the linear part of the data and in the second step, the residuals from the SARFIMA model are analyzed. Linearity in the residuals is checked using the BDS test as suggested by Broock et al. [41]. In order to perform the BDS test, the following steps are being followed.

i. Select embedded dimensions (m) value so that the embed time series transforms into m-dimensional vectors by considering each m succeeding point in the series.

ii. Calculate the correlation coefficient that m-dimensional hyperspace for the proportion of points within a distance $\in$ of each other.

$$U_m L(\in) = \frac{2}{L - m + 1} \sum\nolimits_{t<m} I_\in (y_t^m - y_s^m).$$

Here $I_\in = \left\{ \begin{array}{ll} 1, & \text{if } ||y_t^m - y_s^m|| < \in \\ 0, & \text{otherwise} \end{array} \right\}$ where $||\cdot||$ used for supremum norm. BDS test illustrates that if the null hypothesis demonstrates x_t series is i.i.d., then $U_{m,I}(\in) - U_{m,L}(\in)^m$ with probability one as the sample size tends to infinity and $\in$ tends to zero.

iii. Compute BDS test statistic that is defined as

$$Q_{m,I} = \frac{T^{1/2} U_{m,I}(\in) - U_{m,L}(\in)^m}{\sigma_{m,I}(\epsilon)}$$

where, $\sigma_{m,I}(\epsilon) = 2\left[K^m + 2\sum_{j-1}^{m-1} K^{m-j} U_{I,L}(\in)^{2j} + (m-1)^2 U_{I,L}(\in)^{2m} - m^2 K U_{I,L}(\in)^{2m-2} \right]^{1/2}.$

iv. It is a two-tail test. The statement under the null hypothesis will reject when the BDS test statistic is greater than the critical value.

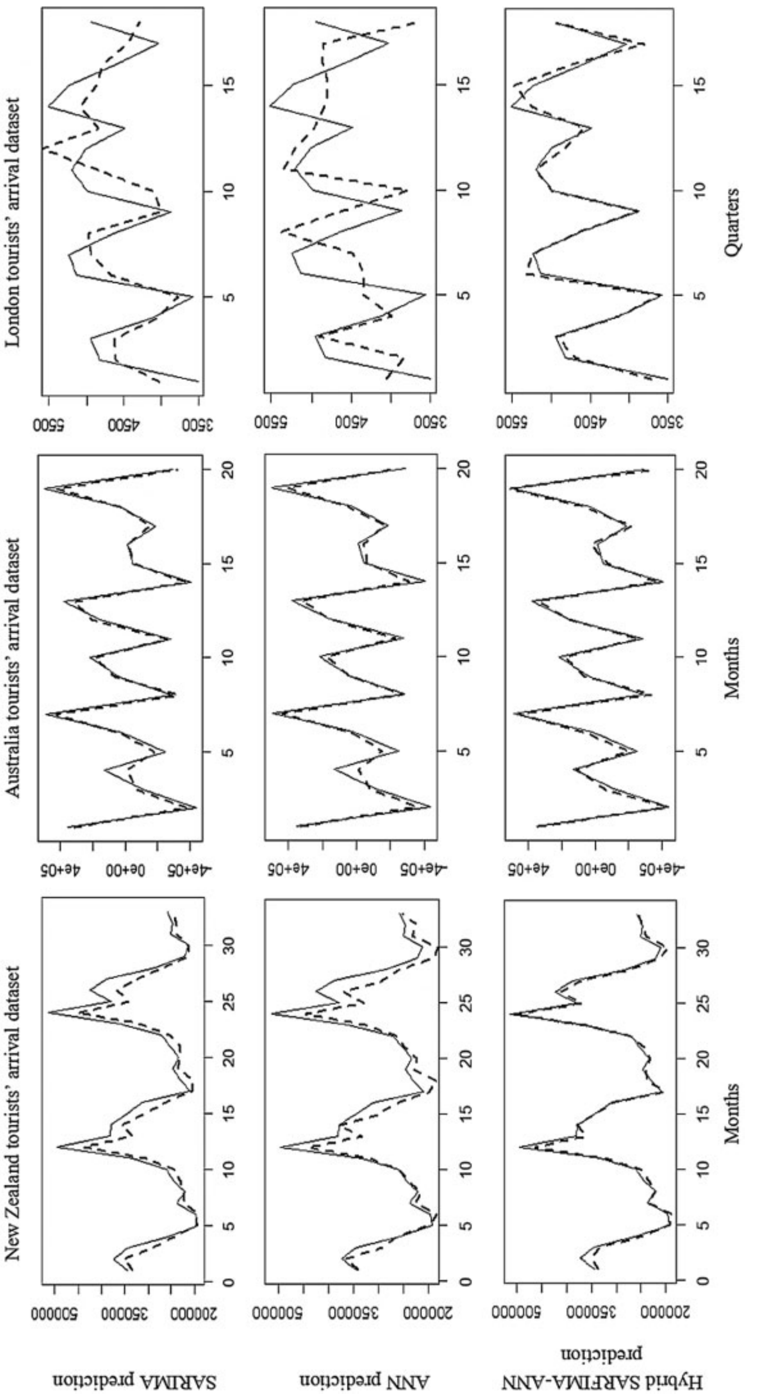


Figure 6: Prediction comparsion among SARFIMA, ANN and hybrid SARFIMA-ANN models (---- predicted and —actual values)

In Tab. 3, the results of the BDS test on residuals from selected models SARFIMA models such as SARFIMA (3, 0, 1)(10, 0, 7)$_6$, SARFIMA (1, 0, 3)(5, 0, 6)$_6$ and SARFIMA (2, 1, 1) (5, 0, 6)$_4$ are presented for tourists' arrival datasets of New Zealand, Australia, and London.

Table 3: BDS test results on residuals from the SARFIMA model

	0.5σ	σ	1.5σ	2σ
New Zealand tourists' arrival dataset				
$m = 2$	8.0854^a	13.9729^a	11.5153^a	4.5385^a
$m = 3$	14.5467^a	15.8293^a	13.5450^a	2.4235^a
$m = 4$	23.0369^a	17.0088^a	14.2159^a	2.8195^a
$m = 5$	41.3932^a	18.6316^a	14.6864^a	3.1481^a
Australia tourists' arrival dataset				
$m = 2$	-10.1289^a	14.0298^a	19.5839^a	10.4037^a
$m = 3$	11.3141^a	20.6787^a	13.7015^a	40.6735^a
$m = 4$	14.0065^a	10.3797^a	10.4214^a	20.4946^a
$m = 5$	14.4796^a	-34.2798^a	-20.1294^a	-17.0037^a
London tourists' arrival dataset				
$m = 2$	-10.1742^a	-20.3702^a	10.4737^a	10.0265^a
$m = 3$	11.6444^a	10.7924^a	11.0642^a	11.4865^a
$m = 4$	21.0811^a	11.4115^a	31.4639^a	10.4585^a
$m = 5$	22.5775^a	21.0564^a	21.4284^a	17.1260^a

Note: BDS: Brock, Dechert, and Scheinkman test for independence of residuals; m: embedding dimension; ϵ: the distance between points measured in terms of several standard deviations of the considered data; σ: standard deviation; a: significance at level 5%.

The numerical results of the BDS test, as in Tab. 3, favour rejecting the null hypothesis about the time series linearity at a 5% level of significance. It demonstrates that the errors (residuals) from the best selected SARFIMA models have nonlinear patterns. This indicates that only the linear model (SARFIMA model) is not adequate to model the data well. Therefore, implementation of nonlinear models also requires, such as ANN, to capture the nonlinearity pattern.

We consider the best selected SARFIMA and ANN models, and build the hybrid SARFIMA and ANN models for New Zealand, Australia, and London tourists' arrival data. Recall Zhang [18] that one can use separate suboptimal models to develop the hybrid method. Following their suggestion, the optimal SARFIMA model is used to model the linear part of the data and the nonlinear patterns fitted by the finalized ANN model. Then the improved prediction is obtained by combining the output of the best fitted SARFIMA and ANN model in the hybrid SARFIMA-ANN model. The performance indicators, that is, MSE and RMSE of the proposed hybrid SARFIMA-ANN and the individuals SARFIMA and ANN models, are presented in Tab. 4. The comparison of the results clearly shows that the proposed hybrid SARFIMA-ANN outperforms than the competing models. Furthermore, the out-of-sample performance of the hybrid and individual models is shown in Fig. 6. We see that the forecasting obtained by the hybrid SARFIMA-ANN model in each dataset is closer to actual than the competing models.

Table 4: The best-fitted models on the training part for all three tourists' arrival datasets

MODELS	MSE	RMSE
New Zealand Tourists arrival		
SARFIMA (3, 0, 1)(10, 0, 7)$_6$	9.70×10^{8}	3.11×10^{4}
ANN(**6 x 2 x 1**)	1.19×10^{9}	3.45×10^{4}
Hybrid SARFIMA-ANN	1.07×10^{8}	1.03×10^{4}
Australia **t**ourists arrival		
SARFIMA (1, 1, 3)(5, 0, 6)$_6$	2.52×10^{9}	5.02×10^{4}
ANN(**4 x 2 x 1**)	3.43×10^{9}	5.86×10^{4}
Hybrid SARFIMA-ANN	1.23×10^{9}	3.51×10^{4}
London **t**ourists arrival		
SARFIMA (2 , 1, 1) (5, 0, 6)$_4$	1.81×10^{5}	4.25×10^{2}
ANN(**10 x 2 x 1**)	5.00×10^{5}	7.07×10^{2}
Hybrid SARFIMA-ANN	1.88×10^{4}	1.37×10^{2}

7 Conclusion

Tourism is a rapidly growing industry in most countries and it is demanding more accurate modeling and forecasting of tourists' arrival data for many purposeful decisions. This, in turn, has grabbed increasing attention to more accurate and advanced forecasting methods. Therefore, the main interest of the study was to establish a possibility for the improvement in the forecast accuracy of tourist arrival using a hybrid modeling approach. Recently, the extension of the SARIMA, which is called the SARFIMA model, has become popular for the linear time series data with long memory processes and periodic patterns. More recently, the ANNs have shown much flexibility in modeling the nonlinear data. Therefore, ANN and SARFIMA can only achieve accurate results in their premises, and generally, none of them is the best model for every forecasting situation. Thus, in this study, a hybrid SARFIMA-ANN approach is established and applied to forecast tourists' arrival in Australia, New Zealand, and London. This approach is outperformed and produced promising results in the actual situation and produces positive results compared to the two competitors, SARFIMA and ANN. Our results have implications both for theory and application. Theoretically, we developed a hybrid SARFIMA-ANN model. In terms of application, the results of hybrid SARFIMA-ANN provide confidence for policymakers in the search volumes of tourists' arrival. Consequently, the proposed hybrid SARFIMA-ANN model and the investigation of this study make a good step in improving the forecast accuracy in tourists' arrival.

Acknowledgement: This research is supported by Artificial Intelligence & Data Analytics Lab (AIDA) CCIS Prince Sultan University, Riyadh 11586 Saudi Arabia. The authors also would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no interest in reporting regarding the present study.

References

[1] H. Song, G. Li, S. F. Witt and G. Athanasopoulos, "Forecasting tourist arrivals using time-varying parameter structural time series models," *International Journal of Forecasting*, vol. 27, no. 3, pp. 855–869, 2011.

[2] T. Saba, I. Abunadi, M. N. Shahzad and A. R. Khan, "Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types," *Microscopy Research and Technique*, vol. 84, no. 7, pp. 1462–1474, 2021.

[3] C. W. Granger, "Long memory relationships and the aggregation of dynamic models," *Journal of Econometrics*, vol. 14, pp. 227–238, 1980.

[4] P. M. Robinson, "Time series with long memory," in *Advanced Texts in Econometrics*, UK: Oxford University Press, 2003.

[5] J. A. Doornik and M. Ooms, "Inference and forecasting for ARFIMA models with an application to US and UK inflation," *Studies in Nonlinear Dynamics & Econometrics*, vol. 8, no. 2, pp. 1–25, 2004.

[6] F. L. Chu, "Forecasting tourism demand with ARMA-based methods," *Tourism Management*, vol. 30, no. 5, pp. 740–751, 2009.

[7] Y. W. Cheung, "Long memory in foreign-exchange rates," *Journal of Business & Economic Statistics*, vol. 11, no. 1, pp. 93–101, 1993.

[8] T. S. Prass, J. M. Bravo, R. T. Clarke, W. Collischonn and S. R. Lopes, "Comparison of forecasts of mean monthly water level in the Paraguay river, Brazil, from two fractionally differenced models," *Water Resources Research*, vol. 48, no. 5, pp. 1–13, 2012.

[9] H. Mostafaei and L. Sakhabakhsh, "Using SARFIMA model to study and predict the Iran's oil supply," *International Journal of Energy Economics and Policy*, vol. 2, no. 1, pp. 41–49, 2012.

[10] L. Peng, L. Wang, X. Y. Ai and Y. R. Zeng, "Forecasting tourist arrivals via random forest and long short-term memory," *Cognitive Computation*, vol. 13, no. 1, pp. 125–138, 2021.

[11] K. J. Waciko and B. Ismail, "Forecasting tourist arrival to bali-Indonesia from 3 continents using thief-mLP hybrid method," *International Journal of Scientific & Technology Research*, vol. 9, no. 2, pp. 1721–1725, 2020.

[12] D. C. Wu, L. Ji, K. He and K. F. Tso, "Forecasting tourist daily arrivals with a hybrid SARIMS–LSTM approach," *Journal of Hospitality & Tourism Research*, vol. 45, no. 1, pp. 52–67, 2021.

[13] T. Saba, A. Rehman and J. S. AlGhamdi, "Weather forecasting based on hybrid neural model," *Applied Water Science*, vol. 7, no. 7, pp. 3869–3874, 2017.

[14] P. Sarvareddy, H. Al-Deek, J. Klodzinski and G. Anagnostopoulos, "Evaluation of two modeling methods for generating heavy-truck trips at an intermodal facility by using vessel freight data," *Transportation Research Record*, vol. 1906, no. 1, pp. 113–120, 2005.

[15] G. P. Zhang and M. Qi, "Neural network forecasting for seasonal and trend time series," *European Journal of Operation Research*, vol. 160, no. 2, pp. 501–514, 2015.

[16] V. R. Prybutok and J. Yi, "Mitchell D. comparison of neural network models with ARIMA and regression models for prediction of houston's daily maximum ozone concentrations," *European Journal of Operation Research*, vol. 122, no. 1, pp. 31–40, 2000.

[17] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for time-series forecasting," *Expert System with Applications*, vol. 37, no. 1, pp. 479–489, 2010.

[18] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, no. 5, pp. 159–175, 2003.

[19] P. F. Pai and C. S. Lin, “A hybrid ARIMA and support vector machines model in stock price forecasting,” *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[20] K. Chen and C. A. Wang, “A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan,” *Expert System with Applications*, vol. 32, pp. 254–264, 2007.

[21] N. Chaâbane, “Electrical power and energy systems a hybrid ARFIMA and neural network model for electricity price prediction,” *International Journal of Electrical Power & Energy System*, vol. 55, pp. 187–194, 2014.

[22] N. A. Chaâbane, “A novel auto-regressive fractionally integrated moving average–least-squares support vector machine model for electricity spot prices prediction,” *Journal of Applied Statistics*, vol. 41, pp. 635–651, 2014.

[23] H. Song and S. F. Witt, “Forecasting international tourist flows to Macau,” *Tourism Management*, vol. 27, no. 2, pp. 214–224, 2006.

[24] C. Phetchanchai, A. Selamat, T. Saba and A. Rehman, “Index financial time series based on zigzag-perceptually important points,” *Journal of Computer Science*, vol. 6, no. 12, pp. 1389–1395, 2010.

[25] S. Mujeeb, T. A. Alghamdi, S. Ullah, A. Fatima, N. Javaid *et al.*., “Exploiting deep learning for wind power forecasting based on big data analytics,” *Applied Sciences*, vol. 9, no. 20, pp. 1–18, 2019.

[26] S. L. Marie-Sainte, T. Saba and S. Alotaibi, “Air passenger demand forecasting using particle swarm optimization and firefly algorithm,” *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 9, pp. 3735–3743, 2019.

[27] M. A. Khan, S. Kadry, Y. D. Zhang, T. Akram, M. Sharif *et al.*, “Prediction of COVID-19-pneumonia based on selected deep features and one class kernel extreme learning machine,” *Computers & Electrical Engineering*, vol. 90, pp. 106960, 2021.

[28] A. M. A. Haimed, T. Saba, A. Albasha, A. Rehman and M. Kolivand, “Viral reverse engineering using artificial intelligence and big data COVID-19 infection with long short-term memory (LSTM),” *Environmental Technology & Innovation*, vol. 22, pp. 101531, 2021.

[29] A. Naz, M. U. Javed, N. Javaid, T. Saba, M. Alhussein *et al.*, “Short-term electric load and price forecasting using enhanced extreme learning machine optimization in smart grids,” *Energies*, vol. 12, no. 5, pp. 1–30, 2019.

[30] C. W. Granger and R. Joyeux, “An introduction to long-memory time series models and fractional differencing,” *Journal of Time Series Analysis*, vol. 1, no. 2, pp. 15–29, 1980.

[31] R. T. Baillie, “Long memory processes and fractional integration in econometrics,” *Journal of Economics*, vol. 73, no. 1, pp. 5–59, 1996.

[32] L. Sakhabakhsh and M. Yarmohammadi, “An empirical study of the usefulness of SARFIMA models in energy science,” *International Journal of Energy Science*, vol. 2, no. 2, pp. 59–63, 2012.

[33] P. Doukhan, G. Oppenheim and M. Taqqu, “Theory and applications of long-range dependence,” Springer Science & Business Media, 2002.

[34] N. Katayama, “Seasonally and fractionally differenced time series,” *Hitotsubashi Journal of Economics*, vol. 48, no. 1, pp. 25–55, 2007.

[35] H. I. Park, “Study for application of artificial neural networks in geotechnical problems,” in *Artificial Neural Networks-Application*, Rijeka in Croatia: IntechOpen, 2011.

[36] H. Rabie, M. El-Beltagy, A. Tharwat and S. Hassan, “Exploring input selection for time series forecasting,” in *DMIN*, Las Vegas, Nevada, USA, pp. 228–232, 2008.

[37] I. S. Markham and T. R. Rakes, “The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression,” *Computers & Operations Research*, vol. 25, no. 4, pp. 251–263, 1998.

[38] N. Kulendran and M. L. King, “Forecasting international quarterly tourist flows using error-correction and time-series models,” *International Journal of Forecasting*, vol. 13, no. 3, pp. 319–327, 1997.

[39] P. Balogh, S. Kovacs, C. Chaiboonsri and P. Chaitip, “Forecasting with X-12-ARIMA: International tourist arrivals to India and Thailand,” *Applied Studies in Agribusiness and Commerce*, vol. 3, no. 4, pp. 43–61, 2009.

[40] S. Sood and K. Jain, "Comparative analysis of techniques for forecasting tourists arrival," *Journal of Tourism Hospitality*, vol. 6, no. 3, pp. 1–4, 2017.

[41] W. A. Broock, J. A. Scheinkman, W. D. Dechert and B. A. LeBaron, "Test for independence based on the correlation dimension," *Econometric Reviews*, vol. 15, no. 3, pp. 197–235, 1996.

Appendix

A hybrid algorithm based on SARFIMA and Neural network

```
    Hypothesis: $Y_t = {Y_t}^{LIN} + {Y_t}^{NLIN}$
    Input: Time series dataset $Y_t = (y_{t1}, y_{t1}, \ldots, y_{N_t})$, test to confirm stationarity (adf.test($Y_t$), pp($Y_t$), kpss($Y_t$))
    # Fitting SARIMA model
    Set: $\alpha = aY_t$, $a$ is a constant, training set size $N_{tr} = N - \alpha$, test set size $N_{ts} = \alpha$, vectors of zeros
        ${Y_t}^{LIN} = 0[, 1 : N_{tr}]$, $\widehat{{Y_{ts}}^{LIN}} = 0[, 1 : N_{ts}]$; $R = 0[, 1 : N_{tr}]$.
      for j in (1: $N_{tr} + N_{ts}$) do
                  $\widehat{Y}_t(j) = SARFIM\widehat{A(p,} d, q)x(P, D, Q)s ,(Y_t, j : j + (N_t - \alpha) - 1)$
        if j $\leq N_{tr}$ Then
          R (j) = $Y_t(j) - \widehat{Y}_t(j)$
        else
          $\widehat{Y_{ts}}(j) = \widehat{{Y_{ts}}^{LIN}} = \widehat{\widehat{Y}_t(j)}$
        end if
    end for
    # Fitting ANN model
    Set: ${Y_{ts}}^{NLIN} = 0[, 1 : N_{ts}]$ and $Y_t = 0[, 1 : N_{tr} + N_{ts}]$
    for k in (1:$N_{ts}$) do
        $\widehat{{Y_{ts}}^{NLIN}}(k) = ANN(R; p \times q \times 1)$
    End for
    Output: $Y_t = \widehat{{Y_{ts}}^{LIN}} + \widehat{{Y_{ts}}^{NLIN}}$
```