Available online at http://scik.org

J. Math. Comput. Sci. 4 (2014), No. 2, 350-362

ISSN: 1927-5307

COMPARING THE UNIVARIATE MODELING TECHNIQUES AND BOX-JENKIN FOR MEASURING OF CLIMATE INDEX IN SITIAWAN, MALAYSIA

SHUHAILI SAFEE* AND SABRI AHMAD

Department of Mathematics, School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Malaysia

Abstract: The purpose of the article is to determine the most suitable technique to generate the forecast models using the data from the series of climate index in Sitiawan, Perak. This study are using univariate time series models and box-jenkin consists of Naïve with Trend Model, Single Exponential Smoothing , Double Exponential Smoothing, Holt's Method, Adaptive Response Rate Exponential Smoothing (ARRES), Holt-winter's Trend and Seasonality and SARIMA model. Using time-series data from 1961-2012 (monthly), there's several data consists missing/outlier value. The issues are overcome with applied the time series model for each missing values and then compare the measure error (mean square error, MSE) for each models. Then, the selection of the most suitable model was indicated by the smallest value of mean square error, MSE. Based on the analysis, SARIMA$(0,1,3)(0,1,2)_{12}$ model is the most suitable model for forecasting the climate index in Sitiawan, Perak.

Keywords: univariate time series models; box-jenkin; mean square error, MSE.

2010 AMS Subject Classification: 00A71.

1. Introduction

*Corresponding author

Received January 29, 2014

Forecasting is the whole process of developing the necessary methods to generate the future. As a decision-making tool, forecasting can act as a scanning device that captures the signals of the future outcomes based on either past events or other related factors considered to influence the outcome of event of interest. The information provided thus, would enable the firm to take the necessary actions to modify the existing plans to suite the expected changing environment so as to avoid following loss in revenues in the future.

A climate index defined as a calculated value that can be used to describe the changes and the state in the climate system. Climate indices allow a statistical study of variations of the dependent climatological aspects, such as analysis and comparison of time series, extremes, means, and trends. The prevalent research strategy in the climate-modelling community has been characterized by Knutti (2008), himself a climate modeller, as “take the most comprehensive model, run a few simulations at the highest resolution possible and then struggle to make sense of the result”. The aim is to produce models as “realistic as possible: (Beven, 2002).

2. Preliminaries

The method of data collection is secondary data collection. The data was obtained via the NASA website. Due the period of time in Sitiawan station is longer from 1931 to 2012 (monthly) compare with the other station, then the data is selected. But due to too many missing data, the study started with 1961 to 2012.

Time series are very frequently plotted via line charts. The time series data can be Separation into components representing trend, seasonality, slow and fast variation, cyclical irregular.

Univariate Modelling Techniques divided into two categories is *naïve model* and *exponential smoothing techniques*. The analysis is divided by two categories which are estimation part (fitting part) and evaluation part (hold-out part).

The naïve trend model is modified to take this characteristic into account. The

application of this model is fairly common among organizations. One reason for its popularity is that it can be used even with fairly short time series. The one step ahead forecast is represented as,

$$F_{t+1} = Y_t\left(\frac{Y_t}{Y_{t-1}}\right)$$

Where Y_t is the actual value at time t, and Y_{t-1} is the actual value in the preceding period. The initial value for this model is taking the first initial data from the actual data.

Exponential Smoothing Techniques:

i) *Single Exponential Smoothing Technique* is the simplest form of model within the family of the exponential smoothing techniques. The model requires only one parameter that is the smoothing constant, α, to generate the fitted values and hence the fitted model forecast for the next and all subsequent periods are determined by adjusting the current period forecast by a portion of the difference between the current forecast and the current actual value.

The equation for single exponential smoothed statistic is given as,

$F_{t+m} = ay_t + (1-a)F_t$

Where,

F_{t+m} is the single exponentially smoothed values in period $t+m$

y_t is the actual value in time period t

a is the unknown smoothing constant to be determined with value lying between 0 to 1 $(0 \leq a \leq 1)$

F_t is the forecast made in period t

ii) *Double exponential smoothing* this technique is also known as Brown's method. It is useful for series that exhibits a linear trend characteristic.To demonstrate the method the following notations will be used.

Let,

S_t be the exponentially smoothed value of Y_t at time t

S'_t be the double exponentially smoothed value of Y_t at time t

As usual, firstly input the initial values of the equation by using the equation below:

$$S_t = \alpha Y_t + (1-\alpha)\, St_{-1}$$

Follow by:

$$S'_t = \alpha Y_t + (1-\alpha) S'_{t-1}$$

$$a_t = 2S_t - S'_t$$

$$b_t = \frac{\alpha}{1-\alpha}\left(S_t - S'_t\right)$$

$$F_{t+m} = a_t + b_t * m$$

This double exponential model requires only one parameter, which is smoothing constant, α, to generate the fitted values and hence forecast.

iii) *Holt's method* this technique not only smoothed the trend and the slope directly by using different smoothing constants but also provides more flexibility in selecting the rates at which the trend and slope are tracked. The application of the Holt's Method requires;

The exponential smoothed series: $S_t = \alpha Y_t + (1-\alpha)(St_{-1} + T_{t-1})$

The trend estimate: $T_t = \beta (S_t - S_{t-1}) + (1-\beta) T_{t-1}$

Forecast m period into the future: $F_{t+m} = S_t + T_t * m$

The α and β are the parameters to be determined with values ranging from 0 to 1.

iv) *Adaptive Response Rate Exponential Smoothing (ARESS)* The development of the adaptive response rate exponential smoothing (ARESS) technique is an attempt to overcome the problem that had discussed in all previous exponential smoothing techniques by incorporating the effect of the changing pattern of the data series into the model.

The ARESS technique comprises of the following basic equations;

$$F_{t+m} = \alpha_t Y_t + (1-\alpha) F_t$$

The value of α_t are estimated using the following equations;

$$\alpha_t = \left| \frac{E_t}{AE_t} \right|$$

where:

$$E_t = \beta e_t + (1-\beta) E_{t-1} \qquad ; 0 \le \beta \le 1$$

$$AE_t = \beta Ie_t I + (1- \beta) AE_{t-1}$$

And that,

$$e_t = Y_t - F_t$$

The value of E_t is defined as the smoothed average error, and AE_t is the smoothed absolute error.

Basically, this equation has similar meaning to the single exponential smoothing equation, except for the parameter value alpha (α_t) which is identified by the subscript t. in the application of the ARRES model one does not require to find the best alpha, α. This is because there is no single best α value which happens to vary over time. For this reason the appropriate symbol used is α_t.

v) *Holt-Winter's Trend and Seasonality* is one of the techniques that take into account the trend and seasonality factors. The Holt-Winter's Trend and Seasonality methodology consists of three basic equations: the level component, the trend component, and the seasonality component. The assumptions of multiplicative effect is been made with regard to the relationship of the components of data. On the assumption that the relationship of these components is multiplicative in nature, the equations are represented as follows:

Level Component:

$$L_t = \alpha \frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

Trend Component:

$$b_t = \beta (L_t - L_{t-1}) + (1-\beta) b_{t-1}$$

Seasonality Component:

$$S_t = \gamma \frac{y_t}{L_t} + (1-\gamma) S_{t-s}$$

The m-step-ahead forecast is calculated as:

$$F_{t+m} = (L_t + b_t \times m)S_{t-s+m}$$

The *Box-Jenkins* approach is synonymous with the general ARIMA (Autoregressive Integrated Moving Average) modeling. ARIMA modeling is commonly applied to time series analysis, forecasting and control. The term ARIMA is in short stands for combination that comprises of Autoregressive/integrated/Moving Average models. The basis of the Box-Jenkins modeling approach consist their main stage. These are;

1. Model Identification
2. Model Estimation and Validation
3. Model Application

The basic model of *Box-Jenkins* that involved was *Autoregressive* (AR) model, *Moving Average* (MA) model and *Mixed Autoregressive* and *Moving Average* (ARMA) model.

The Autoregressive (AR) Model defined that the current value of value of the variable was defined as function of previous value plus an error term. In other words the dependent variable, y_t, is taken as the function of the time-lagged value itself.

Mathematically, it $y_t = \mu + \varnothing_1 y_{t-1} + \varnothing_2 y_{t-2} + \ldots\ldots + \varnothing_p y_{t-p} + \varepsilon_t$ is written as,

Where:

μ and $\varnothing_j$ (j=1,2,….,p) are constant parameters to be estimated.

Y_t is the dependent or current value

Y_{t-p} the p^{th} order of the lagged dependent or current value.

ε_t is the error term which is assumed idd with mean = 0 and varience ,

σ_t^2

The Moving Average (MA) Model links the current values of the time series to random errors that have occurred in the previous period rather than the values of the actual series

themselves. The moving average model can be written as:

$$y_t=\mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots\ldots\ldots - \theta_q\varepsilon_{t-q}$$

Where:

μ is the mean about which the series fluctuates

θ's are the moving average parameters to be estimated

ε_{t-q}' s are the error terms (q=1,2,3,…..,q) assume to be independently distributed over time.

The Mixed Autoregressive Moving Average (ARMA) Model was the combination of AR model and MA model and was assume stationary. In other words the series y_t is assumed stationary (no need differencing) and the ARMA model is written as:

$$y_t=\mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots\ldots + \phi_p y_{t-p} - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots\ldots\ldots - \theta_q\varepsilon_{t-q} + \varepsilon_t$$

The Mixed Autoregressive Integrated Moving Average (ARIMA) Model was formulated when the assumption of stationary assumption was not met. The differencing is required to achieve stationary. The general term as ARIMA (*p.d.q*), where *p* represent AR model, *d* denotes number of time the variable y_t needs to be differenced in order to achieve stationary and *q* represented MA model. In case if data is seasonal the model is writing as SARIMA (*p.d.q*) $(P, D, Q)_s$.

Where:

p and *P* denotes the number of significant spikes in the PACF.

q and *Q* denotes the number of significant spikes in the ACF.

d and *D* denotes the degree of differencing involved to achieve stationary in the series.

s number of periods per season

Selecting Best Forecasting Model

Different models generally give different forecast values. A model that gives a very good fit is based on the smallest error measures.

Mean Squared Error (MSE) this measure is commonly used for comparing model's

forecasting performance. It has the tendency to penalize large forecast errors more severally than other common accuracy measures and therefore is considered as the most appropriate measure to determine which methods avoid large errors. Let assume a series $y_1,y_2,\ldots,y_1$ and the m-step-ahead forecasts made at time t be denoted by F_{t+m}. Hence, for the one-step-ahead forecasts (for m=1) of any series, the MSE is written as:

$$MSE = \frac{\sum_t^n e_{t+1}^2}{n}$$

Where: $e_{t+1}=y_{t+1} - y_{t+1}$

Where:

y_{t+1} is the actual observation at the point t + 1

F_{t+1} is the one-step-ahead forecast of y_{t+1} generated from the origin (t=1,2,3,..,n)

n is the number of out-of-sample terms generated by the model.

The data was separated into two parts. First part is called estimation and second part is called evaluation. Estimation part is 3 per 4 of data series, meanwhile for the rest is evaluation part. Since the total data series for climate index in Sitiawan is 622 series, 465 series was used for estimation part and 157 series for evaluation part.

Figure 1.0 display the estimation and evaluation procedure in this study. Basically, there are three stages involved:

i) In this first stage, the series is divided into two parts. The first part is called model estimation part (fitted part) and the second part is the evaluation part (holdout part), which will be used to evaluate the model's forecasting performance.

ii) In the second stage, the models are tested using various forms of functional relationship and variables selections

iii) In the third stage, the models with the smallest MSE are evaluated by comparing the MSE value of each model.

iv) The model that meets all the criteria is thus selected as the most suitable model. The selection criterion is based on the results of comparing their respective error measures.

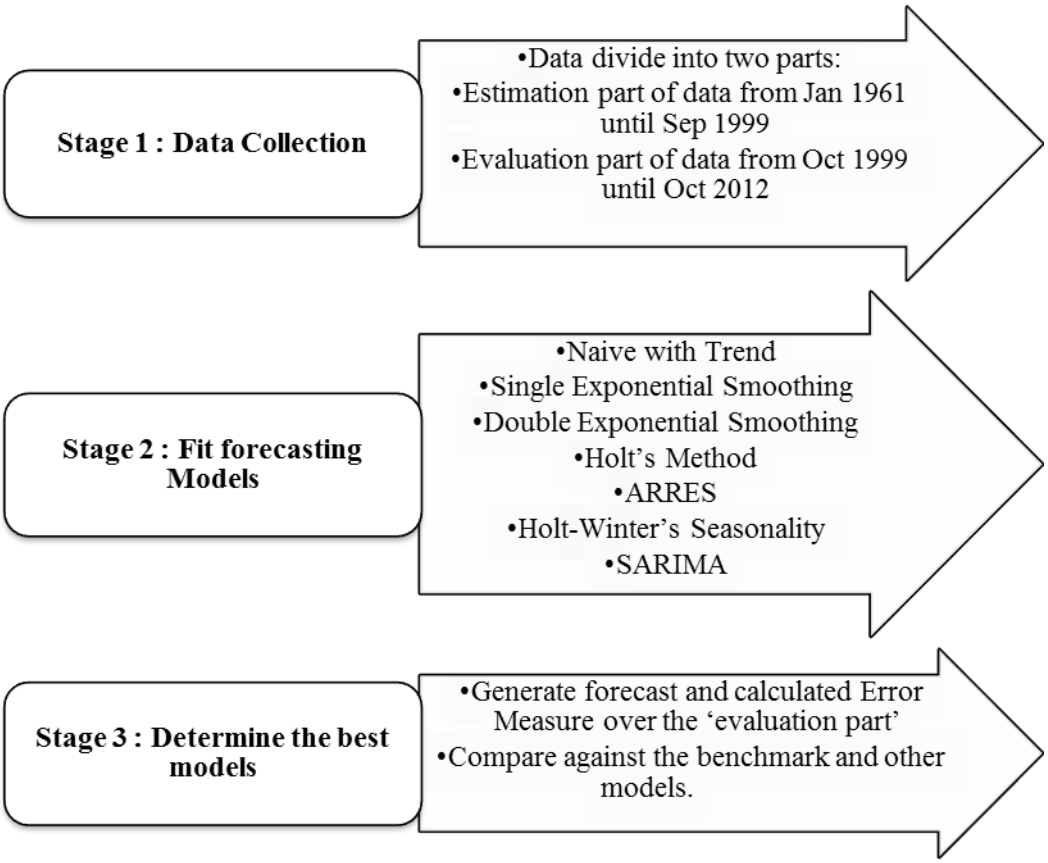

Figure 1.0: Estimation and Evaluation Procedures of Forecasting Model

3. Main results

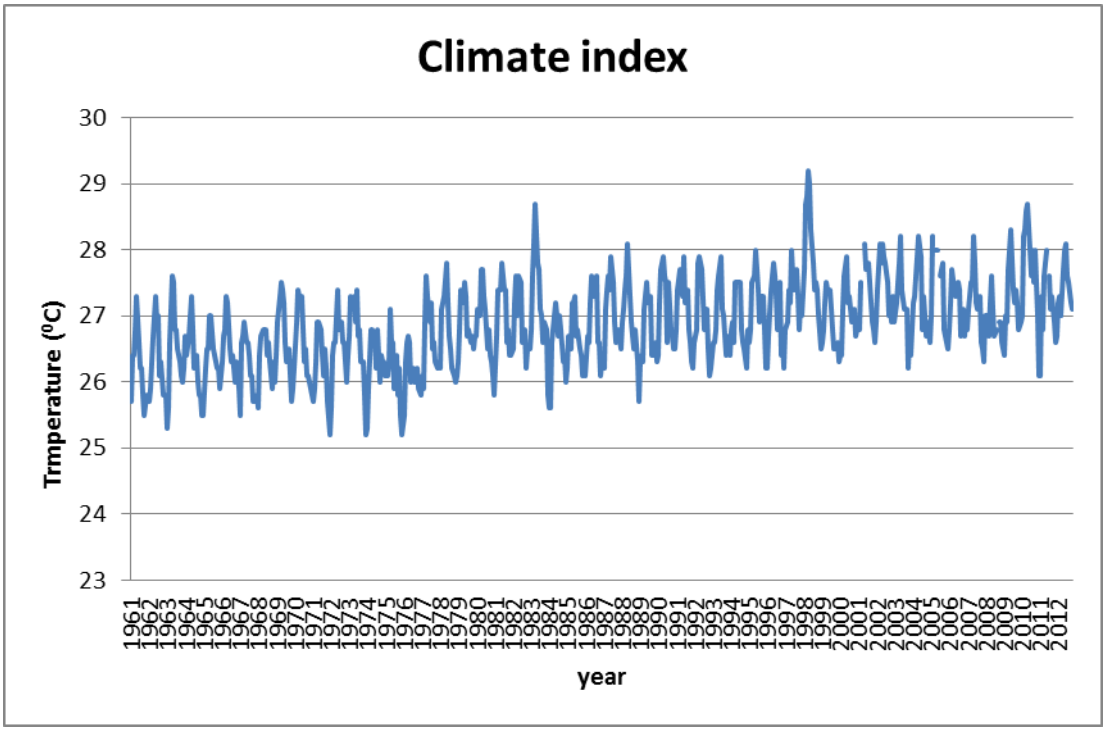

Figure 1.1: The actual graph of climate index, Sitiawan

Based figure 1.1, the data is not stationary, because there is trend line (upward trend) appear. By looking at the flow of the data, there is wave like pattern in the graph, so the data is seasonal.

For the evidence on its seasonality and stationary condition, the graph of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were plotted as in Figure 1.2 and Figure 1.3 respectively. The time plot, ACF and PACF shows a clear seasonal pattern in the data. This is clear in the large values at time lag 12, 24, 36, 48 and 60. The ACF plot slightly decays shows that the series is not stationary. The PACF has a significantly large spike at the lag 1 followed by other smaller spikes at lags more than 1.

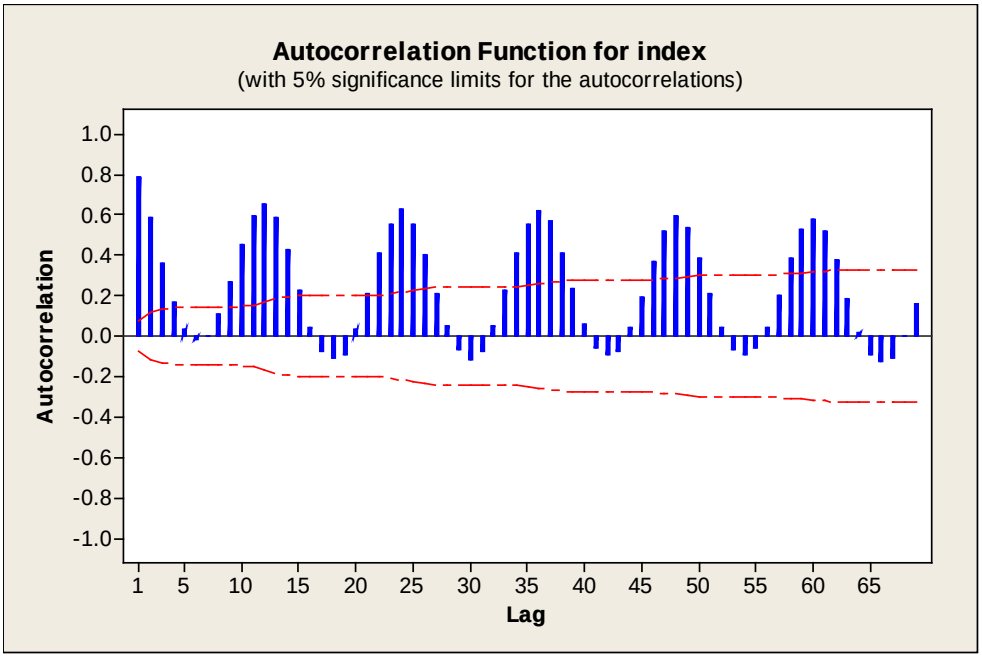

Figure 1.2: Autocorrelation, ACF of climate index, Sitiawan

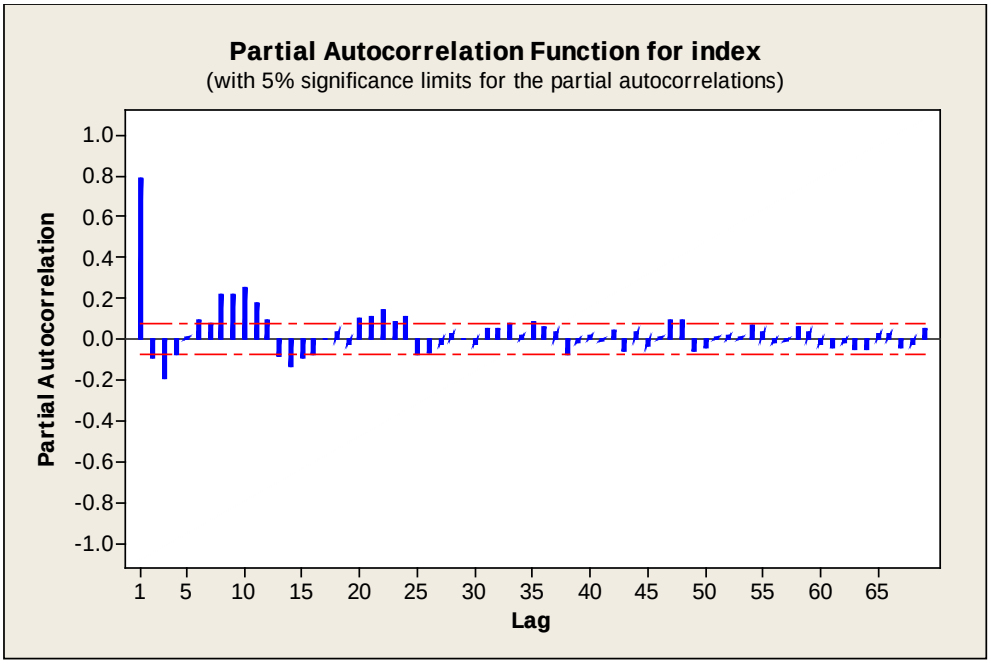

Figure 1.3: Partial Autocorrelation, PACF of climate index, Sitiawan

Table 2: MSE values by type of model

Model type	MSE (without missing value: 1961-2000)	MSE (After compute all the missing values: 1961-2012)
naive with trend	0.3478	0.3780
single explonential smoothing	0.1747	0.1793
double exponential smoothing	0.1878	0.2326
Holt's Method	0.1737	0.1785
ARRES	0.1757	0.1813
Holt-winter's Trend & Seasonal	0.1073	0.1099
SARIMA$(0,1,3)(0,1,2)_{12}$	0.0950	0.0969
SARIMA$(0,1,3)(0,1,1)_{12}$	0.0948	0.0981

By using MSE to identify the smallest error, the model of SARIMA$(0,1,3)(0,1,2)_{12}$ is the best due to value MSE (after compute all missing values of the years 1961 to 2012) are smallest

compare MSE of the other model based on table 2 above.

Using the complete data (without missing values), the estimations were done with the objective of minimising Mean Squared Error (MSE). Results of the corresponding MSE value for each model are shown below

Table 3: Comparison of MSE

Model type	MSE Estimation part: (1961-1999)	MSE Evaluation part: (2000-2012)
naive with trend	0.3515	0.4101
single explonential smoothing	0.1766	0.1918
double exponential smoothing	0.2278	0.2543
Holt's Method	0.1755	0.1921
Adaptive Response Rate Exponential Smoothing (ARRES)	0.1778	0.1957
Holt-winter's Trend & Seasonality	0.1083	0.1125
SARIMA$(0,1,3)(0,1,2)_{12}$	0.0953	0.1014
SARIMA$(0,1,3)(0,1,1)_{12}$	0.0961	0.1014

The criterion used to differentiate between a poor forecast model and good forecast model is called ‘error measure’. Based on the error measures as shown in Table 3, the smallest value in MSE evaluation indicate the same value with SARIMA$(0,1,3)(0,1,2)_{12}$ and SARIMA$(0,1,3)(0,1,1)_{12}$. Since SARIMA$(0,1,3)(0,1,2)_{12}$ also has the smallest value in estimation part, thus the best forecast model is SARIMA$(0,1,3)(0,1,2)_{12}$.

Conclusion

Based on the forecast analysis by using Box-Jenkin, SARIMA$(0,1,3)(0,1,2)_{12}$ model is the most suitable model for forecasting of Climate index in Sitiawan since the value of MSE is

smallest compare with other models for both estimation and evaluation part. The forecast value for best model are shown as below

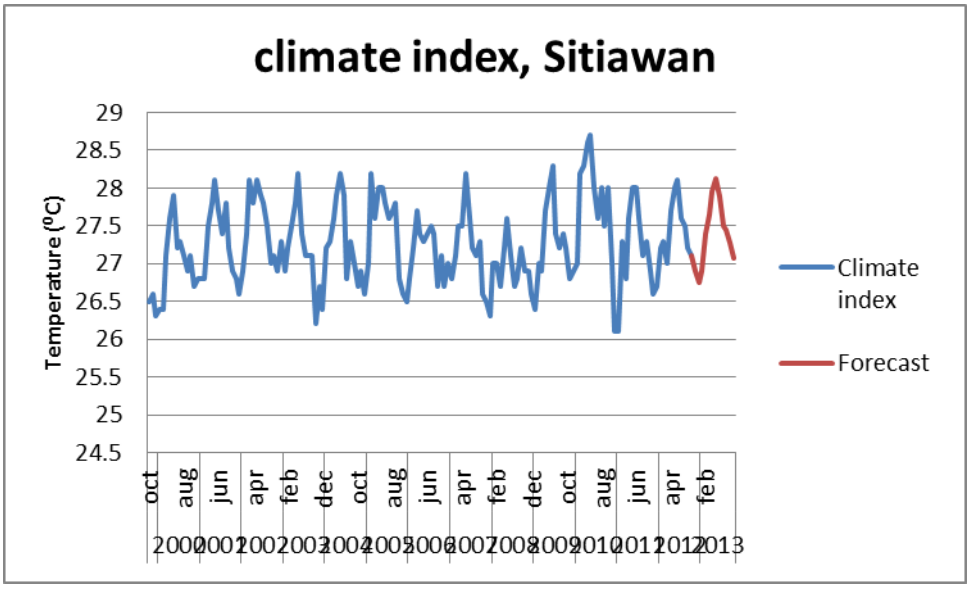

Figure 9: The forecast graph of climate index, Sitiawan

REFERENCES

[1] Beven, K., Towards a coherent philosophy for modeling the environment. Proceedings of the Royal Society a-Mathematical Physical and Engineering Sciences, 458 (2002), 2465-2484.

[2] Knutti, R., Should we believe model predictions of future climate change? Philosophical Transactions of the royal Society a-Mathematical Physical and Engineering Sciences, 366(2008), 4647-4664.

[3] Lazim, M. A. (2011). Introductory Business Forecasting A Practical Approach. University Publication Centre (UPENA).