Expert Systems with Applications 38 (2011) 10368–10376

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Combining linear and nonlinear model in forecasting tourism demand

Kuan-Yu Chen *

Department of Recreation Sport and Health Promotion, National Pingtung University of Science and Technology, 1, Hseuh Fu Rd., Neipu, Pingtung 62248, Taiwan

ARTICLE INFO

Keywords:
Combination forecasting
Tourism demand
Support vector regression
Forecasting accuracy

ABSTRACT

Much research shows that combining forecasts improves accuracy relative to individual forecasts. However, existing non-tourism related literature shows that combined forecasts from a linear and a nonlinear model can improve forecasting accuracy. This paper combined the linear and nonlinear statistical models to forecast time series with possibly nonlinear characteristics. Real time series data sets of Taiwanese outbound tourism demand were used to examine the forecasting accuracy of the combination models. The forecasting performance was compared among three individual models and six combination models, respectively. Among these models, the normalized mean square error (NMSE) and the mean absolute percentage error (MAPE) of the combination models were the lowest. The combination models were also able to forecast certain significant turning points of the test time series. Thus, this paper suggests that forecast combination can achieve considerably better predictive performances and show promising results in directional change detects ability in the tourism context. Besides, the empirical results also clearly show that how a high forecasting accuracy and an excellent directional change detect ability could be achieved by the SVR combination models.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The tourism industry, which benefits the transportation, accommodation, catering, entertainment and retailing sectors, has been blooming in the past few decades. The 20th century witnessed a steady increase in tourism all over the world. Each country wants to know its international visitors and tourism receipts in order to choose an appropriate strategy for its economic well-being. Hence, a reliable forecast is needed and plays a major role in tourism planning.

Accurate forecasts build a sound foundation for better tourism planning and administration. This calls for more efficient and accurate forecasting techniques in tourism demand studies. Time series forecasting is an important area of forecasting in which past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. The model is then used to extrapolate the time series into the future. This modeling approach is particularly useful when little knowledge is available on the underlying data generating process or when there is no satisfactory explanatory model that relates the prediction variable to other explanatory variables. Much effort has been devoted over the past several decades to the development and improvement of time series forecasting models.

In tourism demand studies, the time series forecasting models can be classified into two categories: the linear methods and the nonlinear methods. The most popular of the conventional linear methods are the Naïve method (Burger, Dohnal, Kathrada, & Law, 2001; Chu, 2004), the exponential smoothing (ES) model (Burger et al., 2001; Kim & Ngo, 2001; Law, 2000b), and the autoregressive integrated moving averages (ARIMA) models (Goh & Law, 2002; Law, 2000a, 2004; Qu & Zhang, 1996). Among them, the ARIMA model is the most advanced forecasting model that has been successfully tested in many practical applications. In order to use them, users must specify the model form without the necessary genuine knowledge about the complex relationship in the data. Of course, if the linear models can approximate the underlying data generating process well, they should be considered as the preferred models over more complicated models as linear models have the important practical advantage of easy interpretation and implementation. However, if the linear models fail to perform well in both in-sample fitting and out-of-sample forecasting, more complex nonlinear models should be considered. Based on this point of view, many scholars have also turned to nonlinear methods such as the neural network (NN) (Cho, 2003; Kon & Turner, 2005; Law, 2000b; Law & Au, 1999; Pai & Hong, 2005; Palmer, Montano, & Sese, 2006) and support vector regression (SVR) models (Chen & Wang, 2005, 2007; Pai, Hong, Chang, & Chen, 2006) for tourism demand forecasting. NNs and SVRs are designed to pick up nonlinear patterns from time series. Although there are still a few doubts about neural-network-based tourism demand forecasting, it is generally believed that the nonlinear methods outperform the linear methods in modeling economic behavior and efficiently helping wise decision-making.

* Tel.: +886 6 2010577.
E-mail address: t530903@ms7.hinet.net

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.049

Many stationary phenomena in practice can be described or at least be approximated by linear time series models. However, many nonlinear phenomena such as limit cycles, frequency modulations and animal population cycles cannot be described adequately by linear time series models. Intuitively, if the true model is a nonlinear time series model, then any statistical inferences using analysis for the nonlinear model, which capture the nonlinear characteristic of the data, should be better than those by using linear approximation. However, this is not always the case for the prediction problem. For example, Davies, Petruccelli, and Pemberton (1988) and Pemberton (1989), by numerical simulation, observe the phenomenon that the conditional mean and conditional median forecasts of nonlinear time series models have poor forecast performance compared to those of linear models. For this reason, an important motive to combine forecasts from different models is the fundamental assumption that one cannot identify the true process exactly, but different models may play a complementary role in the approximation of the data generating process.

The concept of combining forecasts started with the seminal work 40 years ago of Bates and Granger (1969). Given two individual forecasts of a time series, they demonstrated that a suitable linear combination of the two forecasts may result in a better forecast than the two original ones, in the sense of a smaller error variance. Different forecasting models can achieve successes each other in capturing patterns of data sets, and many authors have shown that combining the predictions of several models often results in a prediction accuracy that is higher than that of the individual models (Lawerence, Edmundson & O'Connor, 1986; Makridakis, 1989; Makridakis & Winkler, 1983). Throughout the years, applications of combining forecasts have been found in many fields such as meteorology, economics, insurance and forecasting sales and price. Since the early work of Reid (1968), Bates and Granger (1969), the literature on this topic has expanded dramatically. Clemen (1989) provided a comprehensive review and annotated bibliography in this area. Wedding and Cios (1996) described a combining methodology using radial basis function networks and the Box–Jenkins models. Luxhoj, Riis, and Stensballe (1996) presented a hybrid econometric and ANN approach for sales forecasting. Pelikan, de Groot, and Wurtz (1992) and Ginzburg and Horn (1994) proposed to combine several feed-forward neural networks to improve time series forecasting accuracy. Tseng, Yu, and Tzeng (2002) proposed a hybrid forecasting model, which combines the SARIMA and the neural network back-propagation models. Zhang (2003) combined the ARIMA and feed-forward neural networks models in forecasting. Oh and Morzuch (2005) showed that the combined forecasts models always outperform the poorest individual forecasts, and sometimes even perform better than the best individual model. Wong, Song, Witt, and Wu (2007) demonstrated that combined forecasts can generally outperform the worst individual forecasts; thereby risk of complete forecast failure could be reduced through forecast combination.

In this paper, we propose a combined approach to tourism demand forecasting using both linear model and the nonlinear model. The motivation of the combined model comes from the following perspectives. First, it is often difficult to determine whether a time series is generated from a linear or nonlinear underlying process or whether one particular method is more effective than the other in tourism demand forecasting. Second, tourism demand time series are rarely pure linear or nonlinear. They often contain both linear and nonlinear patterns. Third, it is almost generally agreed in the forecasting literature that no single method is best in every situation.

The remainder of this study is organized as follows: In Section 2 the Naïve, the exponential smoothing, the ARIMA, the neural network, the SVR model, and the combination methodology are described. In Section 3 describes the data source. Section 4 discusses the evaluation methods used for comparing the forecasting techniques. Section 5 compares the results obtained from the combination models against the individual models. Section 6 provides concluding remarks.

2. Methodology

The linear and the nonlinear models are summarized in the following as foundation to describe the proposed combination models.

2.1. Naïve method

The Naïve forecasting method simply states that the forecast value for this period (F_t) is equal to the actual value of the last period available (X_{t-1}). More formally:

$$F_t = X_{t-1}, \tag{1}$$

where F = forecast value; X = actual value; t = some time period.

Eq. (1) is equivalent to a random walk model which assumes that trends and turning points cannot be predicted and the forecast is a horizontal line extrapolation.

2.2. Exponential smoothing

Exponential smoothing (ES), the most common method of time series prediction, was developed on the basis of the moving average technique. It forecasts the next value based on current actual value and current forecasted the influence of the nearest actual value on the forecasted value, but not need masses of past values:

$$F_{t+1} = \alpha X_t + (1-\alpha)F_t, \tag{2}$$

where F_{t+1} is the forecast value of period $t+1$; α the smoothing constant ($0 < \alpha < 1$), X_t the actual value now (in period t), and F_t the forecast (i.e., smoothed) value for period t. From Eq. (2), we can see that α can be thought of as the weight given to past history. The larger the value of α, the less weight the past history has in relation to the last actual value of the parameter. The method is called 'exponential', since the forecasted value is the discrete convolution of the observed sequence with an exponential curve with a time constant $1/(1-\alpha)$. Alternately, if the value of X_t becomes fixed, the error ($X_t - F_t$) decays exponentially.

2.3. ARIMA model

ARIMA is the most popular linear model for forecasting time series. It has enjoyed great success in both academic research and industrial applications during the last three decades. A general ARIMA model of order (p, d, q) representing the time series can be written as:

$$\phi(B)\nabla^d x_t = \theta(B)\varepsilon_t, \tag{3}$$

where x_t and ε_t represent the number of visitors and random error terms at time t respectively. B is a backward shift operator defined by $Bx_t = x_{t-1}$, and related to ∇ by $\nabla = 1 - B$, $\nabla^d = (1-B)^d$, d is the order of differencing. $\phi(B)$ and $\theta(B)$ are autoregressive (AR) and moving averages (MA) operators of orders p and q, respectively, and are defined as:

$$\phi(B) = 1\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,$$

and

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q,$$

where $\phi_1, \phi_1, \ldots, \phi_p$ are the autoregressive coefficients and are θ_1, $\theta_2, \ldots, \theta_p$ the moving average coefficients.

When fitting a ARIMA model to the raw data, the ARIMA model involves the following four-step iterative cycles:

(a) Identification of the $ARIMA(p, q, d)$ structure.
(b) Estimation of the unknown parameters.
(c) Goodness-of-fit tests on the estimated residuals.
(d) Forecast future outcomes based on the known data.

The ε_t should be independently and identically distributed as normal random variables with mean = 0 and constant variance σ^2. The roots of $\phi_p(x_t) = 0$ and $\theta_q(x_t) = 0$ should all lie outside the unit circle. It was suggested by Box and Jenkins (1976) that at least 50 or preferably 100 observations should be used for the ARIMA model.

2.4. Back-propagation neural network

Several types of neural architectures are available, among which the back-propagation neural network (BPNN) is the most widely used. As Fig. 1 reveals, a back-propagation network typically employs three or more layers for the architecture: an input layer, an output layer, and at least one hidden layer. The computational procedure of this network is described below:

$$Y_j = f\left(\sum_i w_{ij} X_{ij}\right), \tag{4}$$

where Y_j is the output of node j, $f(\cdot)$ the transfer function, w_{ij} the connection weight between node j and node i in the lower layer and X_i the input signal from the node i in the lower layer.

The back-propagation algorithm is a gradient descent algorithm. It tries to improve the performance of the neural network by reducing the total error by changing the weights along its gradient. The back-propagation algorithm minimizes the square errors, which can be calculated by

$$E = 1/2 \sum_p \sum_j \left[O_j^p - Y_j^p\right]^2, \tag{5}$$

where E is the square errors, p the index of pattern, O the actual (target) output and Y the network output.

The BPNN algorithm is based on a steepest descent technique with a momentum weight/bias function, which calculates the weight change for a given neuron. It is expressed as follows Huang, Hwang, and Hsieh (2002): let $\Delta w_{ij}^p(n)$ denote the synaptic weight connecting the output of neuron i to the input of neuron j in the p th layer at iteration n. The adjustment $\Delta w_{ij}^p(n)$ to $\Delta w_{ij}^p(n)$ is given by

$$\Delta w_{ij}^p(n) = -\eta(n) \frac{\partial E(n)}{\partial w_{ij}^p}, \tag{6}$$

where $\eta(n)$ is the learning rate parameter. By using the chain rule of differentiation, the weight of the network with the back-propagation learning rule is updated using the following formulae:

$$\Delta w_{ij}^p(n) = \eta(n)\delta_j^p(n) X_i^{p-1}(n) + m(n)\Delta w_{ij}^p(n-1), \tag{7}$$

$$\Delta w_{ij}^p(n+1) = w_{ij}^p(n) + \Delta w_{ij}^p(n), \tag{8}$$

where $\delta_j^p(n)$ is the nth error signal at the jth neuron in the pth layer, $X_i^{p-1}(n)$ is the output signal of neuron i at the layer below and m is the momentum factor.

The constant terms of η and m are specified at the start of the training cycle and determine the speed and stability of the network. In brief, the procedure to set up a back-propagation network is:

(1) Select input and define output variables.
(2) Determine layer(s) and the number of neurons in hidden layers. No hard rule is available for determining them, which may depend on trial and error.
(3) *Learning (or training) from historical data.* Learning is the process by which a neural network modifies its weights in response to external inputs in order to minimize the global error. The equation that specifies this change is called the learning rule.
(4) *Testing*: When a neural network is well trained after learning, the neural networks are processed via a test set containing historical data that the network has never seen. If the testing results are in an acceptable range, the network can be considered as fully trained. The next step can then be performed.
(5) *Recalling*: Recalling refers to how the network processes a driving force from its input layer and creates a response at the output layer. Recalling does not do any learning and expects no desired outputs.

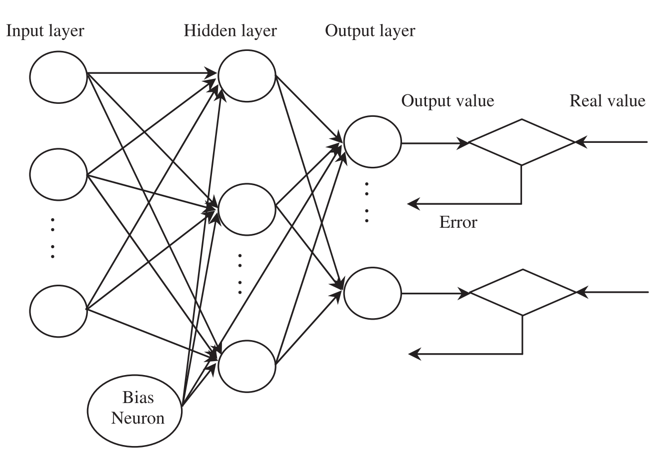

Fig. 1. A back-propagation neural network.

2.5. Support vector regression

Recently, a regression version of SVM has emerged as an alternative and powerful technique to solve regression problems by introducing an alternative loss function. In the sequel, this version is referred to as support vector regression (SVR). Here a brief description of SVR is given. Detailed descriptions of SVR can be found in Vapnik (1995), Schölkopf and Smola (2002).

The SVR formulation follows the principle of structural risk minimization seeking to minimize an upper bound of the generalization error rather than minimize the prediction error on the training set (the principle of empirical risk minimization). This equips the SVR with a greater potential to generalize the input–output relationship learnt during its training phase for making good predictions for new input data. The SVR maps the input data x into a high-dimensional feature space F by nonlinear mapping, to yield and solve a linear regression problem in this feature space (see Fig. 2). The regression approximation estimates a function according to a given data set $G = \{(x_i, d_i)\}_i^n$, where x_i denotes the input vector; d_i denotes the output (target) value, and n denotes the total number of data patterns. The modelling aim is to identify a regression function, $y = f(x)$, that accurately predicts the outputs $\{d_i\}$ corresponding to a new set of input–output examples, $\{(x_i, d_i)\}$. Using mathematical notation, the linear regression function (in the feature space) is approximated using the following function:

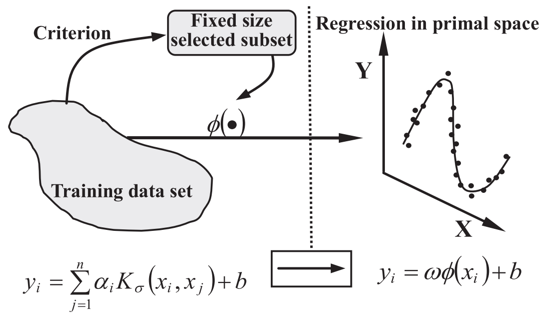

Fig. 2. Mapping input space × into high-dimensional space (from Peng, Wu, & Goo, 2004).

$$f(x) = \omega\phi(x) + b,\qquad \phi : R^n \to F, \omega \in F, \tag{9}$$

where ω and b are coefficients; $\phi(x)$ denotes the high-dimensional feature space, which is nonlinearly mapped from the input space x. Therefore, the linear regression in the high-dimensional feature space responds to nonlinear regression in the low-dimension input space, disregarding the inner product computation between ω and $\phi(x)$ in the high-dimensional feature space. Correspondingly, the original optimization problem involving nonlinear regression is transformed into finding the flattest function in the feature space F, and not in the input space, x. The unknown parameters ω and b in Eq. (9) are estimated by the training set, G.

SVR performs linear regression in the high-dimensional feature space by ε-insensitive loss. At the same time, to prevent over-fitting and thereby improving the generalization capability, following regularized functional involving summation of the empirical risk and a complexity term $\|\omega\|^2/2$, is minimized. The coefficients ω and b can thus be estimated by minimizing the regularized risk function:

$$R_{\mathrm{SVR}}(C) = R_{\mathrm{emp}} + \frac{1}{2}\|\omega\|^2 = C \times \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|\omega\|^2, \tag{10}$$

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon, & |d - y| \geqslant \varepsilon, \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where R_{SVR} and R_{emp} represent the regression and empirical risks, respectively; $\|\omega\|^2/2$ denotes the Euclidean norm, and C denotes a cost function measuring the empirical risk. In the regularized risk function given by Eq. (10), the regression risk (test set error), R_{SVR}, is the possible error committed by the function f in predicting the output corresponding to a new (test) example input vector. In Eq. (10), the first term $C \times (1/n)\sum_{i=1}^{n} L_\varepsilon(d_i, y_i)$ denotes the empirical error (termed "training set error"), which is estimated by the ε-insensitive loss function in Eq. (11). The loss function is introduced to obtain sufficient samples of the decision function in Eq. (9) by using fewer data points. The second item, $\|\omega\|^2/2$, is the regularization term. The regularized constant C calculates the penalty when an error occurs, by determining the trade-off between the empirical risk and the regularization term, which represents the ability of prediction for regression. Raising the value of C increases the significance of the empirical risk relative to the regularization term. The penalty is acceptable only if the fitting error is larger than ε. The ε-insensitive loss function is employed to stabilize estimation. In other words, the ε-insensitive loss function can reduce the noise. Thus, ε can be viewed as a tube size equivalent to the approximation accuracy in training data (see Fig. 3). In the empirical analysis, C and ε are the parameters selected by users.

To estimate ω and b, Eq. (10) is converted to the primal function given by Eq. (12) by introducing the positive slack variables ξ_i and ξ_i^*. According to Fig. 3, the sizes of the stated excess positive and

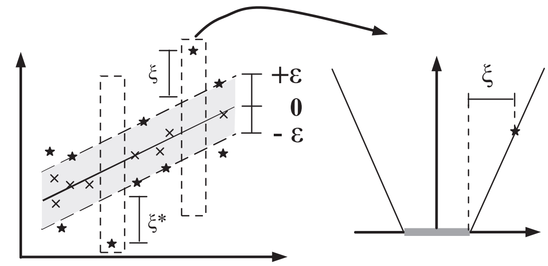

Fig. 3. The soft margin loss setting for a linear SVR (from Schölkopf & Smola, 2002). In SVR, a tube with radius ε is correlated with the data. The trade-off between the model complexity (flatness) and points lying outside the tube (slack variables ξ_i and ξ_i^*) is determined by minimizing Eq. (12). The points lying on or outside the ε-bound of decision function are support vectors (black stars). On the right, the ε-insensitive loss function is shown in which the slope is determined by C (the star represents a support vector).

negative deviations are represented by ξ_i and ξ_i^*, respectively, which are termed "slack" variables. The slack variables assume non-zero values outside the $[-\varepsilon, \varepsilon]$ region. The SVR fits $f(x)$ to the data such that: (i) the training error is minimized by minimizing ξ_i and ξ_i^*, and (ii) $\|\omega\|^2/2$ is minimized to raise the flatness of $f(x)$, or to penalize excessively complex fitting functions. Thus, SVR is formulated as minimization of the following functional:

$$\begin{aligned} &\text{Minimize} \quad R_{\mathrm{SVR}}\left(\omega, \xi^{(*)}\right) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right), \\ &\text{Subjected to} \quad \begin{matrix} d_i - \omega\phi(x_i) - b_i \leqslant \varepsilon + \xi_i, \\ \omega\phi(x_i) + b_i - d_i \leqslant \varepsilon + \xi_i^*, \end{matrix} \quad \xi^{(*)} \geqslant 0 \end{aligned} \tag{12}$$

where ξ_i and ξ_i^* denote slack variables that measure the error of the up and down sides, respectively. The above formulae indicate that increasing ε decreases the corresponding ξ_i and ξ_i^* in the same constructed function $f(x)$, thereby reducing the error resulting from the corresponding data points.

Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, the decision function given by Eq. (9) has the following explicit form (Vapnik, 1995):

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x, x_i) + b \tag{13}$$

2.5.1. Lagrange multipliers

In Eq. (13), α_i and α_i^* are the so-called Lagrange multipliers. They satisfy the equalities $\alpha_i \times \alpha_i^* = 0, \alpha_i \geqslant 0$ and $\alpha_i^* \geqslant 0$ where $i = 1, 2, \ldots, n$, and are obtained by maximizing the dual function of Eq. (12), and the maximal dual function in Eq. (12) which has the following form:

$$\begin{aligned} \text{Max} \quad R(\alpha_i, \alpha_i^*) = \sum_{i=1}^{n} d_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{n} \\ \times \sum_{j=1}^{n}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)K(x_i, x_j), \end{aligned} \tag{14}$$

with the constraints:

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \qquad \begin{matrix} 0 \leqslant \alpha_i \leqslant C & i = 1, 2, \ldots, n, \\ 0 \leqslant \alpha_i^* \leqslant C & i = 1, 2, \ldots, n. \end{matrix}$$

Based on the Karush–Kuhn–Tucker's (KKT) conditions of solving quadratic programming problem, $(\alpha_i - \alpha_i^*)$ in Eq. (13), only some of them will be held as non-zero values. These approximation errors of data point on non-zero coefficient will equal to or larger than ε, and are referred to as the support vector. That is, these data points lie on or outside the ε-bound of decision function. According to Eq.

(13), the support vectors are clearly the only elements of the data points employed in determining the decision function as the coefficient $(\alpha_i - \alpha_i^*)$ of other data points are all equal to zero. Generally, the larger the ε value, the fewer the number of support vectors, and thus the sparser the representation of the solution. Nevertheless, increasing ε decreases the approximation accuracy of training data. In this sense, ε determines the trade-off between the sparseness of representation and closeness to the data (Tay & Cao, 2001).

2.5.2. Kernel function

The term $K(x_i, x_j)$ in Eq. (12) is defined as kernel function, where the value of kernel function equals the inner product of two vectors x_i and x_j in the feature space $\phi(x_i)$ and $\phi(x_j)$ meaning that $K(x_i, x_j) = \phi(x_i) * \phi(x_j)$. The kernel function is intended to handle any dimension feature space without the need to calculate $\phi(x)$ accurately (Tay & Cao, 2001). If any function can satisfy Mercer's condition, it can be employed as a kernel function (Vapnik, 1995). The typical examples of kernel function are the polynominal kernel $(K(x, y) = (x \times y + 1)^d)$ and the Gaussian kernel $(K(x, y) = \exp(-(x - y)^2/2\sigma^2))$. In these equations, d represents the degree of the polynominal kernel, and σ^2 represents the bandwidth of Gaussian kernel. These parameters must be selected accurately, since they determine the structure of high-dimensional feature space and govern the complexity of the final solution.

2.6. The combination methodology

Both linear and nonlinear models have achieved successes in their own linear or nonlinear problems. However, none of them is a universal model that is suitable for all situations. Since it is difficult to completely know the characteristics of the time series data in a real problem, a combination strategy that has both linear and nonlinear modeling abilities is a good alternative for forecasting the seasonal time series data. Both the linear and nonlinear models have different unique strength to capture data characteristics in linear or nonlinear domains, so the combination models proposed in this study is composed of the linear component and the nonlinear component. Thus, the combination models can model linear and nonlinear patterns with improved overall forecasting performance.

It may be reasonable to consider a time series to be composed of a linear autocorrelation structure and a nonlinear component. That is:

$$Z_t = L_t + N_t, \tag{15}$$

where L_t is the linear component and N_t is the nonlinear component of the combination models. Both L_t and N_t have to be estimated from the data set. First, the author let linear model (Naïve, ES, and ARIMA) to model the linear part, and then the residuals from the linear model will contain only the nonlinear relationship. Let E_t represent the residual at time t as obtained from the linear model, then:

$$E_t = Z_t - \widehat{L}_t, \tag{16}$$

where $\widehat{L}_t$ denote the forecast value of the linear model at time t. By modeling residuals using nonlinear model (BPNN, SVR), nonlinear relationships can be discovered. In this study, the author built six various combination models with the following input layers:

$$E_t^{\text{linear}} = f^{\text{nonlinear}}\left(E_{t-1}^{\text{linear}}, E_{t-2}^{\text{linear}}, E_{t-3}^{\text{linear}}, E_{t-4}^{\text{linear}}\right) + e_t, \tag{17}$$

where E_t^{linear} represent the residual at time t from the linear models (Naïve, ES, and ARIMA), $f^{\text{nonlinear}}$ a nonlinear function determined by the nonlinear models (BPNN and SVR) and e_t is the random error. Here, the author proposed six combined models, and called them as Naïve_BPNN, ES_BPNN, ARIMA_BPNN, Naïve_SVR, ES_SVR, and ARIMA_SVR. Therefore, the combined forecast will be:

$$\widehat{Z}_t = \widehat{L}_t + \widehat{N}_t, \tag{18}$$

where $\widehat{N}_t$ is the forecast value of Eq. (17).

3. Data set

Due to rapid economic growth and international tourism promotion, leisure travel is prevailing in Taiwan as well as outbound travel is greatly increasing year by year. In order to thoroughly shed light on Taiwanese outbound tourism demand and to provide a comprehensive overlook, this study covers completely Taiwanese global outbound tourism demands. The monthly outbound traveling population data that Taiwanese travel to 6 areas (Mainland China, Northeast Asia, Southeast Asia, Americas, Europe, Oceania) were collected from Taiwan Tourism Bureau of Ministry. Due to data availability, the study time ranges from 1998 January to 2009 June.

The collected data were divided into two sets, training data (in-sample data) and testing data (out-of-sample data) for each tourism demand time series, in order to testify the performance of the nine suggested forecasting methods. To achieve a more reliable and accurate result, a long period served as the training period. Based on these considerations, the period from January 1998 to December 2007 was selected as the training period, and January 2008 to June 2009 as the testing period. Each data points will be scaled by Eq. (19) within the range of (0, 1). This scaling for original data points will be helpful to improve the forecasting accuracy (Chang & Lin, 2001):

$$\frac{Z_t - Z_{\min}}{Z_{\max} - Z_{\min}} \times 0.7 + 0.15, \tag{19}$$

Z_t, the number of outbound travelers at time t; $Z_{\max}$, the maximum of Z_t during the period of data source; $Z_{\min}$, the minimum of Z_t during the period of data source.

4. Performance criteria

4.1. Quantitative evaluations

Some quantitative statistical metrics such as normalized mean square error (NMSE), mean absolute percentage error (MAPE) and R (correlation coefficient) were used to evaluate the forecasting performance of the forecasting models. Table 1 shows these performance metrics and their calculations. NMSE and MAPE were used to measure the deviation between the actual and predicted values. The smaller the values of NMSE and MAPE, the closer were the predicted values to the actual values. The metric R was adopted to measure the correlation of the actual and the predicted values.

Table 1
Performance metrics and their calculations.

Metrics	Calculation
NMSE	$\text{NMSE} = 1/(\delta^2 n)^* \sum_{i=1}^{n}(a_i - p_i)^2$ $\delta^2 = 1/(n-1)^* \sum_{i=1}^{n}(a_i - \bar{a})^2$
MAPE	$\text{MAPE} = \frac{\sum_{i=1}^{n}\lvert a_i p_i \rvert / a_i}{n} \times 100\%$
R	$R = \frac{\sum_{i=1}^{n}(a_i \times p_i)}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} p_i^2}}$

* a_i and p_i are the actual values and predicted values.

4.2. Turning point evaluations

While the above criteria are good measures of the deviations of the predicted values from the actual values, they can not reflect a model's ability to predict turning points (the directional change). Therefore, the turning points are as important as the forecast value itself. To assess competing models, the model's ability is evaluated using these two directional tests: the directional change accuracy test (DCA) and regression test.

The DCA is a nonparametric test for directional accuracy of forecast that focuses on correct prediction of the directional change for the variable under consideration (Pesaran & Timmermann, 1992). Let $\widehat{Z}_t$ is the predicted value at time t; Z_t is the actual value at time t. This test does not require any quantitative information and only uses the signs of $\widehat{Z}_t$ and Z_t. The following indicator variables can be defined:

$$A_t = \begin{cases} 1, & \text{if } Z_t \text{ change from a low to high value,} \\ 0, & \text{otherwise,} \end{cases}$$

$$F_t = \begin{cases} 1, & \text{if } \widehat{Z}_t \text{ change from a low to high value,} \\ 0, & \text{otherwise,} \end{cases}$$

$$D_t = \begin{cases} 1, & \text{if value of } (A_t, F_t) \text{ are } (1,1) \text{ or } (0,0), \\ 0, & \text{otherwise,} \end{cases}$$

We define P, P_f, P_a as the proportion of times that the signs of D_t, F_t, A_t are, respectively, predicted correctly. Therefore, $P = \sum_{t=1}^{k} D_t/k$, and P_f, and P_a have similar expressions with variables F_t and A_t. Furthermore, under the assumption that F_t and A_t are independently distributed, we have:

$$P^* = P_a P_f + (1 - P_a)(1 - P_f). \tag{20}$$

In general, the standardized test statistics axe given by

$$S_n = \frac{P - P^*}{\{\mathrm{var}(P) - \mathrm{var}(P^*)\}^{0.5}} \sim N(0,1), \tag{21}$$

where

$$\mathrm{var}(P) = k^{-1} P^* (1 - P^*),$$

$$\mathrm{var}(P^*) = k^{-1}(2P_a - 1)^2 (P_f)(1 - P_f) + k^{-1}(2P_f - 1)^2 (P_a)(1 - P_a) + 4k^{-2}(P_f)(P_a)(1 - P_f)(1 - P_a),$$

The regression method had developed by Cumby and Modest (1987). Cumby and Modest (1987) suggest the following regression equation:

$$F_t = \alpha_0 + \alpha_1 A_1 + \xi_t, \tag{22}$$

where ξ_t is error term; and α_1 is the slope of this linear equation. Here, α_1 should be positive and significantly different from 0 in order to demonstrate those F_t and A_t have a linear relationship. This reflects the ability of a forecasting model to capture the turning points of a time series.

5. Experimental results

In total, we have nine forecast models for each tourism demand time series in this study, that is, three individual forecasting methods (Naïve, ES, and ARIMA) and six combined models (Naïve_BPNN, ES_BPNN, ARIMA_BPNN, Naïve_SVR, ES_SVR, and ARIMA_SVR). The combination models integrated the linear model with the nonlinear model and tested with the Taiwanese outbound tourism data. In order to evaluate the performance of the models, we calculate MAPE, NMSE, and R of the testing data set. In additional to NMSE, MAPE, and R measurements, the t value was also used to test the hypothesis that the individual forecasting methods and combination models have the same means of absolute forecast errors. If this hypothesis were statistically significant, we would have demonstrated a better model.

Empirical results in Tables 2–4 compares the forecasting results of different models. Table 2 gives the forecasting results of all the six time series through the Naïve method, Naïve_BPNN model, and Naïve_SVR model, respectively, with performances evaluated by MAPE, NMSE, and R. For the same time series in Table 2, the smaller MAPE, NMSE corresponds to the better forecasting accuracy. From the experimental results given in Table 2, we can find the effectiveness of the combination models (Naïve_BPNN and Naïve_SVR). The performance of the individual method (Naïve) is inferior to those of the combination models. This is due to the fact that the combination models integrate the linear and nonlinear information to forecast, while the individual model only utilizes the linear information of the time series. Besides, of the combination models, the Naïve_SVR model outperforms the Naïve_BPNN. One important reason is that the SVR adheres to the principle of structural risk minimization (Tay & Cao, 2001). Moreover, T tests indicated a rejection of the hypothesis that the MAPE of the combination model are the same as that of the individual models in Mainland China, Northeast Asia and Oceania, but the difference of performance is insignificant in other area.

Respectively, Tables 3 and 4 give the forecasting results of all the six time series through the individual model (ES, ARIMA) and combination model (ES_BPNN, ES_SVR, ARIMA_BPNN, and ARIMA_SVR). The results in Tables 3 and 4 also indicate the combination models are superior to the individual model in terms of three indices, revealing that both the BPNN combination model and the

Table 2
Comparison of prediction error of the combination models and the Naïve model.

		Mainland China	Northeast Asia	Southeast Asia	Americas	Europe	Oceania
Naïve	MAPE	10.74	19.23	13.19	16.56	16.76	19.70
	NMSE	1.6602	1.3766	0.9551	0.8459	1.3480	1.0726
	R	0.9919	0.9790	0.9855	0.9775	0.9818	0.9711
	t-Value	–	–	–	–	–	–
Naïve + BPNN	MAPE	8.01	10.91	11.64	14.98	14.79	15.88
	NMSE	1.0003	0.4341	0.6286	0.5411	1.0910	0.8009
	R	0.9968	0.9938	0.9914	0.9857	0.9856	0.9783
	t-Value	2.1530*	3.5395*	0.8386	0.5650	1.1015	1.3937
Naïve + SVR	MAPE	6.67	8.62	11.42	13.78	13.76	14.42
	NMSE	0.8813	0.3535	0.6711	0.5190	1.0413	0.7151
	R	0.9961	0.9957	0.9916	0.9862	0.9864	0.9817
	t-Value	3.0844*	3.6724*	0.9963	1.3261	1.2908	2.5294*

Null hypothesis of the existence of the same means of the forecasted MAPE generated by the combination model and the Naïve model.
* To be rejected at 5% significance level.

Table 3
Comparison of prediction error of the combination models and the ES models.

		Mainland China	Northeast Asia	Southeast Asia	Americas	Europe	Oceania
ES	MAPE	9.66	14.87	13.12	15.96	16.42	21.06
	NMSE	1.3797	0.8200	0.9426	0.7873	1.2453	0.9660
	R	0.9933	0.9878	0.9857	0.9790	0.9832	0.9741
	t-Value	–	–	–	–	–	–
ES + BPNN	MAPE	7.45	12.13	12.32	14.84	14.03	18.94
	NMSE	0.8212	0.5387	0.7936	0.5502	1.0197	0.8157
	R	0.9981	0.9925	0.9879	0.9852	0.9863	0.9811
	t-Value	2.3057*	1.7610	0.3860	0.5978	1.3792	0.5717
ES + SVR	MAPE	6.14	9.43	11.55	13.04	13.93	17.05
	NMSE	0.8344	0.3669	0.6477	0.4419	1.0778	0.6585
	R	0.9961	0.9950	0.9904	0.9888	0.9865	0.9837
	t-Value	3.4249*	2.1918*	0.9669	1.6036	1.1002	1.2284

Null hypothesis of the existence of the same means of the forecasted MAPE generated by the combination model and the ES model.
* To be rejected at 5% significance level.

Table 4
Comparison of prediction error of the combination models and the ARIMA models.

		Mainland China	Northeast Asia	Southeast Asia	Americas	Europe	Oceania
ARIMA	MAPE	6.67	18.02	14.07	16.88	16.71	13.52
	NMSE	0.6424	1.4677	1.0623	0.7949	1.2852	0.5643
	R	0.9973	0.9788	0.9837	0.9792	0.9831	0.9848
	t-Value	–	–	–	–	–	–
ARIMA + BPNN	MAPE	6.58	11.11	11.76	13.41	12.95	13.46
	NMSE	0.5939	0.4953	0.6832	0.3992	0.8153	0.5327
	R	0.9975	0.9923	0.9896	0.9918	0.9917	0.9856
	t-Value	0.0871	2.1547*	0.9664	1.0613	1.6807	0.0284
ARIMA + SVR	MAPE	5.30	9.34	10.74	11.46	11.37	11.87
	NMSE	0.3669	0.4123	0.7085	0.2878	0.6316	0.5102
	R	0.9982	0.9938	0.9930	0.9923	0.9917	0.9871
	t-Value	2.1587*	2.3065*	1.2200	2.2949*	2.8600*	1.1050

Null hypothesis of the existence of the same means of the forecasted MAPE generated by the combination model and the ARIMA model.
* To be rejected at 5% significance level.

SVR combination model can capture the non-linearity patterns in the data. Of the combination models, the SVR combination model still outperforms the BPNN combination model. Furthermore, *T* tests also indicated a rejection of the hypothesis that the MAPE of the combination model are the same as that of the individual models in some time series, but the difference of performance is insignificant in other time series.

The first column of Table 5 demonstrate the proportion of combination forecasts that outperform the individual forecast among the six time series for each of the forecast combination methods. It is clear that combining forecasts can improve forecasting accuracy for all regions concerned, in all case it is a worthwhile procedure. Here, it is interesting to note that, due to the non-linearity of the Taiwanese outbound tourism demand in the testing data, the high forecasting accuracy of the linear models does not hold. In other words, the linear models provide a poor forecast for tourism demand when there is a drastic change. On the other hand, the forecasting output from the combination models is accurate. Those results indicate that the combination models outperform the other three individual models (Naïve, ES and ARIMA), revealing that the combination models can capture part of the nonlinear patterns in the data.

Furthermore, we are mainly interested in the best possible performance attained, so we performed significance tests to check if the differences between the MAPE of the combination models for each time series and the MAPE of the individual models are significant. The second column of Table 5 shows that the proportion of significant difference between combination forecasts and individual forecasts. The proportion of significant difference is all above 16.7% for the combined forecast and the SVR combination models always make a significant difference in terms of the best attainable performance (50% on average). This indicates that the SVR combination models are typically an excellent model in forecasting tourism demand. Lastly, the third column of Table 5 shows that the combination models have obtained significant improvement on forecasting accuracy compared to the individual models. The average rates of improvement range between 13.17% and 28.11%.

Finally, the turning point evaluation method using DCA and regression Eq. (22) are shown in Tables 6–8 for each combination models. The t value of the slope coefficient α_1 of the combination models show that it is not always statistically different from zero, but in many cases it is positive and significant different from zero and the DCA test is statistically significant. The proportion of

Table 5
Summary of performance of the combination models.

	POO[a] (%)	PSD[b] (%)	IRA[c] (%)
Naïve + BPNN	100	33	20.19
ES + BPNN	100	16.7	13.17
ARIMA + BPNN	100	16.7	16.63
Naïve + SVR	100	50	28.00
ES + SVR	100	33	22.90
ARIMA + SVR	100	66.7	28.11

[a] POO: the proportion of combination forecasts that outperform the individual forecast among the six time series.
[b] PSD: the proportion of significant difference between combination forecasts and individual forecast.
[c] IRA: the improvement rates of forecasting accuracy.

Table 6
Results of turning point forecasting capability of the combination models and the Naïve models.

		Mainland China	Northeast Asia	Southeast Asia	Americas	Europe	Oceania
Naïve	DCA	−1.53	−1.11	0.98	0.49	−0.98	−1.11
	α_1 (t-Value)	−1.50	−1.05	0.92	0.45	−0.92	−1.05
Naïve + BPNN	DCA	2.42*	1.89	1.95	0.98	0.98	1.17
	α_1 (t-Value)	2.66*	1.92	2.00*	0.92	0.92	1.12
Naïve + SVR	DCA	2.40*	2.89*	2.43*	1.46	1.53	2.49*
	α_1 (t-Value)	2.63*	3.54*	2.67*	1.41	1.50	2.77

Null hypothesis of the existence of the α_1 is equal to zero. The slope coefficient α_1 should be positive and significant different from zero and the DCA test is statistically significant. This implies that for the out-of-sample data, the model had turning point forecasting power.
* To be rejected at 5% significance level.

Table 7
Results of turning point forecasting capability of the combination models and the ES models.

		Mainland China	Northeast Asia	Southeast Asia	Americas	Europe	Oceania
ES	DCA	−2.11	−2.11	0.98	0.49	−0.98	−0.11
	α_1 (t-Value)	−2.21	−2.21	0.92	0.45	−0.92	−0.10
ES + BPNN	DCA	2.40*	0.89	0.98	1.46	1.53	2.32*
	α_1 (t-Value)	2.63*	0.83	0.92	1.41	1.50	2.52*
ES + SVR	DCA	2.89*	3.35*	1.95	1.95	2.11*	1.89
	α_1 (t-Value)	3.54*	4.77*	2.00*	2.00*	2.21*	1.92

Null hypothesis of the existence of the α_1 is equal to zero. The slope coefficient α_1 should be positive and significant different from zero and the DCA test is statistically significant. This implies that for the out-of-sample data, the model had turning point forecasting power.
* To be rejected at 5% significance level.

Table 8
Results of turning point forecasting capability of the combination models and the ARIMA models.

		Mainland China	Northeast Asia	Southeast Asia	Americas	Europe	Oceania
ARIMA	DCA	1.73	0.11	−0.50	1.03	−0.89	1.30
	α_1 (t-Value)	1.72	0.1	−0.46	0.97	−0.83	1.25
ARIMA + BPNN	DCA	2.51*	3.35*	1.46	0.98	1.95	2.49*
	α_1 (t-Value)	2.81*	4.77*	1.41	0.92	2.00*	2.77*
ARIMA + SVR	DCA	3.38*	3.87*	2.43*	2.93*	2.11*	2.49*
	α_1 (t-Value)	4.91*	7.66*	2.67*	3.62*	2.21*	2.77*

Null hypothesis of the existence of the α_1 is equal to zero. The slope coefficient α_1 should be positive and significant different from zero and the DCA test is statistically significant. This implies that for the out-of-sample data, the model had turning point forecasting power.
* To be rejected at 5% significance level.

statistically significant reaches 38.89% for the BPNN combination models and 75.00% for the SVR combination models. Furthermore, the individual models always show poor directional change detect ability, as evident from the insignificance of the DCA test and regression test. This implies that the combination models have good turning point forecasting ability and how the SVR combination models are superior to the other methods.

6. Conclusion

The author proposed to use the combination models that combine the linear model and the nonlinear model to predict tourism demand time series data. The results showed that the combination models are superior to the individual models for the test cases of tourism demand time series. The NMSE and MAPE were all the lowest for the combination models. The combination models also outperformed individual models in turning points forecasts. On the whole, the results made by the combination models will be superior to the individual models, in terms of both the prediction errors and directional change detects ability. Thus, the combination models could achieve better predictive performances and show promising results. More importantly, in practical applications for a given data set, unless an additional test is done in advance, there is no way of knowing whether the true model is factually linear or nonlinear. Thus, there arises the question: Under what circumstances should the nonlinear predictors or linear predictors is adopted? In the forecasting literature, it is a conventional fact that no single forecasting model is the best for all situations under all circumstances (Makridakis et al., 1982). Therefore, the "best" model in most real world forecasting situations should be the one that is credible and accurate for a long time horizon and thus users can have confidence to use the model repeatedly. Based on the empirical results, this paper suggests that forecast combination can achieve considerably better predictive performances and show promising results in directional change detects ability. As a conclusion, all of those results indicate that it would be much useful to combine the linear and nonlinear method in forecasting tourism demand. Although, no individual model consistently performs well in all situations (Witt & Song, 2002), the proposed combination models always improve the accuracy of forecasting and are typically a reliable forecasting tool in the tourism context. This conclusion is consistent with the study by Hibon and Evgeniou (2005), where they suggested that using a single method from a set of available methods is more risky than using a combination of method.

Besides, the empirical results clearly suggest that the SVR combination models are able to outperform other models used in this study. The results show that useful prediction could be made for Taiwanese outbound tourism demand without the use of extensive market data or knowledge. It also shows how a high forecasting accuracy and an excellent directional change detect ability could be achieved by the SVR combination models. The superior performance of the SVR combination models over the other models is due to the reason that SVR implement the structural risk minimization principle which minimizes an upper bound of the generalization error rather than minimizes the training error. This eventually leads to better generalization than the neural network which implements the empirical risk minimization principle.

References

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly, 20*, 451–468.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting – A case study of Durban, South Africa. *Tourism Management, 22*, 403–409.

Chang, C. C., & Lin, C. J. (2001). *LIBSVM: A library for support vector machines*. National Taiwan University, Department of Computer Science and Information Engineering. <http://www.csie.edu.tw/~cjlin/papers/libsvm.pdf> Retrieved 20.05.04.

Chen, K. Y., & Wang, C. H. (2005). Using genetic algorithms to optimize support vector regression in tourism demand forecasting. *Journal of Outdoor Recreation Study, 18*(1), 47–72.

Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management, 28*, 215–226.

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management, 24*, 323–330.

Chu, F. L. (2004). Forecasting tourism demand: A cubic polynomial approach. *Tourism Management, 25*, 209–218.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography with discussion. *International Journal of Forecasting, 5*, 559–608.

Cumby, R. E., & Modest, D. M. (1987). Testing for market timing ability: A framework for forecast evaluation. *Journal of Financial Economics, 19*(1), 169–189.

Davies, N., Petruccelli, J. D., & Pemberton, J. (1988). An automatic procedure for identification, estimation and forecasting self exciting threshold autoregressive models. *Statistician, 37*, 199–204.

Ginzburg, I., & Horn, D. (1994). Combined neural networks for time series analysis. *Advances in Neural Information Processing Systems, 6*, 224–231.

Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management, 23*(5), 499–510.

Hibon, M., & Evgeniou, T. (2005). A simple procedure for reliability of repairable systems. *International Journal of Forecasting, 21*, 15–24.

Huang, H. G., Hwang, R. C., & Hsieh, J. G. (2002). A new artificial intelligent peak power load forecaster based on non-fixed neural networks. *International Journal of Electrical Power and Energy Systems, 24*(3), 245–250.

Kim, J. H., & Ngo, M. T. (2001). Modeling and forecasting monthly airline passenger flows among three major Australian cities. *Tourism Economics, 7*, 397–412.

Kon, S. C., & Turner, W. L. (2005). Neural network forecasting of tourism demand. *Tourism Economics, 11*, 301–328.

Law, R. (2000a). Demand for hotel spending by visitors to Hong Kong: A study of various forecasting techniques. *Journal of Hospitality and Leisure Marketing, 6*, 17–29.

Law, R. (2000b). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management, 21*(4), 331–340.

Law, R. (2004). Initially testing an improved extrapolative hotel room occupancy rate forecasting technique. *Journal of Travel & Tourism Marketing, 16*, 71–77.

Law, R., & Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management, 20*(1), 89–97.

Lawerence, M. J., Edmundson, R. H., & O'Connor, M. J. (1986). The accuracy of combining judgmental and statically forecasts. *Management Science, 32*, 1521–1532.

Luxhoj, J. T., Riis, J. O., & Stensballe, B. (1996). A hybrid econometric-neural network modeling approach for sales forecasting. *International Journal of Production Economics, 43*, 175–192.

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting, 5*, 601–603.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., et al. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting, 1*, 111–153.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecast. *Management Science, 29*, 987–996.

Oh, C. O., & Morzuch, B. J. (2005). Evaluating time-series models to forecast the demand for tourism in Singapore: Comparing within sample and post-sample results. *Journal of Travel Research, 43*, 404–413.

Pai, P. F., & Hong, W. C. (2005). An improved neural network model in forecasting arrivals. *Annals of Tourism Research, 32*, 1138–1141.

Pai, P. F., Hong, W. C., Chang, P. T., & Chen, C. T. (2006). The application of support vector machines to forecast tourist arrivals in Barbados: An empirical study. *International Journal of Management, 23*, 375–385.

Palmer, A., Montano, J. J., & Sese, A. (2006). Designing an artificial neural network for forecasting tourism time series. *Tourism Management, 27*, 781–790.

Pelikan, E., de Groot, C., & Wurtz, D. (1992). Power consumption in West-Bohemia: Improved forecasts with decorrelating connectionist networks. *Neural Network World, 2*(6), 701–712.

Pemberton, J. (1989). *Forecast accuracy of nonlinear time series models*. Technical Report 77. Statistics Research Center, Graduate School of Business, University of Chicago.

Peng, K. L., Wu, C. H., & Goo, Y. J. (2004). The development of a new statistical technique for relating financial information to stock market returns. *International Journal of Management, 21*(4), 492–505.

Pesaran, M. H., & Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business and Economics Statistics, 10*(4), 461–465.

Qu, H., & Zhang, H. Q. (1996). Projecting international tourist arrivals in East Asia and the Pacific to the year 2005. *Journal of Travel Research, 35*(1), 27–34.

Reid, D. J. (1968). Combining three estimates of gross domestic product. *Economica, 35*, 431–444.

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels*. Cambridge: MIT Press.

Tay, Francis E. H., & Cao, Lijuan (2001). Application of support vector machines in financial time series forecasting. *Omega, 29*(4), 309–317.

Tseng, F. M., Yu, H. C., & Tzeng, G. H. (2002). Combining neural network with seasonal time series ARIMA model. *Technological Forecasting and Social Change, 69*(1), 71–87.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Wedding, D. K., II, & Cios, K. J. (1996). Time series forecasting by combining RBF networks, certainty factors, and the Box–Jenkins model. *Neurocomputing, 10*, 149–168.

Witt, S. F., & Song, H. (2002). Forecasting tourism flows. In A. Lockwood, & S. Medlik (Eds.), *Tourism and hospitality in the 21st century* (pp. 106–118).

Wong, K. K. F., Song, H., Witt, S. F., & Wu, D. C. (2007). Tourism forecasting: To combine or not to combine? *Tourism Management, 28*, 1068–1078.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175.