Modern Applied Science; Vol. 7, No. 7; 2013
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

A Forecasting Model for Thailand's Unemployment Rate

Kanlapat Mahipan[1], Nipaporn Chutiman[1] & Bungon Kumphon[1]

[1] Mathematics Department, Mahasarakham University, Thailand

Correspondence: Bungon Kumphon, Mathematics Department, Mahasarakham University, Thailand. E-mail: bungon.k@msu.ac.th

Received: April 15, 2013 Accepted: June 1, 2013 Online Published: June 9, 2013

doi:10.5539/mas.v7n7p10 URL: http://dx.doi.org/10.5539/mas.v7n7p10

Abstract

This study deals with two approaches—viz. via Box-Jenkins and artificial neuron network to forecast the unemployment rate in Thailand. The Box-Jenkins approach proves more efficient to estimate the unemployment rate in Thailand, with less MAPE compared to the second model. The forecast values are consistent with the actual values and tend to decrease.

Keywords: Box-Jenkins, artificial neuron network, unemployment rate

1. Introduction

1.1 Introduce the Problem

The unemployment rate (UR) is an important key to indicate economic status, and UR forecasting is a basic tool for planning and risk management in tax, finance, education, agricultural and industrial policies. Two approaches —viz. Box-Jenkins technique that combines moving average (MA) and autoregressive (AR) models, and data mining via an artificial neural network (ANN) model are very popular in prediction. Both approaches are flexible for complicated non-linear data, and their advantages include computational speed, low cost feasibility, and ease of design for operators with little technical experience. Box-Jenkins involves a very strict assumption for residuals during the diagnostic checking stage before proceeding to forecasting. The ANN approach has offers a very good approximation capability, and additional advantages such as fast processing times where the mathematical formulae and prior knowledge on the relationship between inputs and outputs are unknown (Kankal, Akpinar, Kömürcü, & Özşahin, 2011; Sözen, Arcaklioglu, & Ozkaymak, 2005; Sözen & Arcaklioglu, 2007).

The National Statistical Office (NSO) collects the national UR data, but have to take more time to present update reports. This is the motivation for our study. The objective of this study is to evaluate the model to forecast the UR in Thailand based on economic variable defined by the NSO, by using ANN compare to Box-Jenkins techniques. The results from this study employ the important informations in assessing UR patterns and selecting a more accurate approach to estimate the future UR. The remainder of this paper is organized as follows. Section 2 proposes the forecasting methodology of the ANN and Box-Jenkins approaches. Section 3 presents the modelling of Thailand's UR, and some conclusions are stated in the last Section.

2. Methodology and Data

The two different forecasting approaches, via ANNs and SARIMA from Box-Jenkins, are investigated to model the UR in Thailand. Six different models from these two approaches include twelve economic variables defined by the NSO—viz. the total number of workers (x1), the number of seasonal workers (x2), those compulsorally insured (x3), the number employed (x4), the use of electricity (x5), car sales (x6), the industrial production index (x7), the set index (x8), the private investment index (x9), Thailand's economic indicator (x10), the industrial labor productivity index (x11) and the industrial worker index (x12). The response variable is UR(y).

The monthly data used have been collected by the Labour Force in Thailand project of the NSO, from January 2003 to December 2011—cf. Figure 1. The UR is obvious decreasing during this period.

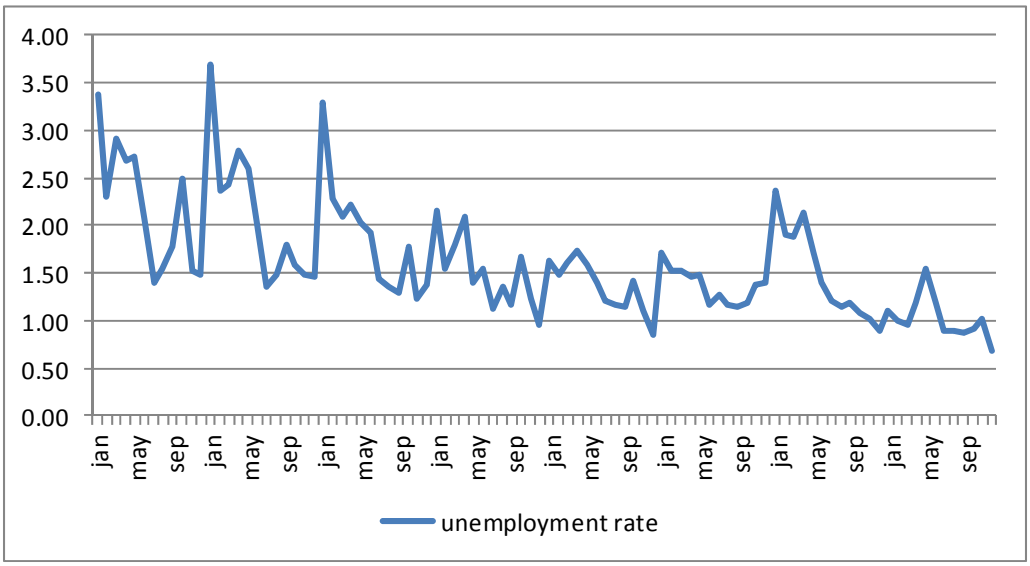


Figure 1. The unemployment rate in Thailand from January 2003 to December 2011

2.1 Artificial Neural Network Model Approach

The processes—viz. training and testing are the methodology of an ANN. The training of ANNs usually involves modifying the connection weights by mean of the learning rule. The total error, based on the squared difference between the predicted and actual output, is computed for the whole training set. Adjustment of the correction weights is carried out using the standard error back-propagation algorithm, which minimizes the total error using the gradient decent method. More details on the back-propagation algorithm shown in Figure 2 are given in Kankal, Akpinar, Kömürcü and Özşahin (2011). Then, testing data are used to check the generalization.

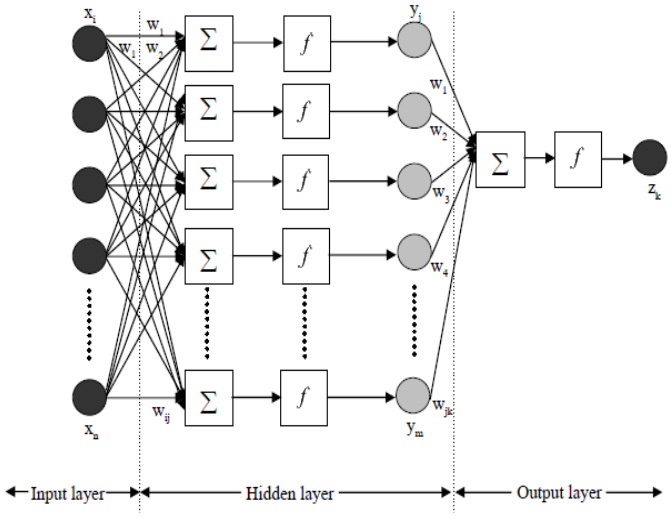


Figure 2. The architecture of the back-propagation network

2.2 Box-Jenkins Approach

The purpose of Box-Jenkins is to find an appropriate model based on statistical concepts. There are both statistical tests to find validity of the model and statistical measures of forecast uncertainty. The iterative approach, with three steps of the model-building, presents in Figure 3.

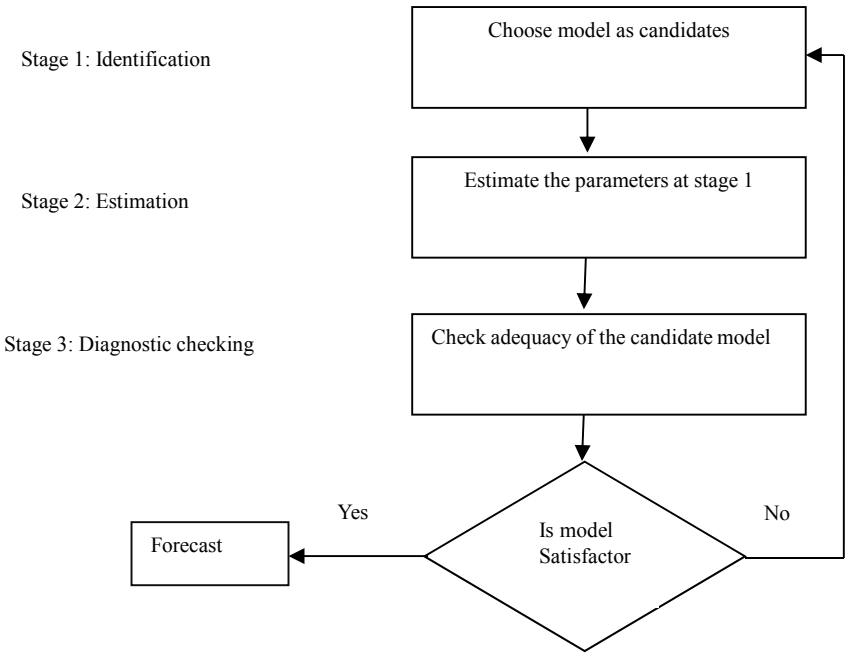


Figure 3. Stages in the Box-Jenkins approach for model building

In mathematical notation, the purely Box-Jenkins ARMA model is a combination of the AR (Autoregressive) and MA (Moving Average) models as

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} \quad (1)$$

where y_t is the observation at time *t*; the φ's and θ's are parameters of the model and a_t is the residual at time *t* with constant mean 0 and variance σ^2, and uncorrelated with each other, which call "white noise" (Dobre & Adeiana, 2008).

Stationary—i.e. with constant mean, constant variance, is necessary in Box-Jenkins model. Differencing of non-stationary series one or more times is required to the achieve stationary series, and "I" stands for integrated. Thus the model becomes ARIMA. The Box-Jenkins approach can be extended to include a seasonal term (S) in the model as the SARIMA. At stage 1, the order for the seasonal autoregressive and seasonal moving average terms can be included in the model—i.e. it is not necessary remove seasonality before fitting the model.

2.3 Accuracy of Model

The accuracy of the forecast is evaluated based on the estimation of error or residual. Thus the smaller the values of the root mean square error (RMSE) and the mean absolute percent error (MAPE), the better the forecast. The MAPE criterion is the decisive factor, because it is expressed in easy generic percentage terms. The following equations are the respective formulas used in computing the RMSE (Mustafa et al., 2012) and MAPE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Predict}_i - \mathrm{Raw}_i)^2} \quad (2)$$

and

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\mathrm{Predict}_i - \mathrm{Raw}_i}{\mathrm{Raw}_i}\right| \times 100 \quad (3)$$

where Raw_i and $\mathrm{Predict}_i$ are the actual and predicted observed at time i respectively, and n is the total number of the predictions. The criterion of MAPE for model evaluation is based on Lewis (1982).

3. Results

3.1 Construction, Teaching and Testing of Artificial Neural Network (ANN)

From the historical data, the appropriate ANN from training data to forecast the UR is discovered, including the

twelve variables as mentioned above. It is not straightforward to determine the best size of the networks for a system, so we apply the correlation between the UR and the other data to consider the size in the networks as 0.2-0.7, greater than 0.7, 0.6-0.7, and 0.2-0.59. Four multilayer networks; ANN1: 12-11-1 (with x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12), ANN2: 2-22-1 (with x1, x3), ANN3: 6-11-1 (with x1, x3, x5, x7, x9, x10) and ANN4: 6-11-1 (with x2, x4, x6, x8, x11, x12) are considered to train the network with the output layer as the UR. Data are split into training and testing as 70:30 percentages. The dashed line in Figure1shows the area between training and testing.

Data normalization is used in the data preprocessing. Transfer functions known as tangent sigmoids are used at the hidden layer. Each group of input and output values are normalized into the range [0.1, 0.9]; the range [0, 1] improves the learning speed, as

$$\text{Normalized value} = \left[\frac{\text{Raw value} - \text{Minimum value}}{\text{Maximum value} - \text{Minimum value}}\right] \times (0.9 - 0.1) + 0.1 \quad (4)$$

The determination of the number of nodes in the hidden layer is not "exact science"—cf. Kankal et al. (2011). The network is therefore tested for different numbers of hidden layer nodes, in order to find the optimum and good convergence for the ANN structure. The problem in the training of an ANN is memorization, which the training is cut when the network starts to memorize. To prevent this, the error values of the training set may be greater than the testing set in the models. The accuracy in training is monitored by RMSE of the training and testing patterns separately

In this study, initial weights for the learning rate are initialized into random values between -0.5 and 0.5, the learning rate equals 0.025, and the momentum is 0.8. After the learning set of data was presented to the ANN models, we stopped the learning process when the epochs reached 50,000 iterations. The best result from Table 1 for our ANN model forecast of the UR is ANN-1: 12-11-1 with the inaccuracy in forecasting where MAPE > 50%.

Table 1. ANN models and their training and testing error

ANN structure	Training error	Testing error	MAPE (%)
Model 1 (ANN-1) 12-11-1	0.0005	0.9428	65.3538
Model 2 (ANN-2) 2-11-1	0.2747	4.3167	214.5998
Model 3 (ANN-3) 6-11-1	0.0356	3.0505	174.8173
Model 4 (ANN-4) 6-11-1	0.0354	2.7852	197.8916

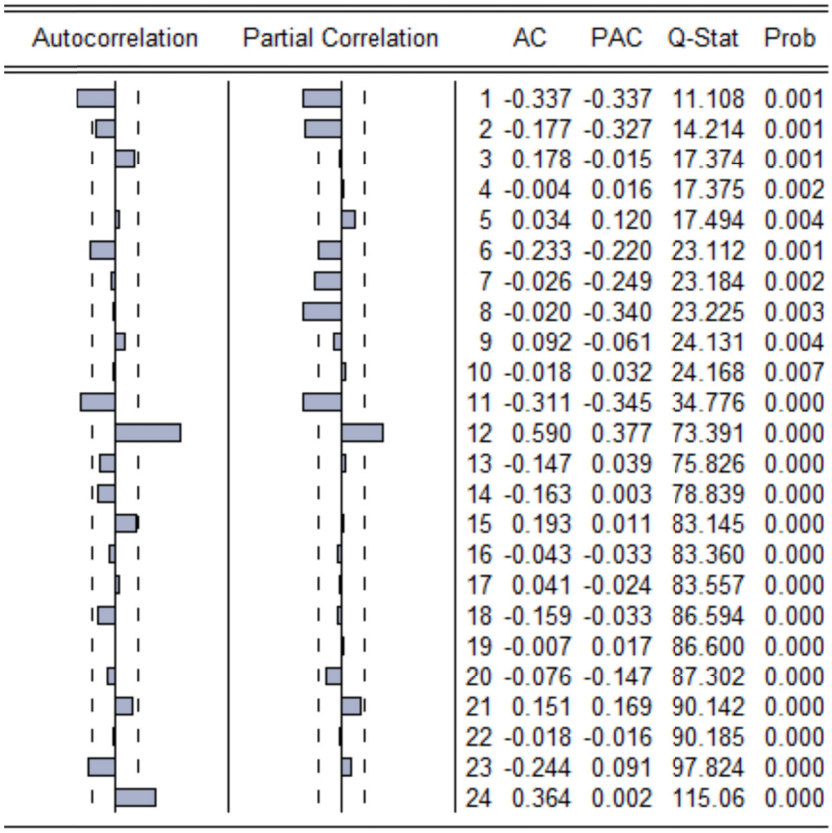
3.2 Model Building by Box-Jenkins

In the Box-Jenkins approach, data are split into two parts (70%) and (30%). The first step is to determine stationary and seasonality—and the Augmented Dickey-Fuller test (ADF) and autocorrelation (ACF) are used, respectively. The unit root test for stationary by the ADF shows that the series has a unit root and it is non-stationary for the original series—cf Table 2. With first order difference, it therefore becomes stationary (p-value < 0.00001).

Table 2. The unit root test for stationary by ADF test

	Statistics	None of intercept and trend	Intercept, none of trend	Intercept and trend
Original series	ADF	-2.6879*	-0.3985 ^{ns}	-1.1217 ^{ns}
	p-value	0.0077	0.9037	0.9188
First order difference	ADF	-7.6006*	-8.3419*	-8.2761*
	p-value	<0.0000	<0.0000	<0.0000

*significant at 0.05



	AC	PAC	Q-Stat	Prob
1	-0.337	-0.337	11.108	0.001
2	-0.177	-0.327	14.214	0.001
3	0.178	-0.015	17.374	0.001
4	-0.004	0.016	17.375	0.002
5	0.034	0.120	17.494	0.004
6	-0.233	-0.220	23.112	0.001
7	-0.026	-0.249	23.184	0.002
8	-0.020	-0.340	23.225	0.003
9	0.092	-0.061	24.131	0.004
10	-0.018	0.032	24.168	0.007
11	-0.311	-0.345	34.776	0.000
12	0.590	0.377	73.391	0.000
13	-0.147	0.039	75.826	0.000
14	-0.163	0.003	78.839	0.000
15	0.193	0.011	83.145	0.000
16	-0.043	-0.033	83.360	0.000
17	0.041	-0.024	83.557	0.000
18	-0.159	-0.033	86.594	0.000
19	-0.007	0.017	86.600	0.000
20	-0.076	-0.147	87.302	0.000
21	0.151	0.169	90.142	0.000
22	-0.018	-0.016	90.185	0.000
23	-0.244	0.091	97.824	0.000
24	0.364	0.002	115.06	0.000

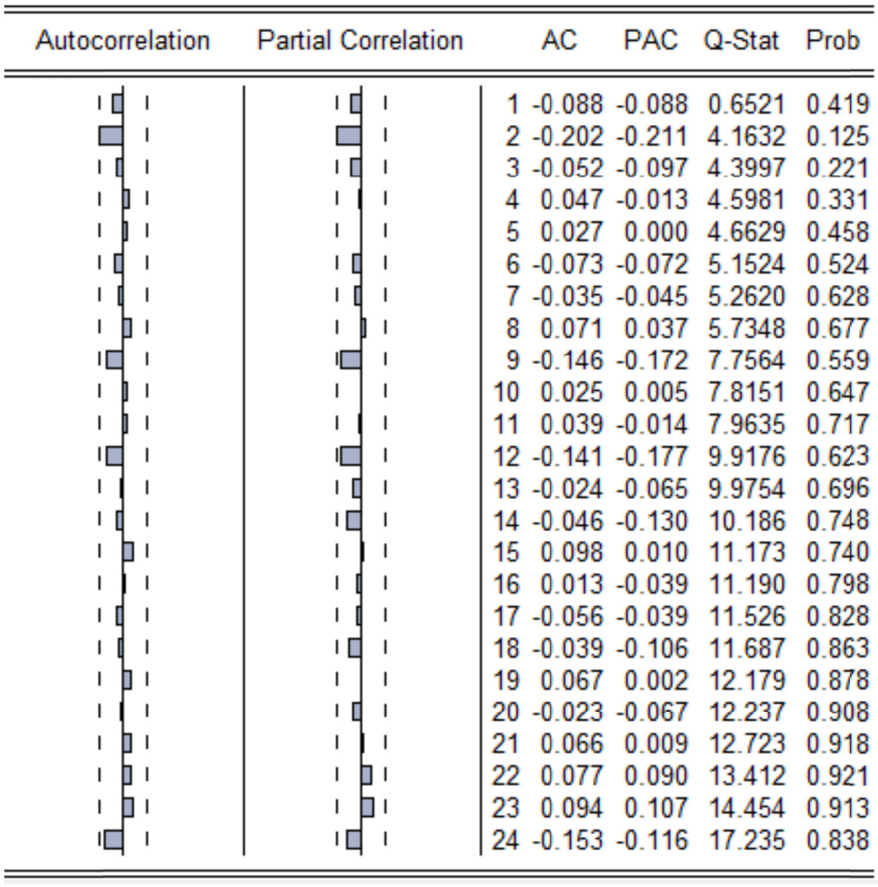
Figure 4. The corelogram for the first order difference of the series

Figure 4 shows the corelogram of the series, with ACF at lag 1, 12 and 24 greater than $Z_{\alpha} / \sqrt{n}$, where α is the significance at 0.05 and n denotes the sample size—so this series has a seasonal property. The SARIMA model is determined for the appropriate order of p, d, and q to determine the right specification—by choosing from both of the difference models, estimated on the information criteria Akaike (AIC) and by generating predictions on the bias of estimate models. The first model starts with the SARIMA$(1,1,2)_{12}$ process and then provides the residual analysis. We found non-normality for the residual term, so the iteration was started until the model was satisfied—cf. Figure 3. Two models SARIMA$(1,1,0)_{12}$ and SARIMA$(0,1,1)_{12}$ are satisfied at stage 2—cf. Table 3. Then considering the residual term in stage 3, SARIMA$(1,1,0)_{12}$ found non-normality, although this model is less AIC and higher R^2 compared to SARIMA$(0,1,1)_{12}$. SARIMA$(0,1,1)_{12}$ is the model that is adequate to forecast the UR, from diagnostic checking with normality, homogeneity of variances, and residual white noise. The Q-stat examines m autocorrelations of the residuals. Figure 5 concludes that the residuals from the SARIMA model have no autocorrelation. As a result, the accuracy of the forecasts are evaluated as RMSE = 0.3177 and MAPE = 16.1076%, i.e. the forecasting is good.

Table 3. The results for two models

Stage 1	Stage 2: Estimation				Stage 3: Diagnostic checking p-value		
Model	coefficient	AIC	DW	R^2	Normality (Jarque-Bera test)	Homogeneity of variance (White test)	Q stat.[a]
SARIMA$(1,1,0)_{12}$	-0.5282	0.493800	1.8965	0.1873	0.001	0.1626	adequate
SARIMA$(0,1,1)_{12}$	-0.3603	0.568853	2.1202	0.1298	0.0689	0.7459	adequate

[a] p-value > 0.05 every lag from Figure 5.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.088	-0.088	0.6521	0.419
		2	-0.202	-0.211	4.1632	0.125
		3	-0.052	-0.097	4.3997	0.221
		4	0.047	-0.013	4.5981	0.331
		5	0.027	0.000	4.6629	0.458
		6	-0.073	-0.072	5.1524	0.524
		7	-0.035	-0.045	5.2620	0.628
		8	0.071	0.037	5.7348	0.677
		9	-0.146	-0.172	7.7564	0.559
		10	0.025	0.005	7.8151	0.647
		11	0.039	-0.014	7.9635	0.717
		12	-0.141	-0.177	9.9176	0.623
		13	-0.024	-0.065	9.9754	0.696
		14	-0.046	-0.130	10.186	0.748
		15	0.098	0.010	11.173	0.740
		16	0.013	-0.039	11.190	0.798
		17	-0.056	-0.039	11.526	0.828
		18	-0.039	-0.106	11.687	0.863
		19	0.067	0.002	12.179	0.878
		20	-0.023	-0.067	12.237	0.908
		21	0.066	0.009	12.723	0.918
		22	0.077	0.090	13.412	0.921
		23	0.094	0.107	14.454	0.913
		24	-0.153	-0.116	17.235	0.838

Figure 5. The Q statistic and white test (homogeneity of variance) for the residuals from SARIMA(0,1,1)12

3.3 Further Comments

From the accuracy criteria with RMSE and MAPE, we formed six different models using two approaches. The most suitable model to forecast the UR in Thailand proved to be SARIMA$(0,1,1)_{12}$. The SARIMA$(0,1,1)_{12}$ has a lower MAPE compared to the ANN1. The training and testing data extended from 2003 to 2011, but more recently we obtained data for the year 2012. The comparison between the actual value of UR and the predicted value from SARIMA$(0,1,1)_{12}$ during January to September 2012 is shown in Figure 6, and the Box-Jenkins model fits the true values quite well.

4. Conclusion and Discussion

Forecasts are very important for policy planning in economic systems. The objective of this study is to find suitable models to predict the unemployment rate in Thailand from two forecasting approaches --- SARIMA from the Box-Jenkins approach, and the Artificial Neuron Network (ANN) method. From MAPE criteria, it is demonstrated that the SARIMA model provides a satisfactory representation of the data. The SARIMA$(0,1,1)_{12}$ shows the good perform and the satisfactory model to forecast the UR in Thailand. The most desirable model was considered, based on the strict assumption of statistical analysis—not only on the actual values but also on the residual values. The weak points of the ANN found in this study are the proper network size, the suitable proportion between the training and testing data set, and memorization. This may be why the ANN approach does not fit the data as well as the SARIMA.

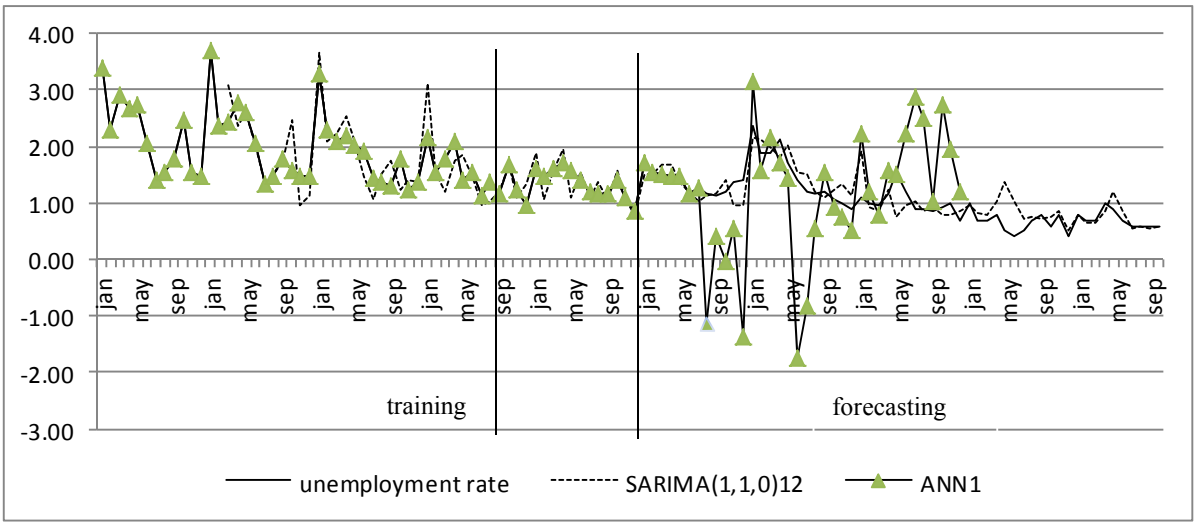


Figure 6. Time plot for the actual value (line), predicted value from ANN1 (line with triangle) and predicted value from SARIMA (dash line)

Acknowledgements

Financial support was provided by the Faculty of Science and graduate studies, Mahasarakham University. The authors would like to thank anonymous reviewers for their comments and suggestions.

References

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis Sorecasting and Control.* New Jersey: Prentice Hall, Englewood Cliffs.

Dobre, I., & Adriana, A. M. (2008). Modeling Unemployment Rate Using Box-Jenkins Procedure. *Journal of Applied Quantitative Method, 3*(2), 156-166.

Kankal, M., Akpinar, A., Kömürcü, M. I., & Özşahin, T. S. (2011). Modeling and forecasting of Turkey's energy consumption using socio-economic and demographic variable. *Applied Energy, 88*(5), 1927-1939. http://dx.doi.org/10.1016/j.apenergy.2010.12.005

Lewis, C. D. (1982). *International and Business Forecasting Methods*. London: Butterworths.

Mustafa, M. R., Bhuiyan, R. R., Isa, M. H., Saiedi, S., & Rahardjo, H. (2012). Effect of Antecedent Conditions on Prediction of Pore-Water Pressure using Artificial Neural Networks. *Modern Applied Science, 6*(2), 6-15. http://dx.doi.org/10.5539/mas.v6n2p6

Sözen, A., & Arcaklioglu, E. (2007). Prediction of Net Energy Consumption Based on Economic Indicators (GNP and GDP) in Turkey. *Energy Policy, 35*, 4981-4992. http://dx.doi.org/10.1016/j.enpol.2007.04.029

Sözen, A., Arcaklioglu, E., & Ozkaymak, M. (2005). Modeling of the Turkey's net energy consumption using artificial neural network. *International Journal of Computer Applications in Technology, 22*(2), 130-136. http://dx.doi.org/10.1504/IJCAT.2005.006944