KKU Res.j. 2016; 21(1) : 37-54
http://resjournal.kku.ac.th

A Hybrid ARIMA and RBF Neural Network Model for Tourist Quantity Forecasting : A Case Study for Chiangmai Province

Rati Wongsathan[1] and Wararat Jaroenwiriyapap[2]*

[1]Department of Electrical Engineering, Faculty of Engineering, North-Chiangmai University
[2]Department of Tourism and Hospitality Management, Faculty of Business Administration, North-Chiangmai University
**Correspondent author: rati@northcm.ac.th, rati1003@gmail.com*

Abstract

Applications of a single model may not be able to capture different data patterns well enough, especially in the tourist forecast problem which is often complex in nature. An autoregressive integrated moving average (ARIMA) is a famous linear model while an artificial neural network (ANN) is a promising alternative to a traditional linear method. The ARIMA model may not be adequate for nonlinear problems while ANN can well reveal the correlation of nonlinear patterns. However, overfitting due to a learning process is the main disadvantage of ANN as well as being trapped in a local optimum for parameters optimization. To improve the forecast performance of both ARIMA and ANN for high accuracy, the two hybridization models, i.e. hybrid ARIMA-RBFNN model and hybrid RBFNN-ARIMA model are employed to examine the Chiangmai's tourist time series data. Statistics test and parameter designed experiments were used to optimize these models and the sum-square of error (SSE) was used to indicate their performances. In this case study, the hybrid RBFNN-ARIMA model has proved that the RBFNN can priori capture the non-stationary non-linear component while the fully linearly stationary residuals were accurately predicted by ARIMA. The experimental results demonstrated that the hybrid RBFNN-ARIMA model outperformed 42% by averaging over the hybrid ARIMA-RBFNN model, an improvement of hybrid ARIMA-RBFNN model, RBFNN model, and ARIMA model.

Keywords: *Radial Basis Function Neural Networks; ARIMA; Hybrid RBFNN-ARIMA; Hybrid ARIMA-RBFNN.*

1. Introduction

In the past few decades, the number of domestic and international tourism has grown rapidly especially in the South East Asia. According to the research data from the World Travel & Tourism Council (WTTC) [1] in 2014, the money spent by the tourist industry increases 10% within this region in 2013. Focus on Chiangmai Province in the northern of Thailand where has a charming in natural, beautiful scenery and fascinating cultural with 720 years old city, is one of the famous and very attractive place for many tourist around the world. It was the first appearance for the World's Best City in year 2005 at the 5th place and 2nd place for Asia's Best City, at the 5th in 2006-2009 and 2nd in 2010 by the readers' survey in Travel and Leisure magazine. The number of tourist is constant growth in this city for various reasons. Development of Chiangmai's tourism may provides an investment in many aspects, including traffic infrastructure, airport, public transport facilities, tourist hotels, restaurants, amusement parks, souvenir shops, shopping mall, etc, which requires a long period of planning. In such cases, the correctness and valid tourist forecast model becomes very significant for effectively future planning. Since, it can reduce a risk for investment and avoid an inadequate construction or waste due to an excessive construction. A properly forecast model would help to sustainable and stable develop for Chiangmai tourism.

Generally, the number of the tourist is ordinarily restricted by many factors e.g. geography, tourism resource, security, infrastructures, facilities, advertising and public relation, image, political situation, epidemic disease, etc. Cooperating all these factors to formulate the forecast model is an inconvenience task while using the historical data to forecast the future data is the most popular in the most research. The universally adopted and widely used for the forecast time series model is an ARIMA model which is usually found in many research due to its statistical properties as well as the well-known Box-Jenkins methodology. The most disadvantage of this model is that it can't capture the nonlinear patterns. In this aspect, a tourist model which has good nonlinear and complex characteristic may not an effective way. Mitigates this problem, an ANN model was introduced as an efficient tool for time series forecast which can well reveal the correlation of nonlinear time series in a delay state space. A significant advantage of an ANN is that one need not have a well-defined physical relationship for systematically converting an input to an output (Hornik et al., 1989). However, the overfitting problem due to a learning strategy is the main disadvantage of this method as well as a local trapped of parameters optimization. To improve the predictive performance of both ARIMA and ANN for high accuracy result, the theoretical and empirical findings have suggested an effective way by combining different models. In this work, the combining strategy of ARIMA and RBFNN is done by the different order of combination for three types of the hybrid model as an ARIMA-RBFNN model, an improvement ARIMA-RBFNN and the RBFNN-ARIMA model. The performance of these hybrid forecast models will be also assessed relative to an ARIMA and the RBFNN model. The aim of this study is to find out the most accuracy model for these forecast models. The rest of the paper is organized as follows. Next section presents

the literature review. Section 3 describes the basic concepts and the modeling approach of an ARIMA and the RBFNN and introduces the three hybrid models. Section 4 presents the construction of the forecast models and gives the forecast results. The last section contains the concluding remarks.

2. Literature Review

An ARIMA and an ANN model are the most commonly used to apply in the forecast area. In the tourist quantity forecast, however, most of the work tends to use a single type method, either an ARIMA model or an ANN model. For a decade, in a breviation, Shan (2002), Huang and Min (2002), Moosa (2005), Coshall (2005), Andrea (2010), Murat (2014), Albert and et al (2015), and etc have already analyzed the forecast by using an ARIMA model. Kon and Turner (2005), Palmer and et al (2006), YinHao and et al (2007), Oscar and Salvador (2014), and etc have also emerged an ANN in this forecast area. The comparison study for valid prediction whether it is an ARIMA or an ANN model was first done by Burger and et al (2001) which was found that the error rate established by an ANN model is the lowest. The comparative between ARIMA and ANN in tourist forecast research is studied by Fernandes and et al (2008). The hybrid ARIMA and ANN has been investigated for various design aspect in this area by Aslanagun and et al (2007), followed by XiPing (2012), and Dursun and Mammadagha (2014). However, the model that discovers the complex pattern of the tourist data nature, whether it is ANN-ARIMA or ARIMA-ANN model, has not yet investigated in any paper. Furthermore, it has a rarely research for forecasting the quantity of tourist in Thailand by the hybrid between an ARIMA and an ANN model.

3. Research Methods

In this work, five forecast methods including an ARIMA, the RBFNN, the hybrid ARIMA-RBFNN, the improvement of hybrid ARIMA-RBFNN and the hybrid RBFNN-ARIMA model were studied. A theory and an implementation of them were be detailed in the next sub-section respectively.

3.1 An ARIMA Model

An ARIMA model has been one of the most popular approaches for time series forecast introduced by Box and Jenkins (1976). It consists of three parts i.e. auto regression AR(p), moving average MA(q) and differencing in order to strip off the integration of the series (d) and then form ARIMA (p, d, q). This linear model is as follows

$$\Delta^d Y_t = \delta+\varphi_1 Y_{t-1}+\varphi_2 Y_{t-2}+\ldots+\varphi_p Y_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\ldots-\theta_q\varepsilon_{t-q}. \quad (1)$$

Where $\Delta=(1\text{-}B)$, B refers to the backward shift operator for $B(Y_t)=Y_{t-1}$, Y_t is the observation data at time t, δ is the constant, φ, φ_2, …, φ_p are the autoregressive parameter, ε_t is the randomly error at time t and $\sim N(0, \sigma^2)$, and Θ_1, Θ_2, …, Θ_q are the moving average parameters. Equation (1) can reduce to

$$\Phi(B)\Delta^d Y_t = \delta+\Theta(B)\varepsilon_t, \quad (2)$$

where

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p \quad (3)$$

$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q \quad (4)$$

and $\Phi(B)$ is the autoregressive polynomial of order p and $\Theta(B)$ is the moving average polynomial order q. A practical approach to building ARIMA model includes three basically iterative steps i.e. identification, parameter estimation, and diagnostic checking.

For the model identification, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the sample data are the basic tools to identify the order q of MA and order p of AR. In the identification step, the differencing and power of data transformation are often required to make the time series data stationary when the observed time series present trend and hetero scedasticity. The set of ϕ and Θ parameters are estimated after tentative model is identified such that an overall measure of error is minimized. This can be accomplished by using a nonlinear optimization procedure. Diagnostic checking by the several statistics assumption of the residuals such as Box-Pierce Chi-Square test or the correlation of the residual plot is used to examine at the last step. An ARIMA model is not sufficient if there are still linear correlations remain in the residuals [2]. If the model is not adequate, a new tentative or candidate model will be replaced by the three steps mentioned above until a satisfactory model is finally selected. Due to the seasonal of tourist data nature, ARIMA lacks of the efficiency to fit the model while the performance of NN is outstanding than ARIMA relative to various tourism time series models [3]. By this approach, both time series for Thai and foreigner tourist data were preliminary examined on the stationary by ACF and PACF and unit root test by augmented Dickey-Fuller (ADF) test and K wiatkowski-Phillips-Schmidt-Shin (KPSS) test. By the graphic and hypothesis test, both time series data are non-stationary affected with trend and intercept. The first differencing was applied to the time series data to generate 1^{st} order difference time series (Z_t). After Z_t is identified as the stationary, the best fit parameters of an ARIMA model were estimated according to its order p and q by PACF and ACF considering.

3.2 A Radial Basis Function Neural Networks (RBFNN)

ANN is an imitation of simulating human brain cells using a computational model. By learning the mapping data, ANN model can be built without explicit formulating the possible relationship that exist between variables. It is regarded as multivariate, nonlinear and nonparametric method which can well reveal the correlation of nonlinear time series in delay state space. The Kolmogorov continuation theorem guarantees the prediction feasibility of the ANN for time series from the view of mathematics [4]. Various ANNs in forecast field are presented in [5], a feed-forward NN model was used to forecast Japanese tourism demand [6] while RBFNN was applied in forecast tourist quantity in Hainan province of China [7]. In this work, the RBFNN is selected to adopt in the time series forecast instead of the multi-layered perceptron (MLP) which is quite a slow learning and usually trapped in the local minima. Due to their universal approximation,

compact topology, and faster learning speed make RBFNN have attracted considerable attention. The RBFNN is a three-layer feed-forward neural network which consists of an input layer, a hidden layer with radial activation function neurons commonly Gaussian function and an output layer with linear neurons, and is shown in Fig. 1.

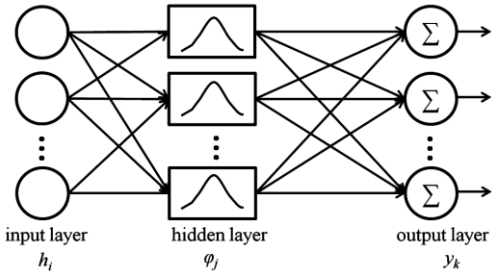

Figure 1. The network structure of the RBFNN.

The output of RBFNN is the weighted summation of each hidden layer neuron's output which can be expressed as

$$\hat{y}_j = \sum_{i=1}^{I} w_{ij}\varphi\left(\left\|\mathbf{x}-\mathbf{c}_i\right\|\right)+\beta_j, \qquad (5)$$

where the radial basis function (j) refers as

$$\varphi(r)=\exp(-\left\|\mathbf{x}-\mathbf{c}_i\right\|^2/2\alpha_i^2), \qquad (6)$$

and I is the number of node in the hidden layer by i Î {1, 2, …, I}, J is the number of node in an output layer by j Î {1, 2, …, J}, w_{ij} is the weight between i^{th} node in the hidden layer and j^{th} node in an output layer, a_i and $\mathbf{c}_i$ is the spread and center of i^{th} node in hidden layer, $\mathbf{x}$ is the data input vector, b_j is the bias j^{th} node in output layer. RBFNN learning process is divided into two stages. At the first stage, the unsupervised learning process such as K-mean algorithm [8], orthogonal least squares (OLS) algorithm [9], etc are used to solve the center and variance of Gaussian function. In the second stage, the supervised learning process such as gradient descent algorithm, least square algorithm , etc are used to adjust the weight between hidden and output layer. For time series data, input variable of RBFNN is the past quantity data of several months. The number of neuron of the input layer corresponding to the dimension of input vectors is determined and selected by the design experiment. A large amount of input node makes the learning is relatively difficult whereas a bit of input node can not reflect the value of the correlation between the precursors and forecast. The number of pattern learning of the observation data (y_t) of n samples which are fed into the r input node and one output node of RBFNN generates n–r types e.g. the 1^{st} pattern consists of data input [y_1, y_2, …, y_r] and data output is y_{r+1}, the 2^{nd} pattern consists of data input [y_2, y_3, …, y_{r+1}] and data output is y_{r+2} and the last pattern consists of data input [y_{t-r}, y_{t-r+1}, …, y_{t-1}] and output data is expressed as follows

$$y_t = \sum_{i=1}^{Q} w_i a_i + \beta_0, \qquad (7)$$

where

$$a_i = \exp\left(-\sum_{j=1}^{r} \left(y_{t-j} - \hat{y}_{t-j} \right)^2 \Big/ \alpha_i^2 \right), \qquad (8)$$

and y_{t-j} is an input for j^{th} node for an input layer, $\hat{y}_{t-j}$ is the center which is the preceding forecast value of j^{th} input, a_i is the spread parameter of i^{th} node in hidden layer, w_i is the weight between i^{th} node in the hidden layer and the output layer, and b_0 is the bias of the output node.

For an overfitting problem of ANN model, the model memorized only the training data rather than learning to generalize from trend. To avoid an overfitting, many techniques e.g. cross-validation [10], regularization [11], early stopping [11], pruning [11], Bayesian priors on parameters [12] were applied. In this paper, however, multiple rounds of cross-validation method is applied i.e. the training and test set data are performed using different partitions then select the best results from the validate test sample to specify the suitable training and test samples ratio.

3.3 The Hybrid ARIMA-RBFNN Model

Since, there is not a universal model that has successfully achieved both of linear and nonlinear domains. ARIMA model may not be adequate the nonlinear problem while RBFNN have yielded mixed results. Hybrid methodology which has both linear and nonlinear modeling capabilities can be a wise for prediction the real problem. It can reduce the unstable and trace the change of data pattern that frequently occurred in such time series data [13]. In this research, the tourist series data is assumed to be a composition between a linear autocorrelation structure and nonlinear component as,

$$Y_t = L_t + N_t + e_t. \qquad (9)$$

Where L_t denotes the linear component, N_t denotes the nonlinear component and e_t is the residuals at time t. There are two hybrid types corresponding to the order of combination: the hybrid ARIMA-RBFNN model which is discussed in this section, an improvement of hybrid ARIMA-RBFNN model which is detailed in section 3.4 and hybrid of RBFNN-ARIMA model which is detailed in section 3.5.

For an ARIMA model, the diagnostic check of the residuals step is not able to detect any nonlinear patterns in the time series data. Even if the ARIMA model has passed the diagnostic check for the residuals, it's not guarantee the sufficient condition since the limitation of nonlinear pattern still not reveal. Therefore, the residuals can be modeled by the RBFNN to discover nonlinear pattern which is composed to the hybrid ARIMA-RBFNN. This kind of method observed that both linear and nonlinear relationship still existed in both residuals (e_t) and original data (z_t). In order to yield more accurate results according to the proposed hybrid model by [14], the hybrid ARIMA-RBFNN model considered the time series as a nonlinear function of several past observations and random errors expressed as

$$\hat{Y}_t = f[(z_{t-1}, z_{t-2}, ..., z_{t-m}), (e_{t-1}, e_{t-2}, ..., e_{t-n})], \qquad (10)$$

where $z_t = (1\text{-B})^d(y_t)$ is the difference of time series data, and f is a nonlinear function determined by RBFNN, e_t is the residuals at time t, and m is the number

input node of z_t, and n is the number of the input node of e_t. The model configuration is shown in Fig. 2.

This hybrid model has the structure similar to the RBFNN model except for the joint between the residual (e_t) from the ARIMA model with the first difference of ordinary data (z_t) as the input. The residuals still generated from the ARIMA(p, d, q) in section 3.1 which used as the partial input altogether with z_t. The number of m input node, for the z_t and n input node for e_t are determined by the design experiment as well as the number q hidden node of RBFNN which denoted hybrid of ARIMA-RBFNN ([m, n], q).

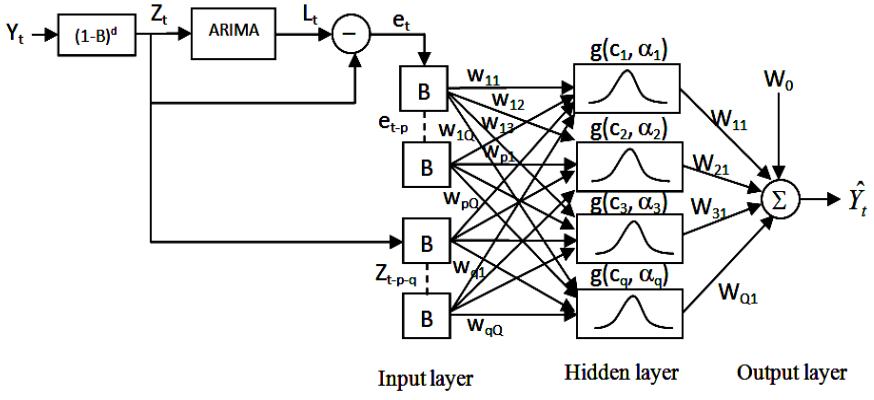

Figure 2. The hybrid ARIMA-RBFNN model.

At the first stage, the ARIMA model generated the residuals (e_t) which altogether entered with the original data (z_t) into the RBFNN model in the second stage. The mapping between input and output value by RBFNN is expressed as,

$$y_t = w_0 + \sum_{j=1}^{Q} w_j \cdot g\left(w_{0,j} + \sum_{i=1}^{p} w_{i,j} \cdot z_{t-i} + \sum_{i=p+1}^{p+q} w_{i,j} \cdot e_{t+p-i} \right) + \varepsilon_t, \quad (10)$$

where Q is the number of node in the hidden layer, $w_{i,j}$ is the weight between input layer and hidden layer (i=0, 1, 2, …, p+q,j=1, 2, …, Q), w_j is the weight between hidden layer and output layer (j,=0, 1, 2, …, Q), and p is the number of input node e_t ,q is the number of input node z_t. All parameters p, q, and Q will be determined in the design process of RBFNN.

3.4 An Improvement of Hybrid ARIMA-RBFNN Model

The configuration of the improvement of hybrid ARIMA-RBFNN model which is denoted I-ARIMA-RBFNN is shown in Fig. 3. For this hybrid model, an ARIMA model first captures the L_t term, then the residuals will remain N_t term. The residuals at time t from ARIMA model is referred as

$$e_t = Y_t - YF_t, \quad (12)$$

where YF_t is the output of the ARIMA model at time t. With p input nodes, the RBFNN model for residuals is as follows,

$$N_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-p}) + u_t, \quad (13)$$

where f is the nonlinear function determined by the RBFNN, e_t is the residuals generated by an ARIMA model at time t, and u_t is the randomly error. The mapping between input and output of RBFNN is expressed as,

$$N_t = w_0 + \sum_{j=1}^{Q} w_j g\left(w_{0,j} + \sum_{i=1}^{p} w_{i,j} e_{t-i} \right) + u_t, \quad (14)$$

where Q is the number of node in the hidden layer, $w_{i,j}$ is the weight between input layer and hidden layer (i=0, 1, 2, …, p, j=1, 2, …, Q), w_j is the weight between hidden layer and output layer, and p and Q are the integer which are determined by the design experiment process.

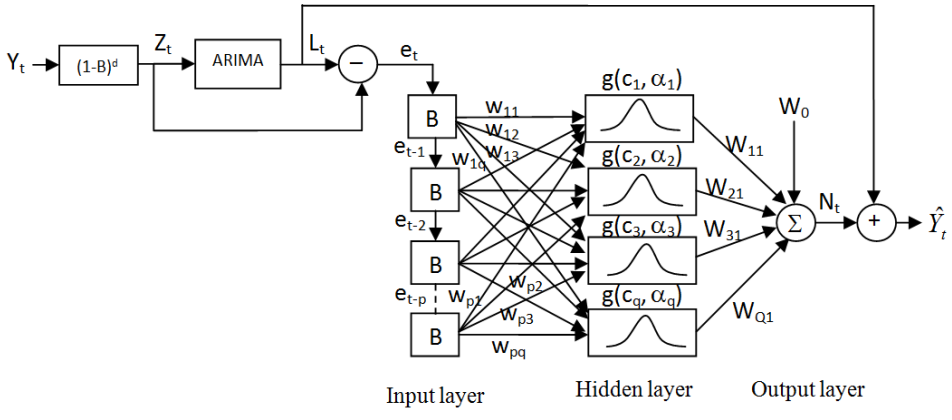

Figure 3. The structure of an improvement of hybrid ARIMA-RBFNN model.

3.5 The Hybrid of RBFNN-ARIMA Model

The diagram of the hybrid RBFNN-ARIMA model is shown in Fig. 4, the designed RBFNN of both Thai and foreigner tourist data in section 3.2 were adopted to forecast the solution N_t in the first stage. The residuals will be treated as the linear model (L_t) in the second stage which is statistically investigated and modeled by an ARIMA model. The unit root test of the residuals performed by ADF test for stationary checking. In order to estimate the tentative ARIMA model, the ACF and PACF will be used to evaluate the order in MA and AR model respectively.

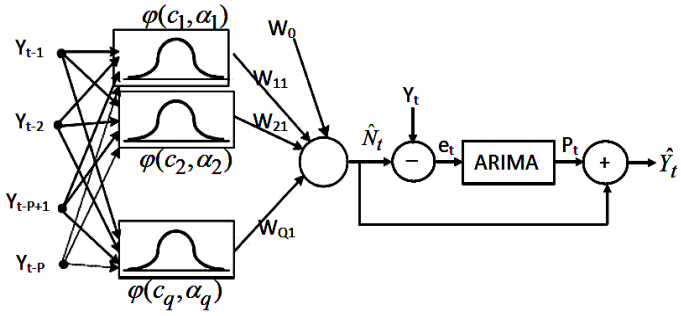

Figure 4. The structure of hybrid RBFNN-ARIMA model.

From the Figure 4, the forecast value at time t for P input node, Q hidden node and one output node of RBFNN can be expressed as

$$NF_t = w_0 + \sum_{j=1}^{Q} w_j \cdot g\left(w_{0,j} + \sum_{i=1}^{P} w_{i,j} \cdot y_{t-i} \right). \tag{15}$$

The residual at time t from the nonlinear model, then follows by,

$$e_t = Y_t - NF_t, \tag{16}$$

where NF_t refers to the forecast value for time t in (15). This linearly residuals will be modeled by the ARIMA model, then linear relationships can be discovered. The residual forecast is as follows,

$$\Phi(B)\Delta^d e_t = \mathrm{d} + \Theta(B)\mathrm{e}_t \tag{17}$$

Once, the linear correlation will be removed from the residuals for sufficient condition of ARIMA model. Combining of (14) and (17), then the forecast of the hybrid RBFNN-ARIMA model will be hold.

The hybrid model utilizes the unique feature and strength of RBFNN model as well as ARIMA model in determining complex patterns. It could be advantageous to model linear and nonlinear patterns separately. The test will clearly evident by the experimental results of the forecast in the next section.

4. Empirical Results and Analysis

In the experiment, the sample data is categorized in two cases of monthly time series data i.e. Thai and Foreigner tourist. Due to the difference in each travel behavior, considering the tourist into two groups gives more useful detail in the future planning. The number of the tourist continues growing as was shown in Fig. 5 by the recorded time series data [15] during 2008-2013 except only in 2010 which has occurred the critical in political. A total 72 months data were used to examine in this work, the preceding 60 data during January 2008–December 2012 were used to train for the forecast model while the 24 data in January 2013- December 2014 were used to test the validation. The forecast model established in this work includes an ARIMA, the RBFNN, the hybrid

ARIMA-RBFNN, an improvement of hybrid ARIMA-RBFNN and the hybrid RBFNN-ARIMA model will be determined the best one performance. The experimental results and analysis are discussed in the following sub-section.

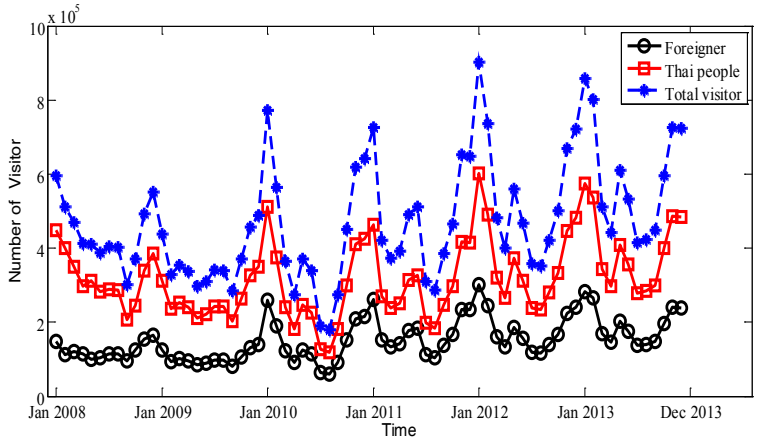

Figure 5. A Time series of Thai and foreigner tourist data in Chiangmai during 2008-2013.

4.1 An ARIMA Model's Design and Forecast

From the test, the graph of the ACF died off smoothly at a geometric rate after 1 time lag and the PACF declined geometrically after 1 time lag for Thai tourist data. For foreigner tourist data, the ACF died off smoothly at a geometric rate after 1 time lag and the PACF declined geometrically after 4 time lag. Then, the tentative ARIMA model was introduced as ARIMA(1,1,2) for Thai tourist data and ARIMA(4,1,2) for foreign tourist data. However, after diagnostic checking and model selection by the criteria i.e. adjust R^2, Akaike Information Criterion (AIC), and Schwarz's Bayesian Information Criterion (SBC) for 10 candidate models which summarizes in Table 1.

Table 1. Comparison results of ARIMA models.

Candidate model	Thai tourist's statistics			Foreigner tourist's statistics		
	Adjusted R^2	AIC	SBC	Adjusted R^2	AIC	SBC
ARIMA(1,1,0)	0.016	25.44	25.51	0.001	24.12	24.18
ARIMA(1,1,1)	0.029	25.44	25.55	0.006	24.13	24.22
ARIMA(1,1,2)	0.245	25.20	25.33	0.225	23.89	24.02
ARIMA(1,1,3)	0.250	25.21	25.37	0.233	23.89	24.05
ARIMA(0,1,1)	0.037	25.41	25.47	0.012	24.10	24.17
ARIMA(0,1,2)	0.053	25.41	25.50	0.155	23.96	24.05
ARIMA(0,1,3)	0.257	25.18	25.30	0.241	23.86	23.99
ARIMA(2,1,0)	0.076	25.40	25.50	0.045	24.10	24.20
ARIMA(2,1,1)	0.287	25.15	25.28	0.272	23.84	23.97
ARIMA(2,1,2)	0.275	25.19	25.35	0.246	23.89	24.05

Finally, ARIMA(2,1,1) is proved as suitable model for both Thai and foreigner tourist data which has the most greater adjusted R^2 value and smaller AIC and SBC value than other candidate models. At 99% statistic test of confidence interval level was shown in Table 2 for both tourist data, further the standard error is less twice than the coefficient of parameters value and all P-values is less than 0.01. It can conclude that the selected coefficient of parameters has the statistical significance. The ARIMA model for Thai and foreigner tourist data with these coefficients can be expressed in (18) and (19) respectively as

$$YT_t - YT_{t-1} = 2435.43 + 0.94(1-B)YT_{t-1} - 0.43(1-B)YT_{t-2} - 0.99B\mathrm{e}_{Tt}, \quad (18)$$

$$YF_t - YF_{t-1} = 1706.86 + 0.85(1-B)YF_{t-1} - 0.34(1-B)YF_{t-2} - 0.99B\mathrm{e}_{Ft}. \quad (19)$$

Table 2. Statistics test for the parameter's coefficient for an ARIMA (2, 1, 1) model.

Thai torist					Foreigner tourist				
Variable	Coefficient	Std. Error	t-Statistic	Prob.	Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2435.43	585.82	4.16	0.0001	C	1706.86	354.47	4.81	0.0000
AR(1)	0.95	0.12	8.05	0.0000	AR(1)	0.85	0.11	7.80	0.0000
AR(2)	-0.43	0.11	-3.84	0.0003	AR(2)	-0.34	0.12	-2.87	0.0056
MA(1)	-0.99	0.043	-22.85	0.0000	MA(1)	-0.99	0.03	-33.46	0.0000

The forecast results by the ARIMA (2, 1, 1) model by (18) and (19) are shown in Fig. 6 (left) and (right), respectively. The forecast is quite worst corresponds with the adjust R^2 statistics value which less than 30% for all ARIMA models in Table 1. Especially, the error is occurred in Thai forecasting rather than in Foreigner forecasting implies more complex pattern existed in Thai tourist data.

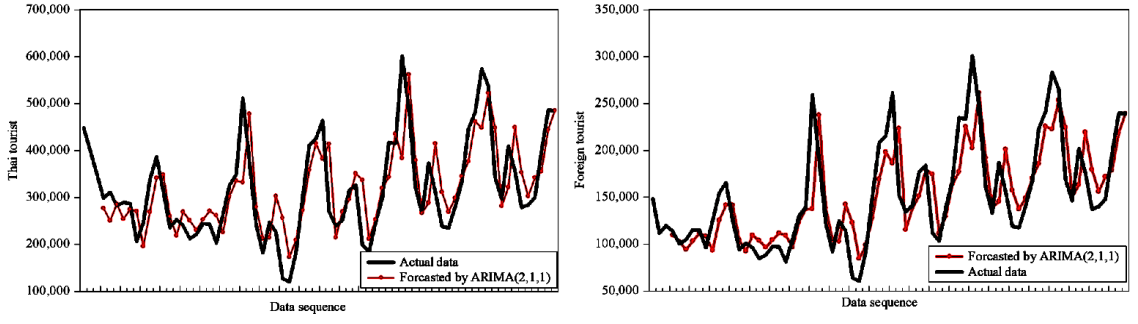

Figure 6. Forecast results by ARIMA(2,1,1) model for (left) Thai tourist and (right) foreigner tourist data.

4.2 The RBFNN Model's design and Forecast

In the design experiment of generalized RBFNN, several factors including number of input node (in this case is the time lag length), number hidden node, and center and spread parameters are properly selected for accuracy and rapidly convergence of solution. To avoid the over fitting problem, the train and test samples ratio will be also first specified. The designed parameters by using RMSE performance in Fig. 7 (left) and (right) have found that the train:test ratio, number of input node, number of hidden node and spread parameter are 80:20, 7, 10, and 1.4 for Thai tourist data which denoted by RBFNN(7, 9) and 80:20, 7, 9 and 1.0 for the foreigner tourist data which denoted by RBFNN(7, 10). The forecast results and post residuals by selected RBFNN are shown in Fig.8 (top left) and (top right) for Thai and foreigner tourist, respectively. The forecast by this approach is still worst, however the residuals reflect a stationary time series with values vibrate around zero. Further, the forecast error is also occurred in Thai tourist data rather than Foreign tourist data like ARIMA model resulted in Fig. 8(bottom left) and (bottom right), respectively.

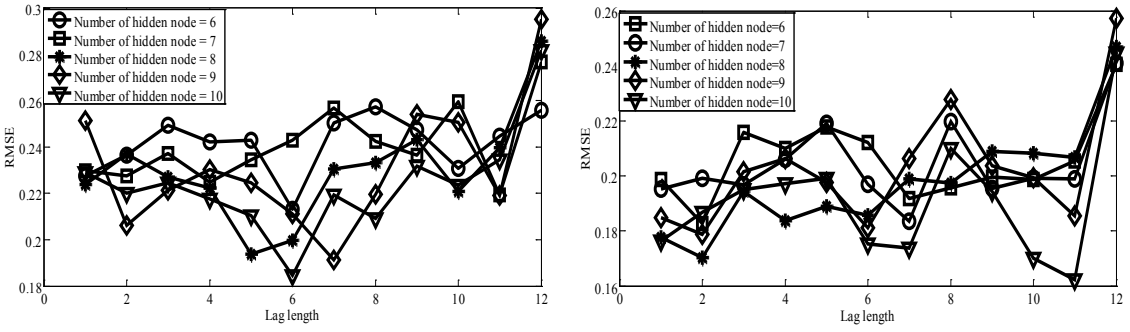

Figure 7. Selection input and hidden node of RBFNN for (left) Thai and (right) foreigner tourist data.

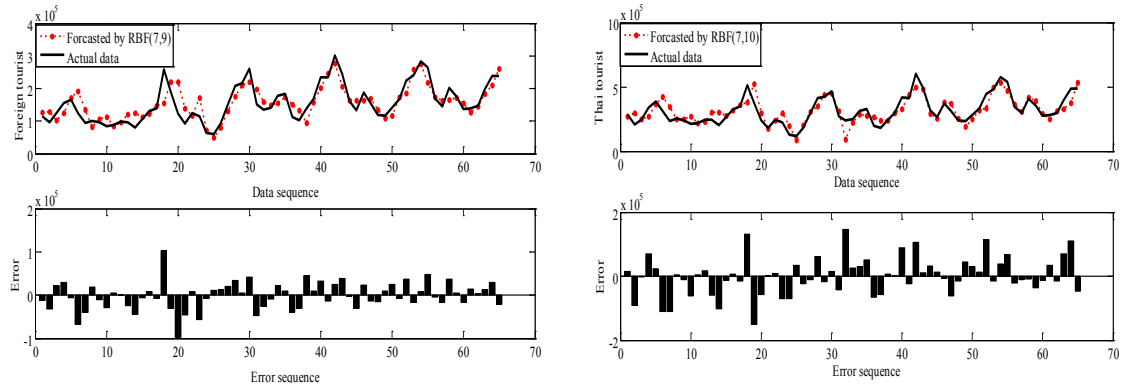

Figure 8. Forecast and residuals results by RBFNN (7, 10) for (left) Thai and (right) foreigner tourist data.

4.3 The hybrid ARIMA-RBFNN Model's Design and Forecast

This hybrid model, the residuals generated from the ARIMA(2, 1, 1) in section 4.1 are used as the partial input altogether with z_t. The number of m input node, for the z_t and n input node for e_t are determined by the design experiment as well as the number q hidden node of RBFNN which denoted hybrid of ARIMA-RBFNN([m, n], q). To design parameter m, n and q, the 64 combinations of (m, n ,q) were set for the test where m= [1 5 10 15], n=[1 5 10 15] and q=[1 5 10 15] by fixing φ equals 1. From the results showed in Fig. 9 (left), it found that combination number 22 (5, 5, 5), 23 (5, 5, 10), and 24 (5, 5, 15) are suited for Thai tourist data and combination number 18 (1, 5, 5), 19 (1, 5, 10) and 20 (1, 5, 15) are suited for foreigner tourist data. Then, the hybrid ARIMA-RBFNN([5, 5], q) and ([1, 5], q) model are initially optimized structure for Thai and foreigner tourist data respectively.

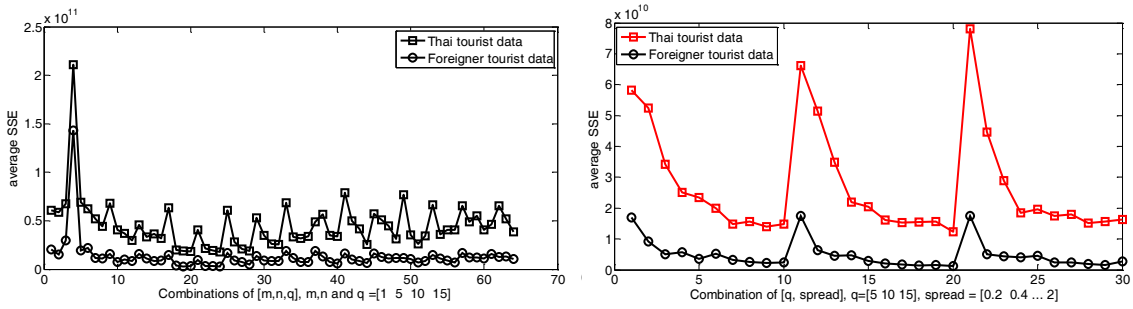

Figure 9. The design experiment for (left) data input node and residual node number (m,n) and hidden node number (q) and of (right) hidden node (q) and spread parameter (φ) for RBFNN.

Next step, number of hidden node and φ value were determined by 30 combinations of (q,φ) where q equal to 5, 10, and 15 and φ equals to 0.2, 0.4, …, 2. From the results shown in Fig. 9 (right), it has shown that the combination number 10(5, 2) is suited for Thai and foreigner tourist data. Furthermore, it can observe that the large φ *value* at q equals 5, 10 and 15 is provided the minimum of SSE for both time series data. The forecast results by the hybrid ARIMA-RBFNN([5, 5], 5) and the hybrid ARIMA-RBFNN([1, 5], 5) for Thai and foreigner tourist data are shown in Fig. 10 (left) and (right), respectively. The forecast results for both time data series quite well than ARIMA model and RBFNN model. The error forecast is still occurred in Thai tourist data rather than Foreign tourist data.

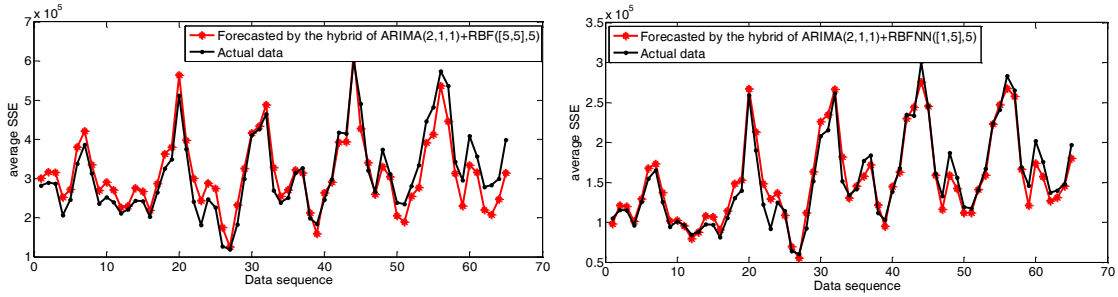

Figure 10. Forecast results by the hybrid ARIMA-RBFNN for (left) Thai and (right) Foreigner tourist data.

4.4 An Improvement of Hybrid ARIMA-RBFNN Model's Design and Forecast

For this hybrid model, in the first step, ARIMA model has priori set to forecast and follows by post residuals forecast by RBFNN according to the architecture depicted in Fig. 3.

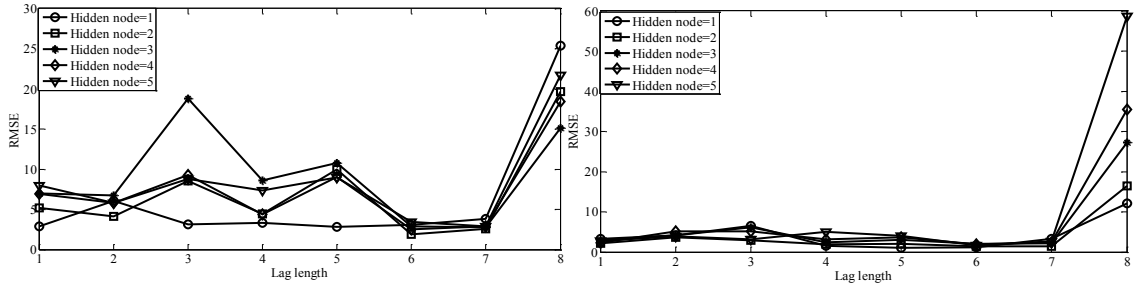

Figure 11. Selection number of input node and hidden node for RBFNN in an improvement of the hybrid ARIMA-RBFNN of (left) Thai tourist data and (right) foreigner tourist data.

An ARIMA(2,1,1) model expressed by (18) and (19) for Thai and foreigner tourist data entranced the input data and then generated one output (post residual). The lag residuals enter as the input to the RBFNN which generate the nonlinear solution by (14). The design experiment mentioned in section 4.3 for optimized RBFNN (p, q) structure has resulted the optimized RBFNN (6, 1) and RBFNN (7, 1) for Thai and foreigner tourist data which are shown in Fig. 11 (left) and (right), respectively. The forecast results by an I-ARIMA(2, 1, 1)-RBFNN(6, 1) and an I-ARIMA(2, 1, 1)-RBFNN(7, 1) model obtained are shown in Fig. 12 (left) and (right) for Thai tourist and foreigner tourist data, respectively by comparing with the hybrid ARIMA (2, 1, 1)-RBFNN ([5, 5], 5) and ARIMA (2, 1, 1)-RBFNN ([1, 5], 5) model mentioned in section 4.3. By the simple explore of the comparison test with them, an I-ARIMA-RBFNN has more the accuracy by keep the forecast closer to the actual data than the hybrid ARIMA-RBFNN.

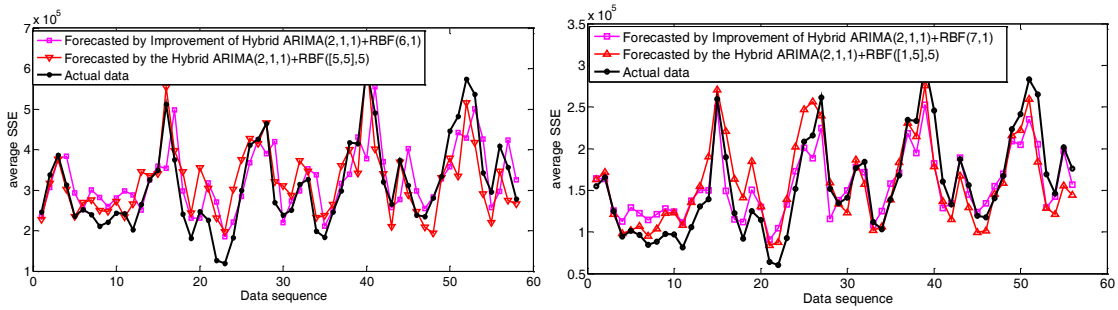

Figure 12. Comparison on forecast results between hybrid ARIMA-RBFNN and improvement hybrid ARIMA-RBFNN for (left) Thai tourist data and (right) foreigner tourist data.

4.5 The Hybrid of RBFNN-ARIMA Model's Design and Forecast

In this hybrid model, the designed RBFNN (7, 10) for Thai tourist data and RBFNN (7, 9) for foreigner tourist data in section 4.2 were adopted to forecast solution N_t in the first step. The residuals from RBFNN model will be fed into ARIMA model in the second step. For Thai tourist data, it found that between time lag 10 and time lag 13 of PACF and between time lag 10 and time lag 17 of ACF correlogram has the value more than critical value which basic indicated the order of AR is 3 and MA is 7, therefore the tentative model is ARIMA (3, 0, 7). For foreigner tourist, between time lag 2 and time lag 8 of PACF and between time lag 2 and time lag 12 of ACF has the value more than critical value, which basic indicate the order of AR is 7 and MA is 11, therefore the tentative model is ARIMA (7, 0, 11). However, after diagnostic checking and model selection by the criteria mentioned in section 4.1 are summarized in Table 3 for the 12 candidate models. It found that the suitable statistic model is ARIMA (4, 0, 8) for Thai tourist data and ARIMA (9, 0, 9) for foreigner tourist data.

Table 3. Comparison results of selection model for ARIMA models of the hybrid RBFNN-ARIMA.

ARIMA model	Thai tourist			ARIMA model	Foreigner tourist		
	Adjusted R^2	AIC	SBC		Adjusted R^2	AIC	SBC
(3, 0, 5)	0.115	24.80	25.07	(7, 0, 8)	0.357	23.31	23.84
(3, 0, 6)	0.147	24.78	25.08	(7, 0, 9)	0.338	23.35	23.91
(3, 0, 7)	0.411	24.42	24.76	(7, 0, 10)	0.352	23.33	23.94
(3, 0, 8)	0.342	24.54	24.92	(7, 0, 11)	0.580	22.91	23.55
(3, 0, 9)	0.260	24.67	25.08	(8, 0, 8)	0.305	23.41	23.98
(4, 0, 5)	0.200	24.71	25.02	(8, 0, 9)	0.298	23.43	24.04
(4, 0 ,6)	0.184	24.74	25.09	(8, 0, 10)	0.479	23.14	23.79
(4, 0, 7)	0.227	24.70	25.08	(8, 0, 11)	0.464	23.18	23.86
(4, 0, 8)	**0.445**	**24.38**	**24.80**	(9, 0, 8)	0.279	23.47	24.09
(4, 0, 9)	0.434	24.41	24.86	**(9, 0, 9)**	**0.711**	**22.57**	**23.22**
(5, 0, 5)	0.280	24.68	24.98	(9, 0, 10)	0.615	22.87	23.55
(5, 0, 6)	0.163	24.80	25.18	(9, 0, 11)	0.475	23.19	23.91

At a 99% confidence interval level for statistics test, the selected parameter has the statistical significance. The ARIMA model with these parameters of residuals for Thai and foreigner tourist data can be expressed respectively as

$$e_t = 2435.429 + 0.9488(1-B)YT_{t-1} - 0.4277(1-B)YT_{t-2} - 0.9999B\varepsilon_{Tt}, \quad (20)$$

$$e_t = 1706.86 + 0.8541(1-B)YF_{t-1} - 0.3386(1-B)YF_{t-2} - 0.9998B\varepsilon_{Ft}. \quad (21)$$

The forecast results by the hybrid RBFNN(7, 10)-ARIMA (4, 0, 8) model and hybrid RBFNN (7, 9)-ARIMA (9, 0, 9) model obtained are shown in Fig. 13 (left) and (right) for Thai tourist data and foreigner tourist data, respectively. It can be seen that the hybrid RBFNN-ARIMA model gives the more accuracy forecast results than RBFNN for both time series data.

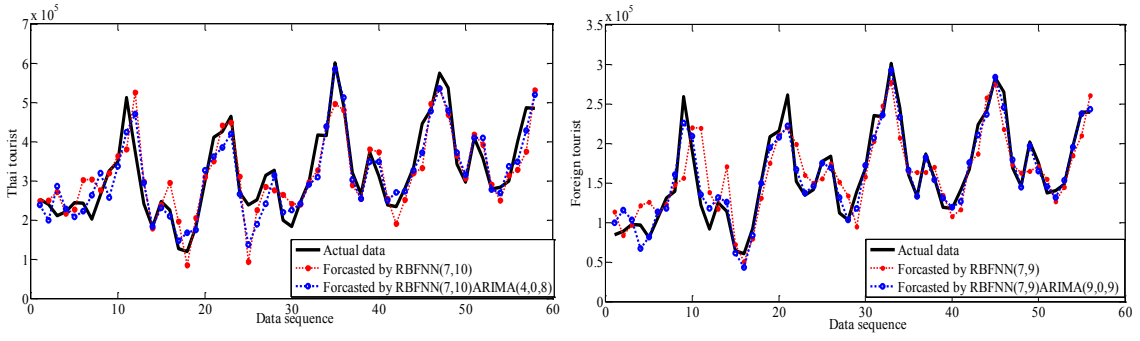

Figure 13. The forecast results of (left) the hybrid RBFNN(7, 10)-ARIMA(4, 0, 8) model by comparing with the RBFNN(7, 10) model for Thai tourist time series and (right) of the hybrid RBFNN(7, 9)-ARIMA(4, 0, 8) model by comparing with the RBFNN(7, 9) model for foreigner tourist time series.

4.6 Comparisons Forecast Results

After complete the design experiment, the forecast models are used to test with the test data of tourist time series in 2013-2014. The comparison results have shown in Table 4.

Table 4. Comparison of the forecast results for the optimized forecast model.

Model	Architecture of the forecast model		Average SSE of the forecast results	
	Thai tourist	Foreigner tourist	Thai tourist ($\times 10^{-9}$)	Foreigner tourist ($\times 10^{-9}$)
ARIMA	(2,1,1)	(2,1,1)	4.66	1.34
RBFNN	(7,10)	(7,9)	3.39	0.96
ARIMA-RBFNN	(2,1,1)-([5,5],5)	(2,1,1)-([1,5],5)	2.63	0.85
Improvement ARIMA-RBFNN	(2,1,1)-(6,1)	(2,1,1)-(7,1)	2.32	0.68
RBFNN-ARIMA	(7,10)-(4,0,8)	(7,9)-(9,0,9)	2.08	0.43

The valid model with a smaller average SSE is the proposed hybrid RBFNN-ARIMA model following by an I-ARIMA-RBFNN model, the hybrid ARIMA-RBFNN model, the RBFNN model, and an ARIMA model respectively. From the experiment results in Table 4, it has seen that a nonlinear ANN model can gives better result than a linear ARIMA model which exhibit more nonlinear pattern on this tourist time series data. Further, a hybrid linear and nonlinear forecast model i.e. hybrid ARIMA-RBFNN, an improvement ARIMA-RBFNN and hybrid RBFNN-ARIMA can clearly give more accuracy than a single linear or nonlinear model. However the priority processing between linear and nonlinear is the significant issue which will be considered. Since main pattern of this tourist problem is nonlinear then the first priority processing should be done by RBFNN and managed the remaining by ARIMA model. In this case, a hybrid RBFNN-ARIMA model is the best forecast model. In general case, there is no any theoretical guarantee which hybrid model is better. The best of combination of hybrid method depends on the nature of the problem.

5. Conclusions

The numbers of tourist was set as a dependent variable to the forecast model. An ARIMA, the RBFNN, the hybrid of ARIMA-RBFNN, an improvement of hybrid ARIMA-RBFNN, and the hybrid RBFNN-ARIMA model were designed and optimized to make up the forecast model and a two-year forecast was made. The experiment results showed that the optimized structure for the model i.e. an ARIMA (2, 1, 1), the RBFNN(7, 10), the hybrid of ARIMA (2, 1, 1)-RBFNN (6, 1), an IARIMA (2, 1, 1)-RBF ([5, 5], 5) and the hybrid of RBFNN (7, 10)-ARIMA (4, 0, 8) for Thai tourist model, and an ARIMA (2, 1, 1), the RBFNN (7, 9), the hybrid of ARIMA (2, 1, 1)-RBFNN (7, 1),an IARIMA (2, 1, 1)-RBF ([5, 1], 5), and the hybrid of RBFNN (7, 9)-ARIMA (9, 0, 9) for foreigner tourist model. The average SSE was used to indicate the performance of these models which indicated that the hybrid RBFNN-ARIMA model perform better than an IARIMA-RBFNN, the hybrid of ARIMA-RBFNN, RBFNN and ARIMA model by average 61.64%, 46.92%, 35.61% , and 23.55%. By the comparison, the proposed hybrid RBFNN-ARIMA model is the optimal model, an I-ARIMA-RBFNN and the hybrid ARIMA-RBFNN model are the sub-optimal model and RBFNN and ARIMA are the worst models. The hybrid RBFNN-ARIMA model, has the capability to learn the non-linear pattern and results the highly linear error output in this case study. The supplement forecast by ARIMA then can keep more the accuracy. Unlike the hybrid of ARIMA-RBFNN model, an ARIMA cannot trace the non-linear pattern well and then produce the uncertain form to the second stage forecast by RBFNN which made the errorneous prediction. However the hybridization of both linear and non-linear model can effective forecast than use the single one as ARIMA model and RBFNN model.

6. References

(1) Rochell T., "Travel and Tourism Economic Impact 2014 South East Asia," *Report*, 2014.

(2) Wang X., "A Hybrid Neural Network and ARIMA Model for Energy Consumption Forecasting," *J. of Comp*., vol. 7, no. 5, May 2012.

(3) Claveria O. and Torra S., "Forecasting Tourism Demand to Catalonia: Neural Networks vs. Time Series Models," Economic Modelling, 36(2014), pp. 220-228.

(4) Zhang G., Patuwo B.E., and Hu M. Y., "Forecasting with Artificial Neural Networks: The State of the Art," Int. J. of Forecasting 14(1998), 35-62.

(5) Marquee L., Connor M.O. and Remus W., "Artificial Neural Network Models for Forecasting and Decision Making," *Int. J. of Forecasting*, vol. 10, issue 1, 1994.

(6) Law R. and Au N., "A Neural Network Model to Forecast Japanese Demand for Travel to Hong Kond," Tourism Management, 20(1999), pp. 89-97.

(7) Zhang H.Q. and Li J.B., "Prediction of Tourist Quantity Based on RBF Neural Network," J. of Computers, vol. 7, no.4, 2012.

(8) Sing J.K., Basu D.K., Nasipuri M. and Kundu M., "Improved K-means Algorithm in the Design of RBF Neural Networks," *IEEE*, vol.2, 2003.

(9) Chen S., Cowen C.F.N and Grant P.M, "Orthogonal Least Square Learning Algorithm for Radial Basis Function Networks," *IEEE Trans. on NN*, vol.2, no. 2, 1991.

(10) Seymour G., Predictive Inference, New York, NY:Chapman and Hall, ISBN 0-412-03471-9, 1993.

(11) Ehem A., Introduction to machine learning, Cambridge Mass.[u.a.]: MIT Press, 2004.

(12) James O. B., Statistical decision theory and Bayesian analysis, Berlin:Springer-Verlag, 1985.

(13) Chatfield C., "Model uncertainty and forecast accuracy," *J. Forecasting*, vol. 15, 1996, pp. 495-508.

(14) Khashei M. and Bijari M., "An Artificial Neural Network (p,d,q) Model for Timeseries Forcasting," *J. of Expert Systems with Application*, vol. 37, 2010, pp.479-489

(15) http://www.tourism.go.th/home/listcontent/11/221/276 (last accessed 12/9/2014).