International Journal of Mathematics and
Computer Science, **16**(2021), no. 4, 1341–1354

A New Hybrid Forecasting Using Decomposition Method with SARIMAX Model and Artificial Neural Network

Chalermrat Nontapa[1], **Chainarong Kesamoon**[1], **Nicha Kaewhawong**[1], **Peerasak Intrapaiboon**[2]

[1]Department of Mathematics and Statistics
Faculty of Science and Technology
Thammasat University
Pathum Thani, Thailand

[2]Corporate Innovation Office
Siam Cement Group
Bangkok, Thailand

email: chalermrat.non@dome.tu.ac.th, chainarong@mathstat.sci.tu.ac.th, nicha@mathstat.sci.tu.ac.th, peerasai@scg.com

(Received February 11, 2021, Accepted March 16, 2021)

Abstract

In this paper, we present a new hybrid forecasting model using a decomposition method with SARIMAX model and Artificial Neural Network (ANN). The proposed model has combined linear and non-linear models such as a decomposition method with SARIMAX model and ANN. The new hybrid model is compared to SARIMA, SARIMAX, decomposition methods with SARIMA/SARIMAX models and ANN. We applied the new hybrid forecasting model to real monthly data sets such that the electricity consumption in the provincial area of Thailand and the SET index. The result shows that the new hybrid forecasting using a decomposition method with SARIMAX model and ANN performs well. The best hybrid model has reduced average error rate for 3 months and 12 months lead time forecasting of 47.3659% and

Key words and phrases: Time series, decomposition method, SARIMAX, artificial neural network, hybrid model.
AMS (MOS) Subject Classifications: 68T07.
ISSN 1814-0432, 2021, http://ijmcs.future-in-tech.net

33.1853%, respectively. In addition, the new hybrid forecasting model between decomposition method with SARIMAX models and ANN has the lowest average MAPE of 1.9003% for 3 months and 2.2113% for 12 months lead time forecasting, respectively. The best forecasting model has been checked by using residual analysis. We conclude that the combined model is an effective way to improve more accurate forecasting than a single forecasting method.

1 Introduction

Improving the forecasting accuracy has become vital for decision makers and managers in various fields of science especially time series forecasting. Many researchers believe that combining different models or using hybrid models can be an effective solution to improve forecasting accuracy and to overcome limitations of single models. Theoretical, as well as empirical, evidence in the literature suggest that by combining in homogeneous models, the hybrid models will have lower generalization variance or error. On the other hand, the main aim of combined models is to reduce the risk of failure and obtain results that are more accurate [1].

Many researchers have devoted their time to develop and improve the hybrid time series models since the early work of Reid [2] and Bates and Granger [3]. In pioneering work on combined forecasts, Bates and Granger showed that a linear combination of forecasts would give a smaller error variance than any of the individual methods. Since then, the studies on this topic have expanded dramatically. Combining linear and nonlinear models is one of the most popular and widely used hybrid models, which have been proposed and applied in order to overcome the limitations of each component and improve forecasting accuracy. Chen and Wang [4] constructed a combination model incorporating SARIMA model and SVMs for seasonal time series forecasting. Khashei et al. [5] presented a hybrid autoregressive integrated moving average and feedforward neural network to time series forecasting in incomplete data situations, using fuzzy logic. Pai and Lin [6] proposed a hybrid method to exploit the unique strength of ARIMA models and SVMs to forecast stock prices.

In this paper, we present the performance of a new series hybrid model for time series using a decomposition method with SARIMAX model and ANN model. The main aim of this paper is to determine the relative predic-

tive capabilities of the decomposition method with SARIMAX-ANN models. On the other hand, this paper aims to conclude which sequence of decomposition method with SARIMAX and ANN is better for constructing series hybrid models for time series forecasting of the electricity consumption in the provincial area of Thailand and the SET index. The rest of this paper is organized as follows: In the following section, we review the decomposition method, SARIMA, SARIMAX, ANN and the proposed method. In section 3, we explain data preparation and model evaluation criteria which we experimented in this research. Empirical results for comparing the forecasting techniques from two monthly real data sets are illustrated in section 4. The final section provides a conclusion and directions for future research.

2 Methodology

2.1 Decomposition Method

An important goal in time series analysis is the decomposition of a series into a set of non-observable components that can be associated to different types of temporal variations such as trend, seasonal, cycle and irregular. The most common forms are known as additive and multiplicative decomposition, which are expressed in equations (2.1) and (2.2), respectively.

$$y_t = (T_t + S_t + C_t + I_t) + \varepsilon_t \tag{2.1}$$

$$y_t = (T_t \times S_t \times C_t \times I_t) + \varepsilon_t \tag{2.2}$$

2.2 Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

Seasonal Autoregressive Integrated Moving Average (SARIMA) Model is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality. The SARIMA$(p,d,q)(P,D,Q)_S$ model is expressed as follows:

$$\phi_p(B)\,\Phi_P\left(B^S\right)(1-B)^d\left(1-B^S\right)^D y_t = \theta_q(B)\,\Theta_Q\left(B^S\right)\varepsilon_t \tag{2.3}$$

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)\left(1-\sum_{k=1}^{P}\Phi_k B^{kS}\right)z_t=\left(1-\sum_{j=1}^{q}\theta_j B^j\right)\left(1-\sum_{l=1}^{Q}\Theta_l B^{lS}\right)\varepsilon_t \tag{2.4}$$

2.3 Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) Model

Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) Models is a SARIMA model with Exogenous Variables (X), called SARIMAX$(p,d,q)(P,D,Q)_S$, where X is the vector of exogenous variables. The exogenous variables can be modeled by a multiple linear regression equation which is expressed as follows:

$$y_t=\beta_0+\beta_1X_{1,t}+\beta_2X_{2,t}+\cdots+\beta_kX_{k,t}+\omega_t, \tag{2.5}$$

where β_0 is a constant parameter and $\beta_1,\beta_2,\ldots,\beta_k$ are regression coefficient parameters of exogenous variables, $X_{1,t},X_{2,t},\ldots,X_{k,t}$ are observations of k exogenous variables corresponding to the dependent variable y_t; ω_t is a stochastic residual; i.e., the residual series that is independent of input series.

$$\omega_t=\frac{\theta_q(B)\,\Theta_Q\left(B^S\right)}{\phi_p(B)\,\Phi_P\left(B^S\right)(1-B)^d\left(1-B^S\right)^D}\varepsilon_t \tag{2.6}$$

The general SARIMAX model equation can be obtained by substituting Equation (2.6) into Equation (2.5) [7].

$$y_t=\beta_0+\sum_{i=1}^{k}\beta_iX_{i,t}+\frac{\theta_q(B)\,\Theta_Q\left(B^S\right)}{\phi_p(B)\,\Phi_P\left(B^S\right)(1-B)^d\left(1-B^S\right)^D}\varepsilon_t \tag{2.7}$$

2.4 Artificial Neural Network

The Artificial Neural Network is a computing algorithm that can solve complex problems consisting of artificial neurons or nodes which are information processing units arranged in layers and interconnected by synaptic weights (connections). Neurons can filter and transmit information in a supervised fashion in order to build a predictive model that classifies data stored in memory. The ANN model is a three-layered network of interconnected nodes: the

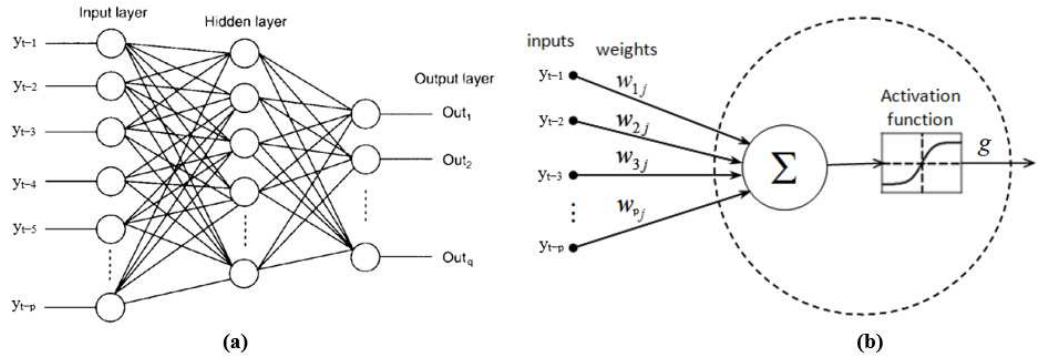

Figure 1: Architecture of Neural network and active node. (**a**) Neural network architecture; (**b**) Neural network active node.

input layer, the hidden layer, and the output layer. The nodes between input and output layers can form one or more hidden layers. Every neuron in one layer has a link to every other neuron in the next layer, but neurons belonging to the same layer have no connections among them (see Figure 1(a)).

Input layer phase: The input layer receives information from the outside world, the hidden layer performs the information processing and the output layer produces the class label or predicts continuous values. For the time series problems, an ANN is fitted with past lagged value of actual data $(y_{t-1}, y_{t-2}, y_{t-3}, ..., y_{t-p})$ as an input vector. Therefore, an input layer is composed of p nodes that are connected to the hidden layer.

Hidden layer phase: The hidden layer is an interface between input and output layers. The ANN models which are designed in this paper have two hidden layers with q nodes. In this step, one of the important tasks is determining the type of activation function $g(y)$ which is identifying the relationship between input and output layer (see Figure 1(b)). Neural networks support a wide range of activation function such as linear, quadratic, tanh and logistic. In this research, the tanh function is used as the hidden layer activation function that is shown in the following equation:

$$g(y) = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}} \tag{2.8}$$

Output layer phase: In this step, by selecting an activation transfer function and the appropriate number of nodes, the output of neural network is used to predict the future values of time series. In this paper, the output

layer designed by neural network contains one node because the onestep-ahead forecasting is considered. Also, the linear function as a non-linear activation function is introduced for the output layer. The formula that expresses the relationship between the input and output layers follows.

$$y_t = w_0 + \sum_{j=1}^{q} w_j \cdot g\left[w_{0,j} + \sum_{i=1}^{p} w_{i,j} \cdot y_{t-i}\right] + \varepsilon_t, \tag{2.9}$$

where $w_{i,j}(i = 0, 1, ..., p \quad and \quad j = 1, 2, ..., q)$ and $w_j(j = 0, 1, ..., q)$ are referred as connection weights. It should be noted that deciding the number of neurons in the hidden layer (q) and the number of lagged observations (p) and the dimension of the input vector in input layer are vital parts of neural network architectures, but no methodical rule exists in order to select these parameters and the only possible way to choose an optimal number of p and q is trial and error [5].

2.5 Proposed Method

The proposed method is applied in order to combine the linear and non-linear models. This method includes three steps as follows (see Figure 2). We note first Zhang's hybrid model [8]:

$$y_t = L_t + N_t + \varepsilon_t \tag{2.10}$$

In the first step, a linear component is estimated by the decomposition with SARIMA/SARIMAX model. The decomposition with SARIMA/SARIMAX starts by decomposing data in time series into four parts: irregular (I), trend-cycle (TC), trend-cycle-irregular (TCI) and seasonality (S) components by using multiplicative decomposition. SARIMA and SARIMAX are applied to the trend-cycle-irregular part to find the model that best describes it. After that, each of SARIMA and SARIMAX trend-cycle-irregular are then combined with a seasonal index to make a series of forecast values. Let e_t denote the residual of the decomposition with SARIMA/SARIMAX model at time t :

$$e_t = y_t - \hat{L}_t \tag{2.11}$$

In the second step, an artificial neural network is used to model the residuals from the decomposition with SARIMA/SARIMAX. We claim that the residuals from the decomposition with SARIMA/SARIMAX can capture

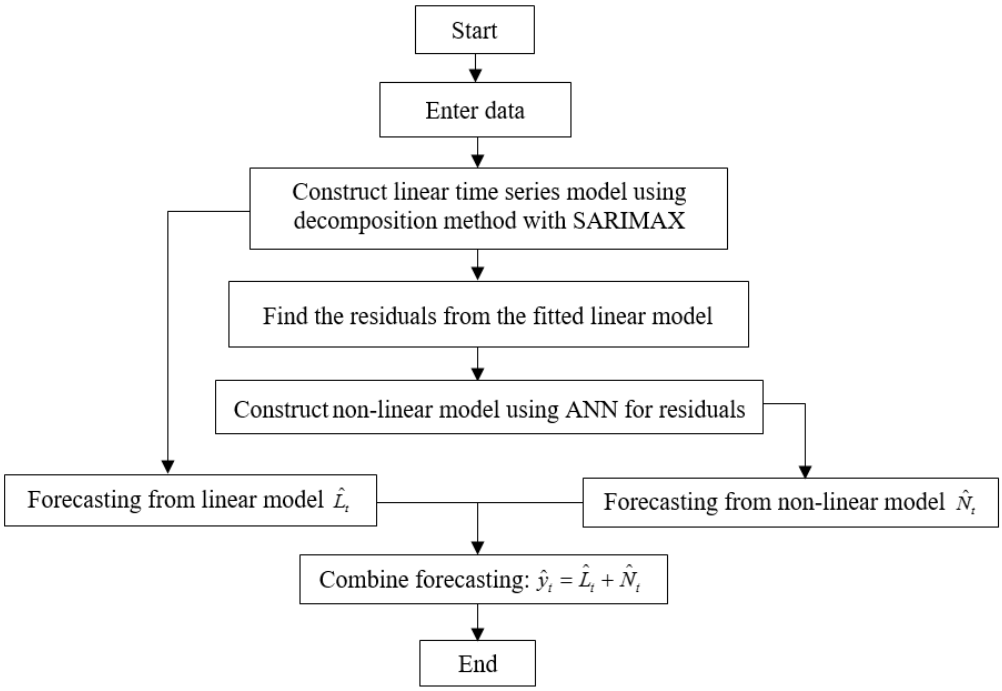

Figure 2: Hybrid model chart between the decomposition with SARIMAX model and ANN.

the artificial neural network to modeling nonlinear relationships. With p input nodes, the ANN model for the residuals is:

$$e_t = f\left(e_{t-1}, e_{t-2}, e_{t-3}, ..., e_{t-p}\right) + \varepsilon_t \tag{2.12}$$

$$e_t = w_0 + \sum_{j=1}^{q} w_j \cdot g\left[w_{0,j} + \sum_{i=1}^{p} w_{i,j} \cdot e_{t-i}\right] + \varepsilon_t \tag{2.13}$$

In the third step, the linear $(\hat{L}_t)$ and non-linear $(\hat{N}_t)$ forecasting values obtained from the first and second steps, denoted as linear non-linear components respectively, are combined as follows:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{2.14}$$

3 Data and Model Evaluation Criteria

3.1 Data Descriptions

The first data set is an electricity consumption in the provincial area of Thailand [9], where the monthly electricity consumption from January 2002 to December 2019 shows a seasonal pattern which has a change in seasonal fluctuations, giving a total of 216 observations. The second data set is the index stock SET of Thailand [10]. We consider the monthly time series data from January 1998 to December 2019.

Table 1: Details of the time series data sets.

Series	**Electric**	**SET**
Sample Size	216	264
Training Set	2002 2018 (204)	1998 2018 (252)
Test Set (3 months)	2019 (January March)	
Test Set (12 months)	2019 (January December)	

3.2 Data Preparation

To assess the forecasting performance of different models, each data set is divided into two samples of training and testing. The training data is used exclusively for model development and then the test sample for 3 months and 12 months are used to evaluate the established model. The data compositions for the two data sets are given in Table 1.

3.3 Model Evaluation Criteria

3.3.1 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a measure of prediction accuracy of a forecasting method in Statistics. It usually expresses the accuracy as a ratio. MAPE is defined by the formula:

$$MAPE\% = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100 \tag{3.15}$$

3.3.2 Reduce Error Rate (RER)

Reduce Error Rate (RER) is a measure of reduce errors for proposed model when compared with original model as a ratio. RER is defined by the formula:

$$RER\% = \left(1 - \frac{MAPE_{Proposed}}{MAPE_{Original}}\right) \times 100 \tag{3.16}$$

4 Empirical results and discussion

The results for two real monthly data sets using ANN, SARIMA, SARIMAX, the decomposition method with SARIMA/SARIMAX models and hybrid model between the decomposition method with SARIMA/SARIMAX and ANN models for 3 months and 12 months lead time forecasting were showed in Tables 2 and 3. We have conducted experiments on RStudio and SPSS package.

Table 2: MAPE and average MAPE of two data sets for 3 months lead time forecasting.

Models	Electric	SET	Average
ANN	3.6314	4.3499	3.9907
SARIMA	2.8308	4.3900	3.6104
SARIMAX	1.6829	4.0506	2.8668
DEC-SARIMA	1.2400	3.3090	2.2745
DEC-SARIMAX	1.5601	3.2035	2.3818
Hybrid (DEC-SARIMA-ANN)	1.4107	2.9756	2.1932
Hybrid (DEC-SARIMAX-ANN)	**1.1135***	**2.6871***	**1.9003****

Tables 2 and 3 show the average MAPE of seven time series forecasting methods for 3 months and 12 months lead time forecasting. We found that the hybrid model between the decomposition method with SARIMAX model and ANN has the lowest MAPE for two data sets and average MAPE of 1.9003% and 2.2113% for 3 months and 12 months lead time forecasting, respectively.

In Figure 3, the network diagram that SPSS used to predict residuals from the decomposition method with SARIMAX model of the electricity consumption in the provincial area of Thailand is shown in Figure 3(a). The diagram shows 14 input nodes ($e_{t-1}, e_{t-2}, e_{t-3}, ..., e_{t-12}, S_t, S_{t-1}$), 8 nodes in the first hidden layer, 6 nodes in the second hidden layer and 1 output node.

Table 3: MAPE and average MAPE of two data sets for 12 months lead time forecasting.

Models	Electric	SET	Average
ANN	4.2456	3.1430	3.6943
SARIMA	2.8440	3.7751	3.3096
SARIMAX	2.5369	3.6173	3.0771
DEC-SARIMA	2.1757	3.5516	2.8637
DEC-SARIMAX	2.1200	3.4791	2.7996
Hybrid (DEC-SARIMA-ANN)	2.0571	2.9983	2.5367
Hybrid (DEC-SARIMAX-ANN)	**1.8795***	**2.5430***	**2.2113****

In Figure 3(b), the network diagram used to predict residuals from decomposition method with SARIMAX model of the SET index. The diagram shows 14 input nodes ($e_{t-1}, e_{t-2}, e_{t-3}, ..., e_{t-14}$), 8 nodes in the first hidden layer, 6 nodes in the second hidden layer and 1 output node.

Moreover, the hybrid model between the decomposition method with SARIMAX model and ANN fitting was adequate for two real monthly data sets with the Portmanteau Statistic Q of Box-Ljung. This model has been checked by using residual analysis. We conclude that the random errors are normally distributed, no autocorrelated, zero mean and constant variance.

Figure 4 shows that the sequence chart between the actual value (blue line) and forecast value (red line) for two real monthly data sets (a) Electric and (b) SET index using hybrid model between the decomposition method with SARIMAX model and ANN.

Table 4: RER% for 3 months and 12 months lead time forecasting of two data sets using hybrid model between the decomposition method with SARIMAX model and ANN.

RER%	Electric	SET	Average
3 months	60.6648	38.7904	47.3659
12 months	33.9135	32.6375	33.1853

Table 4 shows the average Reduce Error Rates (RER) of hybrid model between the decomposition method with SARIMAX model and ANN for two real monthly data sets in 3 months and 12 months lead time forecasting. We found that the proposed model has average RER of 47.3659% and 33.1853%, respectively. Moreover, the proposed model for 3 months lead time forecast-

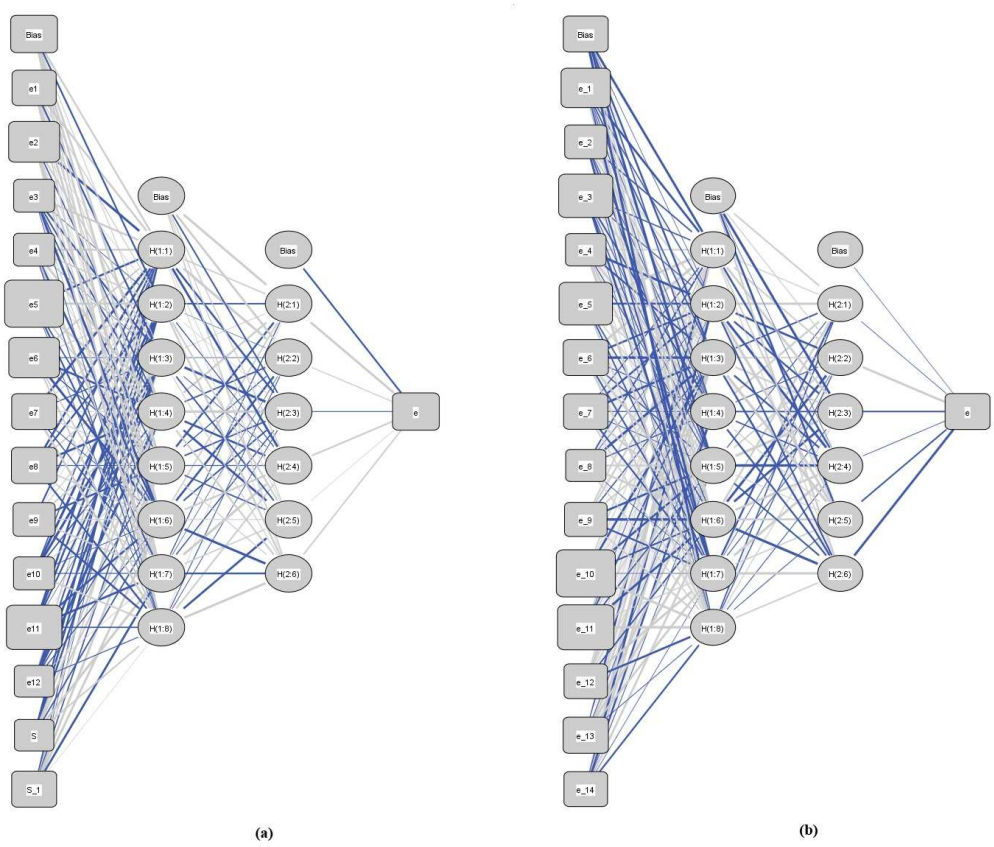

Figure 3: Network diagram, which construct from the residuals of decomposition method with SARIMAX model for two data sets. (a) ANN (14×8×6×1) of Electric; (b) ANN($14 \times 8 \times 6 \times 1$) of SET.

ing is more accurate than for 12 months lead time forecasting.

Hybrid model between the decomposition method with SARIMAX model and ANN for two monthly data sets are expressed as follows:

1. Electric: Hybrid model (R% = 98.6935)

DEC-SARIMAX$((13), 0, (2, 15, 44, 47, 51))(0, 0, 0)_{12} - ANN(14 \times 8 \times 6 \times 1)$

SARIMAX$((13), 0, (2, 15, 44, 47, 51))(0, 0, 0)_{12}$ with TCI_{t-1} and TCI_{t-12}

$L_t = [\beta_0 + \beta_1(TCI_{t-1}) + \beta_2(TCI_{t-12}) + \phi_{13}\omega_{t-13} - \theta_2\varepsilon_{t-2} - \theta_{15}\varepsilon_{t-15}$

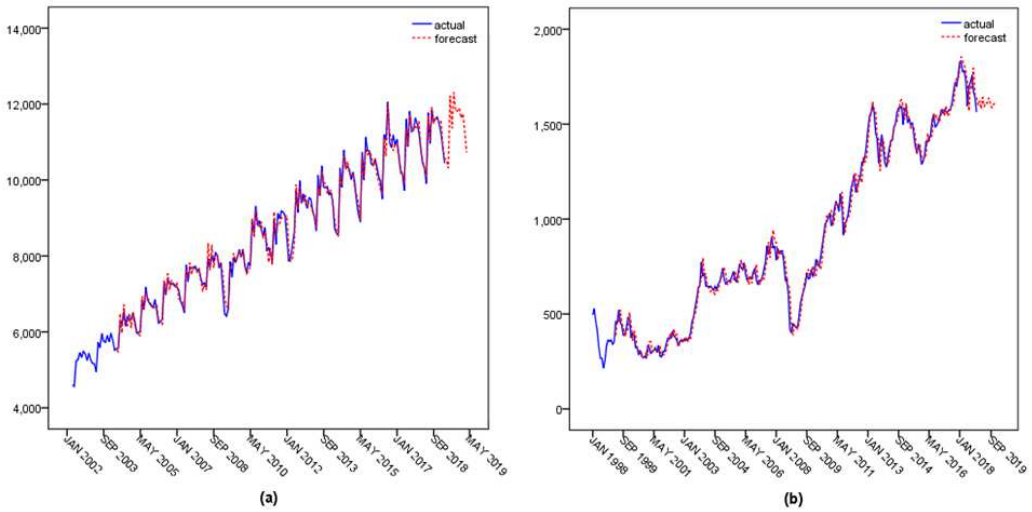

Figure 4: The time plots of actual value (blue line) and forecast value (red line) for two real monthly data sets: (a) Electric and (b) SET index using hybrid model between the decomposition method with SARIMAX model and ANN.

$$- \theta_{44}\varepsilon_{t-44} - \theta_{47}\varepsilon_{t-47} - \theta_{51}\varepsilon_{t-51} + \varepsilon_t] \times S_t + \varepsilon'_t$$

$$\hat{L}_t = [119.9010 + 0.8468(TCI_{t-1}) + 0.1486(TCI_{t-12}) \ -0.2896\omega_{t-13} - 0.2338\varepsilon_{t-2} - 0.1924\varepsilon_{t-15} - 0.1738\varepsilon_{t-44} - 0.2690\varepsilon_{t-47} + 0.1908\varepsilon_{t-51}] \times \hat{S}_t$$

$\hat{N}_t = ANN(14{\times}8{\times}6{\times}1)$ with 14 input nodes $(e_{t-1}, e_{t-2}, e_{t-3}, ..., e_{t-12}, S_t, S_{t-1})$, 8 nodes in the first hidden layer, 6 nodes in the second hidden layer and 1 output node.

2. SET: Hybrid model (R% = 98.9211)

DEC-SARIMAX$((0,0,(33))(0,0,0)_{12} - ANN(14 \times 8 \times 6 \times 1)$

SARIMAX$((0,0,(33))(0,0,0)_{12}$ with TCI_{t-1}

$$L_t = [\beta_1(TCI_{t-1}) - \theta_{33}\varepsilon_{t-33} + \varepsilon_t] \times S_t + \varepsilon'_t$$

$$\hat{L}_t = [1.0037(TCI_{t-1}) - 0.1689\varepsilon_{t-33}] \times \hat{S}_t$$

$\hat{N}_t = ANN(14 \times 8 \times 6 \times 1)$ with 14 input nodes $(e_{t-1}, e_{t-2}, e_{t-3}, ..., e_{t-14})$, 8 nodes in the first hidden layer, 6 nodes in the second hidden layer and 1 output node.

5 Conclusion and Future Research

5.1 Conclusion

Time series forecasting is one of the very demanding subjects during the last few decades since it has many applications in financial, economics, engineering and scientific modeling. The standard statistical techniques in the literature, namely the decomposition and the Box-Jenkins methods are well known for many researchers. In this research, we preferred a new hybrid method; namely, a decomposition method with SARIMA-ANN and decomposition method with SARIMAX-ANN models with application to two real monthly data sets such that the electricity consumption in the provincial area of Thailand and the SET index. The proposed methods were compared to ANN, SARIMA, SARIMAX, the decomposition with SARIMA and the decomposition with SARIMAX. The performance evaluation results indicated that the decomposition method with SARIMAX and ANN performed well. In addition, the decomposition method with SARIMAX and ANN model have average reduced error rate of 47.3659% and 33.1853% for 3 months and 12 months lead time forecasting, respectively compared to the SARIMA model.

5.2 Future Research

We can try to use the trend-cycle-irregular (TCI) in Exponential Smoothing method or the Holt-Winters method by applying this technique to some other problems and big data sets with various numbers of features.

Acknowledgment. This research was supported by the Government of Canada, Canada-ASEAN Scholarships and Educational Exchanges for Development (SEED 2019-2020).

References

[1] M. Hibon, T. Evgeniou, To combine or not to combine: selecting among forecasts and their combinations, International Journal of Forecasting, **21,** (2005), 15–24.

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

[3] J. M.Bates, C. W. J. Granger, Combination of forecasts, Operational Research, **4,** (1969), 451–468.

[4] K. Y. Chen, C. H. Wang, A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, Expert Systems with Applications, **32,** (2007), 254–264.

[5] M. Khashei, M. Bijari, G. A. Raissi Ardali, Improvement of Auto-Regressive Integrated Moving Average models using Fuzzy logic and Artificial Neural Networks, Neurocomputing, **72,** (2009), 956–967.

[6] P. F. Pai, C. S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, **33,** (2009), 497–505.

[7] M. Cools, E. Moons, G. Wets, Investigating variability in daily traffic counts using ARIMAX and SARIMA (X) models: assessing impact of holidays on two divergent site locations, Journal of the Transportation Research Board, **72,** (2009), 1–22.

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

[9] Energy Policy and Planning office (EPPO) Ministry of Energy, electricity consumption in the provincial area of Thailand, http://www.eppo.go.th/index.php/en/en-energystatistics/electricity-statistic.

[10] Stock Exchange of Thailand, the SET index, https://www.set.or.th/en/market/market_statistics.html