DEPARTMENT OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE

PROJECT REPORT

SSCU 4902 (UNDERGRADUATE PROJECT 1)

Semester 1, 2020/2021

NAME : **VENESSA TAY SIN YI**

PROGRAMME : **4 SSCM**

RESEARCH TITLE : **FORECASTING CONSUMER PRICE INDEX USING SINGULAR SPECTRUM ANALYSIS**

SUPERVISOR'S VERIFICATION

I declare that I have read through this proposal and in my opinion this proposal has fulfilled the requirements in terms of the scope and quality for the Final Year Project of the Department of Mathematical Sciences.

Signature:

Name of Supervisor:

(Stamp)

DR. NORHAIZA AHMAD
Senior Lecturer
Department of Mathematical Sciences
Faculty of Science
Universiti Teknologi Malaysia
81310 UTM Johor Bahru
Email : norhaiza@utm.my

To examiner: Please return this research report to the student (through the student's supervisor if necessary) for their FYP 2 perusal.

LIST OF ABBREVIATIONS

CPI	-	Consumer Price Index
ARIMA	-	Autoregressive Integrated Moving Average
SSA	-	Singular Spectrum Analysis
PNN	-	Process Neural Network
SVM	-	Support Vector Machine
NARDL	-	Nonlinear Autoregressive Distributed Lag
VECM	-	Vector Error Correction Model
MIDAS	-	Mixed Data Sampling
MSSA	-	Multivariate Singular Spectrum Analysis
VAR	-	Vector Autoregression
R	-	R programming language of statistical software
DOSM	-	Department of Statistics Malaysia

CHAPTER 1

INTRODUCTION

1.1 Introduction/Research Background

Consumer Price Index (CPI) is used to estimate the average price changes of representative basket of goods and services purchased by consumer in a specific time period. CPI can also be called as cost of living index which serves as the economic indicator of inflation in measuring and aggregating the price level and reflect to the country's economy. There are many categories of CPI such as food, clothes, medical care, transportation, education and communication. Nowadays, financial analysts and economists always track CPI indicator because CPI play an important role in economy that affects the economic growth of country. CPI helps the government to determine financial policy such as controlling the nation's money supply and increase the interest rates. From an investor perspective, CPI is a critical measurement on total return and on a nominal basis to meet their financial goals.

According to the trading economics of Malaysia, the increase of CPI is caused by the most major group of food and non-alcoholic beverages with total thirty percent of total weight lead to inflation rate. This show that it is important to make pricing and economic decision to predict future rate of inflation. Consumer price index covers all the category of groups in measuring the pricing and economic rate with the consumption expenditure which are changing from period to period. The price can be collected from the retail stores. The coverage of shops and retail outlets can also be used to determine the purchasing behaviour among the households. In recent years, the importance of internet as the media to purchase consumer goods and services has been rising significantly should be included in a CPI to maintain the consumer purchasing habits. The rapid growth of technology has influenced more households making purchase on the web-based retailers compare to the market stalls.

Economic is the most fundamental determinant the buyers on purchasing behaviour in Malaysia. A study on Malaysian Journal of Management Studies shows that the factor of consumer's purchase decision has huge impact on the retail market in Malaysia. An increase in income surely lead to the rising of consumption expenditure on shopping goods which

greatly improved the quality of life. This situation indicate that CPI is the most appropriate macroeconomic indicator to measure inflation rate in Malaysia. Due to the elevated volatility in the consumer price indices that cause the changes of inflation trends more than one month in advance, therefore it is necessary to employ the forecast technique in predicting the CPI to give a potential insight to the economic sector.

Modelling CPI by using time series approach has been widely applied by many researchers. The historical data from the economic prices has taken into consideration and used to generate a prediction model for the consumer price indices. The most common and extensive model that have been used to analyse the CPI is Autoregressive Integrated Moving Averages (ARIMA). Seasonal ARIMA (SARIMA) model is used to analyse the time series data containing seasonal component to develop a forecasting model. Box-Jenkins ARIMA is the classical time series analysis that used to predict future data that mostly considered in the forecast research. However, this approach is not model-free and have the disadvantage due to its restriction to normality, linearity and stationarity assumption (Silva, 2016). It also may involve the transformation of data that is from non-stationary to stationary data by method of differencing for further analysis. Alternatively, a non-parametric forecasting model which is Singular Spectrum Analysis (SSA) has applied to overcome such restriction that imposed in ARIMA model.

In general, SSA is a powerful tool which can perform two complementary stages known as decomposition and reconstruction. This technique is capable to carry out filtering process which decomposing the original signals into smoothing trend, seasonality and noise series components based on the spectrum of eigenvectors in the singular valued decomposition of the covariance matrix to form the reconstructed one dimensional series for forecasting while this procedure does not present in traditional analysis.

Nowadays, SSA methods has been very important and useful in time series analysis since it shows a great understanding of nonlinear dynamics system underlying time series by inspecting trend, seasonal fluctuation and input signal. Thus, forecasting CPI data lead to inflation dynamics becoming more important particularly for economists, financial analysts, and statisticians. The high accuracy of price indices for future prediction was particularly important to facilitate the decision-makers to make strategic decisions in order to consider

between the decision on output and the actual output of the inflation rate. This study will deal with the time series model of overall CPI data in Malaysia.

1.2 Problem Statement

The most common type that have been used as forecasting techniques is Box-Jenkin ARIMA methodology**.** Nevertheless, there is a contrast in modelling data by using SSA methodology and traditional time series procedure. This is because the classic time series techniques are not model free and not capable in handling linearity, normality and stationarity time series assumptions. Besides, it is not able to carry out any further analysis without involving any transformation of data which from non-stationary to stationary data by method of differencing.

Furthermore, ARIMA model is not applicable in identifying a true signal based on the spectrum of eigenvectors in the singular valued decomposition of the covariance (trajectory) matrix. As a results, it cannot carry out filtering stage in decomposing the original series into extracted trend, seasonality components and reconstructed noise series for smoothing. Hence, SSA takes advantage as an effective implement in CPI modelling and forecasting in the upcoming time.

1.3 Objective of Study

The objectives of the study are as follows:

i. To select a post-period structural break in the CPI time series data based on Chow test to reduce forecasting errors

ii. To construct a trajectory matrix from the one dimensional CPI time series based on the window length

iii. To identify the true signals from noise components based on the spectrum of eigenvectors in the singular valued decomposition of the covariance (trajectory) matrix

iv. To reconstruct the one dimensional time series data for forecasting

1.4 Scope of Study

This study will focus on forecasting the monthly data of consumer price index in Malaysia from January 2005 to October 2020. The data source is obtained from the website https://www.mef.org.my/kc/monthlycpi.aspx?year=2020 which revealed the statistics of consumer price index by Malaysians Employers Federation from the source of Department of Statistics Malaysia (DOSM) official portal. In this study, determination the appropriate window length in changing the dimension of Hankel matrix by the technique of embedding and singular valued decomposition in decomposition stage of singular spectrum analysis in order to predict accuracy of time series. The methods of grouping and diagonal averaging in reconstruction stage are used to estimate the suitable singular value of the parameter as well as carry out the filtration process of signal components results in measuring the stability of model. It is very important to understand that the number of window length, the dimension of trajectory matrix, and the number of singular values of parameter will influencing the sensitivity and precision on inflation when forecasting the data of consumer price index. The statistical programming R software will be used to model the monthly CPI data.

1.5 Summary of Study

This chapter described the background of CPI problem and showed the outline of this thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will review literatures on the consumer price index and the forecasting analysis approaches which include Autoregressive Integrated Moving Average (ARIMA), Vector Autoregression (VAR), Mixed Data Sampling (MIDAS), Multivariate Singular Spectrum Analysis (MSSA), Process Neural Network (PNN), Vector Error Correction Model (VECM), Hybrid Neuro-Fuzzy Model, Support Vector Machine (SVM) and Nonlinear Autoregressive Distributed Lag Model (NARDL).

2.2 Consumer Price Index (CPI)

An actuary is business professional to manage the financial risks normally gets sufficient information regarding a piece of evidence and use statistics to make predictions for the future. Based on Murdipi and Law (2016), a large number of empirical studies have been conducted to investigate inflation on their specific country area or group of countries using various econometric techniques. It is important to study and estimate the future economic growth in CPI that cause the inflation and to understand the factors influence the inflation.

In the study of Young et al. (2004), CPI measures changes in the prices of goods and services that households consume. Such changes affect the real purchasing power of consumers' incomes and their welfare. It is also widely used as a proxy for a general index of inflation for the economy as a whole, partly because of the frequency and timeliness with which it is produced. The main factor influence CPI is consumer behaviour. For example, there are some of the consumer need to consider pricing, stores choices, brand or quality to look for their products leads to higher inflation. Malaysia's official statistics states that the rise in the CPI was due to increases in alcoholic beverages and tobacco by 22.6%, miscellaneous goods and services (+5.2%), food and non-alcoholic beverages (+4.8%), furnishing, household equipment and routine household maintenance (+4.7%), restaurants and hotels (+4.7%), and

health (+4.5%). There are many data to be used in estimating the model. Hence, identifying the most appropriate time series forecasting model is very important.

2.3 Autoregressive Integrated Moving Average (ARIMA)

In statistics and econometrics, ARIMA is one of the most popular time series forecasting model. Based on Wang and Zhao (2009) paper, ARIMA has been originated from the autoregressive model (AR) proposed by Yule in 1927, the moving average model (MA) invented by Walker in 1931 and the combination of the AR and MA, the ARMA (p, q) models. ARIMA uses particle swarm optimization algorithm (PSO) which is a complex traditional estimation to forecast the monthly CPI of thirty-six big or medium-sized cities in China. PSO can be implemented with ease and has a powerful optimizing performance is employed to optimize the coefficients of ARIMA. ARIMA (4, 3, 1) formed from the number of autoregressive terms, the number of non-seasonal differences and the number of lagged forecast errors are obtained after plotting ACF and PACF graphs are drawn based on 1~M lag numbers which provide information about ARMA (4, 1) since the autocorrelation coefficient starts at a very high value at lag one and then statistically rapidly declines to zero while PACF up to four lags are individually statistically significant different from zero but then statistically rapidly declines to zero. The method of moment estimation by Yule-Walker equation and linear iterative method to approximate the results of the parameter with MSE value is 1.3189. Then, PSOARIMA model used to optimize the coefficient of ARIMA with range $[-1, 1]$ and obtained the value of MSE is 1.2935. This results showed that the accuracy of PSOARIMA (4, 3, 1) model is better than the ARIMA (4, 3, 1) model because the relative error of PSOARIMA (4, 3, 1) is smaller than ARIMA (4, 3, 1) in general. Similar study was also conducted by using Box-Jenkins method to forecast annual datasets of CPI in controlling inflation for the upcoming 10 years. This study has gathered the data from Belgium CPI data from year 1960 to 2017. The findings of this study shows that ARIMA (0, 2, 1) is best fit to CPI data and there is an upward trend for the forecast period (Nyoni, 2019). Other than that, Boniface and Martin (2019) carried out their research on forecast CPI data from Ghana which CPI act as the indicator in influencing economic plan in order to affect the purchasing rate of the residents. This research used the monthly data series from March 2013 to November 2018 by applying the SARIMA model. Based on the analysis, SARIMA (2, 1, 1) (1, 0, 0) as most fitted time series model to forecast the CPI for next nine months and the parameters are estimated by Minitab software. There is

another study found that the implementation time series ARIMA method in forecasting CPI data using monthly data from January 2005 to December 2015 of Indonesia (Ahmar et al., 2018). The findings revealed ARIMA (1, 0, 0) model is most suitable to fit the CPI data using forecast package in R software and the efficiency of this forecast models is measured with RMSE and MAPE which is 5.695 and 1.625. Besides, Mia et al. (2019) researched the development of ARIMA approach to estimate the CPI data from period 2019 to 2025 by applying the Akaike information criteria (AIC), corrected Akaike information criteria (AICc) and Bayesian information criteria (BIC). The results of analysis indicate that there is an upward trend shown by the best model of ARIMA (2,2,0). Another research paper by Akpanta and Okorie (2015) have applied SARIMA model in analysing CPI average monthly data of Nigeria from January 2014 to December 2015. The ACF and PACF plotting graph initially showed that non-stationary CPI data are determined by Augmented Dickey-Fuller test then followed by carried out the diagnostic plot of ARIMA and seasonal ARIMA model. The t- test statistics results have discovered SARIMA (1, 2, 1) (0, 0, 1) is the best fit to CPI data because there does not exists any significance different between the observed data and predictive values when 5% of significance level is used.

2.4 Vector Autoregression (VAR)

A study was conducted by Murdipi and Law (2016) on dynamic linkages between price indices and inflation in Malaysia by using monthly data among CPI, Producer Price Index (PPI), Industrial Production index (IP) and Import price index (IM) of sample period spans from January 2005 to December 2013. In this research paper, unit root test is the time series analysis to determine whether the variable in stationary or not as well as the integration order. This analysis required Augmented Dickey Fuller (ADF) and Phillips-Perron (PP) unit root tests respectively. The Schwartz Bayensian Criterion (SBC) was also applied to determine the appropriate lag lengths of the models and the results revealed that all series are non-stationary and integrated of order one. Moreover, this study employed Johansen Multivariate Co-Integration Test to examine the existence of the long-run relationship among the variables within a multivariate framework which is a vector autoregressive (VAR) based test of restrictions imposed by co-integration on the unrestricted VAR carried out by two test statistics namely trace and maximum eigenvalue used to determine the number of co-integrating vectors. Once the co-integrating relationship is present, then analyse the short-run Granger Causality

test which incorporate the error correction term (ECT) for the adjustment for the deviation from its long run equilibrium. This modified model can refer to Vector Error Correction Model (VECM) framework based on the first lag differences variable. However, the analysis may be conducted as a standard VAR model if co-integration does not exist. Therefore, the findings have shown the industrial production and import price are statistically significant determinants of CPI in the long run indicates that 1% increase in IP and IM will result 0.237% and 0.303% increase in CPI respectively. Generalized impulse response function (GIRF) and variance decomposition (VDC) also been applied to trace the impact of a one standard deviation shock on variables in the system and to estimate the forecast error variance due to shocks or innovations in other variables in providing dynamic interaction of variables. The findings indicate the response of CPI to a one standard error which is positive shock in IP, PPI, IM, oil and money supply. This implies that the CPI responds positively and is statistically significant to shocks in PPI, IM and money supply.

2.5 Mixed Data Sampling (MIDAS)

According to Harchaoui and Janssen (2018), a study on how the big data Billion Price Project (BPP) enhance the timeliness of official statistics on consumer price index in US have conducted using MIDAS time series model that accommodate data sampled at different frequencies. These models generate estimates that remain robust to the variety of time periods considered in the forecast accuracy of official consumer price inflation figures. BPP CPI data runs from 1 July 2008 to 31 July 2015 with only a three-day lag with respect of the reference month to predict the official CPI that is available with an average of three-weeks delay from the reference month. Under the MIDAS approach, a lower-frequency dependent variable is regressed on a higher-frequency lagged independent variable. There are three polynomial weighting function used for baseline estimates and robustness checks to determine the order of polynomial degree which are Almon polynomial lag, beta polynomial lag, and exponential polynomial lag. The forecast results from the MIDAS model showed has a lower root mean square error (RMSE) compared to those from the autoregressive model with one lag that is considered as the benchmark model. In addition, Qiu (2020) have also conducted a research in forecasting the consumer confidence index (CCI) with tree-based MIDAS regressions to predict US economic activity. CCI consists of nine regular predictors which are five monthly macroeconomic variables and four daily financial variables. The monthly data from January 2013 to March 2017 while the daily data span from 1 January 2013 to 22 March 2017 where

the monthly confidence data are related to daily, or even hourly measures of US sentiment index (USSI) to predict changes in the Conference Board's CCI. Nonlinear least squares method used to estimate on the aggregate effect of lags and MIDAS-Almon method used for varying the weight lag polynomial. Heterogeneous MIDAS (H-MIDAS) method takes weighted averages of lagged high-frequency USSI observations to enhance parameter stability across various horizons and permits a weaker discounting of the very past sentiment data. Ordinary Least Square (OLS) method was used to estimate the lag weight when define weighted regressors and weight vector with constraints. The paper examined that an improvement to generate stable forecasts by identifying the lowest of mean squared forecast error (MSFE) and mean absolute forecast error (MAFE) are the Bagging Tree (BAG), Random Forest (RF) method, Least Squares Boosting Tree of RT ensembles (BOOST) and Support Vector Regression (SVR). The results revealed that MIDAS-BAG can improve the forecast accuracy by up to 52.5% relative to the best-performing conventional MIDAS method.

2.6 Multivariate Singular Spectrum Analysis (MSSA)

Hassani and Silva (2018) have applied the MSSA model in forecasting UK consumer price inflation to improve the accuracy of prediction. The historical monthly data from January 2006 to May 2018 collected via the Office for National Statistics used to generate the forecast data. This paper used a variety of parametric and nonparametric test to optimize univariate forecasting model which are Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing (ETS), Neural Network (NN), Trigonometric Box–Cox ARMA Trend Seasonal (TBATS). ARIMA model was applied to generate univariate forecasts for UK consumer price inflation with Akaike Information Criterion (AIC) which minimised the number of seasonal difference, d and determine the number of autoregressive terms, p and the number of lagged forecast error, q. ETS method was used to consider the error, trend and seasonal components by optimizing initial values and parameters using Maximum Likelihood Estimator (MLE) in selecting the best model based on the AIC. The selected parameter based on a loss function embedded in nnetar algorithm trains 25 networks by using random starting values and obtains the average of the resulting predictions to compute the forecast in NN model. TBATS technique aimed to provide accurate forecasts for time series with complex seasonality. Then, select the best performing forecast within the Multivariate Singular Spectrum Analysis with Auxiliary Information, MSSA(AI) framework forecast after generating from univariate model. MSSA model begins with the decomposition stage which has two steps known as

embedding and Singular Value Decomposition (SVD). MSSA(AI) with the help of Vertical MSSA Recurrent (VMSSA-R) and Vertical MSSA Vector (VMSSA-V)) consider data with different series lengths and a time lag in the future for developing an improved accuracy multivariate forecast distinguished based on the RMSE and the ratio of the RMSE criteria. The results indicate that the MSSA(AI) forecasts UK consumer price inflation are statistically significant better than the forecasts from ARIMA, NN, and BATS models in a long run. Furthermore, Hassani et al., (2013) had proposed using univariate Singular Spectrum Analysis (SSA) and MSSA to predict inflation dynamics in USA. There are two indicators that used for the inflation prediction which are CPI and gross domestic product (GDP) price index. The indicator of CPI used the datasets from January 1986 to December 1996 while the GDP index used the datasets from 1970 to 1985. The model was carried out in short run by quarterly and long run by yearly. The accuracy of prediction was measured by root mean square errors (RMSE) and statistical significance test namely Diebold-Mariano was performed in order to measure the forecast performance on the rate of inflation. In this research paper, the result shows that SSA and MSSA are statistically significant outperforms to other methods in forecasting inflation and price indices. From the empirical results, it can conclude that MSSA is the best performance in predicting the direction of change at 1% of confidence level and consider that MSSA is a most effective approach in economic forecasting. Next, Caporale and Skare (2018) had proposed about the application of univariate and multivariate of SSA in Netherlands for analysing Gibson's paradox from the period of 1800 to 2012 which the data observations including 73 macroeconomics variables. This study analyse the co-movement between the long term and short term of interest rates and CPI. The spectrum was used to examine the correlation between the interest rates and CPI and to shows the inflation dynamics pattern of statistical framework. This technique used three types of spectral measures namely Squared Coherency, Gain and the Phase spectrum and the accuracy of prediction is measured by MSE. This approach also to perform decomposition of the interest rates and CPI data into the oscillatory components in a long term bond yields. Based on the spectral analysis, it displayed that the interest rates and CPI are highly correlated both in the short and long run by using coherency squared function. Besides, the MSSA shows that it improved the accuracy of one-step ahead forecast compared to the forecast by univariate SSA.

2.7 Process Neural Network (PNN)

Ge and Yin (2018) conducted a research on consumer price index prediction using time series analysis which is the application of process neural network. The China's CPI historical data are composed of eight categories residents' basic consumption data, and has very strong nonlinear characteristics. The seasonal and non-seasonal factors have affected the monthly economic time series. The first structure of PNN is weighted operator can be time-varying has an advantage in CPI short-term prediction, the second structure of PNN is aggregation operator is composed of multi-input aggregation in space and cumulative aggregation of time and the third structure is activation operator. The input function vector of structures in this process is weight function and activation function which may take linear function, Sigmoid function and Gauss-type function. The topology structure of feedforward PNN has the input of time-varying function, the constant output, and the topology structure of network is n-m-L-1. Concrete measures were adopted to improve overall CPI prediction accuracy. The normalized raw data was directly expressed as a set of orthogonal basis expanded form to reduce errors and speed up network convergence. In PNN, the concrete implements of combined type improved BP algorithm have applied momentum method, adaptive learning rate method and steepness factor method to get better convergence accuracy and test errors. The results revealed that the training parameters selected are 8 input nodes, 100 is the first hidden layer nodes, 1 is the second hidden layer nodes, learning rate is 0.01, the largest iteration number is 10000, learning accuracy is 0.001, momentum factor is 0.8 activation function of the first hidden layer is the tangent sigmoid function, activation function of the second layer is a linear function, the minimum run gradient is 1e-010. The network was terminated after 305 generations because the gradient does not meet the minimum run gradient value. The average relative error of test samples in the traditional neural network model is -4.591% and the average of relative errors absolute value is 4.591% respectively. This implies that a multivariable and nonlinear PNN provides a short term prediction of time series compared to traditional neural network prediction model.

2.8 Vector Error Correction Model (VECM)

Venkadasalam (2015) used Augmented Dicker-Fuller (ADF) to examine the stationarity of CPI with the macroeconomic variables then estimated the data by VECM model to determine the long term equation in identifying the long run causality results. The purpose of the study is to show the significance of broad money, export of good and services, gross

domestic product, and household final consumption expenditure to the CPI on long run economy. The output of this study indicate that VECM model faced restriction on short run interactions between the CPI and others variables.

2.9 Hybrid Neuro-Fuzzy Model

In addition, a study on forecast of inflation level from US markets by using Hybrid Neuro-Fuzzy system of neural network has been conducted (Enke & Mehdiyev, 2014). This study used average monthly CPI data of 169 data values which are from January 2000 to January 2014 to predict the future change of CPI for the upcoming 12 months. There are three stages have applied in this model which are data collection and analysis, subtractive clustering and fuzzy inference neural networks. Graphical representation showed clearly for the hybrid model in CPI prediction and is measured with the root mean square error (RMSE) which yield the lowest RMSE value with 0.829 compared to others model.

2.10 Support Vector Machine (SVM)

Qiong and Zhong (2013) researched the development of a model to estimate the residents of China's CPI yearly data from 2010 to 2012 by using the idea of least squares SVM model with genetic anneal simulation algorithm. This model performs three stages in adjusting the accurate parameters and optimizing characteristics which are embedding, nested-based hybrid framework and improved genetic algorithm. The efficiency of the forecast models is evaluated using the average relative error which compare between the actual and prediction value. The results exhibit the predictive accuracy performance of SVM in time series analysis is relatively large in contrast with grid network search method with 95% to 87% respectively.

2.11 Nonlinear Autoregressive Distributed Lag Model (NARDL)

Alsamara et al., (2020) modelled the import price index into CPI and inflation rate in Gulf Corporation Countries (GCC) by using the NARDL approach in order to examine the

short run and long run asymmetric pass through. The nonlinear empirical analysis indicates that gross domestic products on CPI have highly influence the overall economic growth of GCC in a long run. The dynamic multiplier was used to reveal the long run pattern and make a stability adjustment of CPI data to a new equilibrium state after a shock. This technique is indeed impact the exchange rate on domestic consumer price over both short- term and long-term horizons.

2.12 Summary

There are many studies that have been applied in CPI forecasting by using different techniques. Since Singular Spectrum Analysis is the powerful tool that used in predicting CPI, thus, we will develop the univariate nonparametric forecast model to predict the monthly average consumer price indices in Malaysia in this paper.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter is focused on the research methodology. The steps of embedding and singular valued decomposition are used in choosing the suitable window length in changing the dimension of the Hankel matrix in decomposition stage of singular spectrum analysis. Then, the selection of the singular value of parameter and signal filtration process by the grouping and diagonal averaging methods required to carry out to maintain the stability structures in reconstruction stage. The forecast accuracy of model should be measured based on identifying the appropriate parameters and approximating the dimension of rank of trajectory matrix. The method mentioned above will be further discussed in this chapter.

3.2 Research framework

Figure 3.2.1 shows research framework.

3.3 Stationarity

Augmented Dicker-Fuller (ADF) test is used to check the unit root for stationarity in time series analysis (Stephanie, 2016). The DF test statistic formula is:

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \tag{3.1}$$

There exists a unit roots and the data is non-stationary when the null hypothesis is rejected where $\gamma = 0$. This indicate that the p-value is less than 0.05 or the DF_τ statistic is more negative compared to the critical value of the Dicker-Fuller t distribution. Besides, the autocorrelation function (ACF) was also be used in identifying and interpreting the correlogram for stationarity

of data by using the formula defined as below:

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t-k}-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2} \quad (3.2)$$

where r_k is the autocorrelation coefficient for a k period lag, $\bar{x}$ is the mean of the time series, x_t is the value of the time series at period t and x_{t-k} is the value of time series k period before period t.

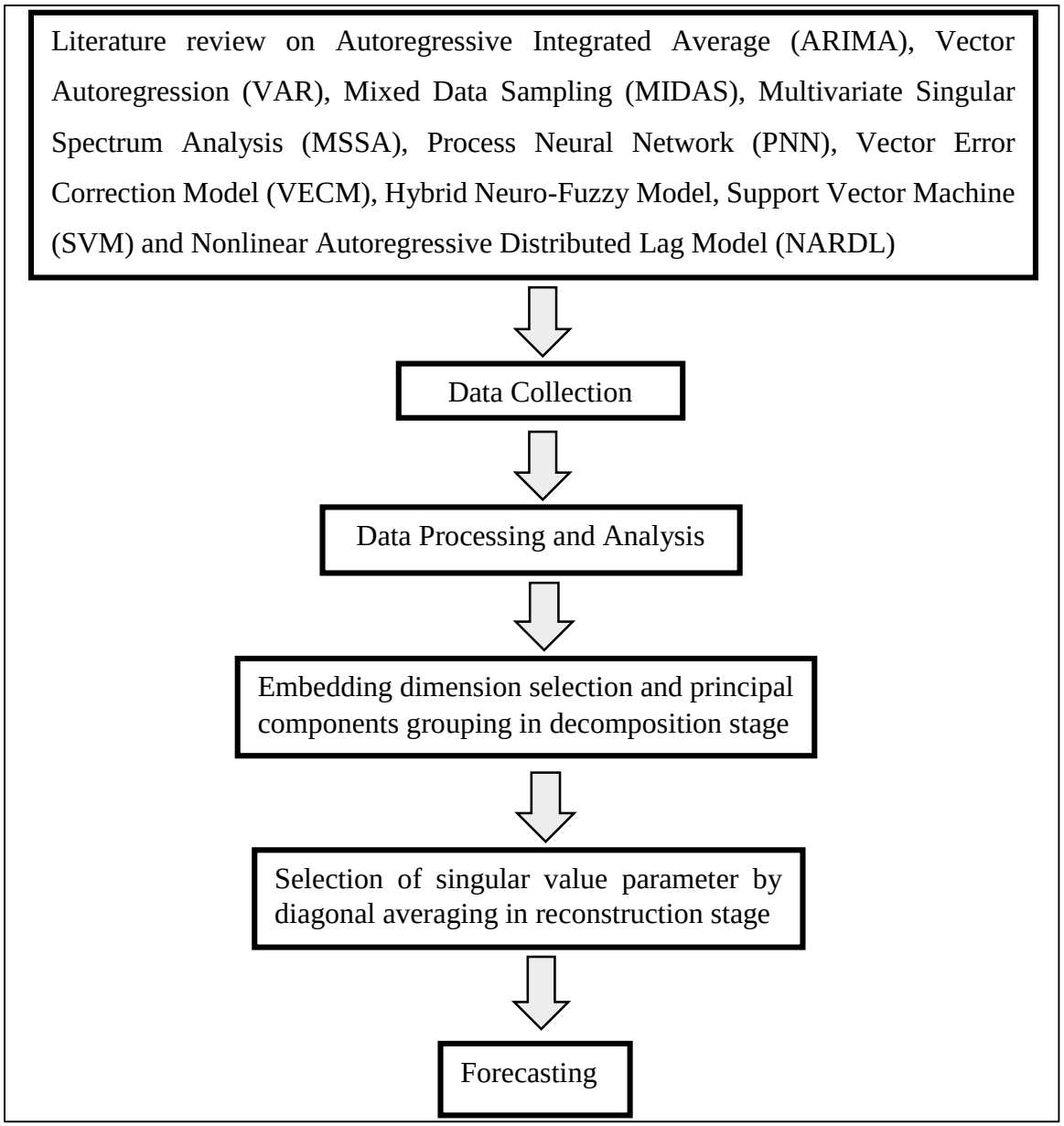


Figure 3.2.1 Research Framework

3.4 Structural Break

In identifying whether there is a structural break in the model, Chow test statistics are computed. Chow test is a useful tool that commonly used to analyze the changes in periods of time series data. The splitting of a model into samples at a prior breakpoint in the two linear regression equations as follow:

$$y_t = X_1 b_1 + \mu_1 \quad (3.3)$$

$$y_t = X_2 b_2 + \mu_2 \quad (3.4)$$

If the coefficients are equal which are $b_1 = b_2$ and $\mu_1 = \mu_2$, then the datasets can be defined as there exists a single regression line (Stephanie, 2016). The null hypothesis is the data has no breakpoint and the F statistical test is calculated by using the formula as shown:

$$F = \frac{(RSS_p - (RSS_1 + RSS_2))/k}{(RSS_1 + RSS_2)/(N_1 + N_2 - 2k)} \quad (3.5)$$

where RSS_p is the pooled regression line, RSS_1 is the regression line before break and RSS_2 is the regression line after break. The null hypothesis is rejected when the calculated p-value of F statistics value is lower than the critical value of 0.0005.

3.5 Singular Spectrum Analysis

Singular spectrum analysis (SSA) is a powerful tool of statistical analysis in forecasting consumer price index data. Based on research paper from Sun and Li (2017) states that the paper of SSA model is first published by Broomhead and King in 1986. Gill, Vautard and their colleagues deal with the paper on the methodological aspects of analogy between the trajectory matrix and the application of SSA. The natural extension from univariate framework to multivariate SSA (MSSA) take advantage in obtaining similar idea as SSA with larger matrices for multi-vintage data. SSA is mainly based on matrix and a univariate analysis which can decompose a time series data into many component parts such as trend, seasonal, cyclical and random. Since SSA is a nonparametric model, therefore it is well appropriate for exploratory analysis of time series. SSA can be applied in solving many types of problems such as time series decomposition, trend extraction, noise reduction, parameter estimation, signal extraction,

data mining and forecasting. This revealed that SSA method is slightly difference when comparing with the conventional time series analysis. The powerful model-free technique of SSA is unrestricted with any pattern of time series data. For instance, SSA enable to generate the forecast by using less volatile data or does not have to assumed the stationarity-type condition for the time series. The main purpose of this method is to predict the nonlinear dynamics for reconstructing the attractor of a system in time series analysis.

This statistical technique then is widely used in dealing with spectral profile which can illustrate how the features can be applied to analyse the extensive signals in order to attain narrow band signals or to deduce the rate of occurrence. SSA is usually used to reduce the additive noise level and detect damage on the structural components that containing noise in processing measurement of nonlinear vibration systems. In addition, SSA also simply served to extract the underlying deterministic dynamics and nonlinear filtering. This method is incorporating with the elements of traditional time series analysis, multivariate statistics and signal processing. Indeed, it is often to be applied in meteorological, economics, social sciences, financial mathematics and dynamical systems.

3.5.1 Procedure of Singular Spectrum Analysis

The powerful nonparametric SSA method consists of two important complementary stages which are decomposition and reconstruction (Osmanzade, 2017). Both together incorporate with two separate steps. At the beginning stage, decompose the series into the sum of components such as trend, periodic and noise while at the second stages carry out reconstruction on the original series then apply the resulting reconstructed components for forecasting new data points by filtration and parameter estimation within SSA framework. In this paper, SSA methodology is being described and this implementation is used for analysis of real world statistics time series data. Then, it is adaptable in applying this useful technique to the original time series which is the monthly data of consumer price index in Malaysia.

(I) Decomposition stage

Based on the singular spectrum of time series analysis, two steps are required in the decomposition stage of SSA namely Embedding and Singular Valued Decomposition (SVD). Embedding is the fundamental process of mapping the sample time series into a vector space

of multi-dimensional time series, $X_1, X_2, \dots, X_k$ with lagged vectors $X_i = (x_i, \dots, x_{i+L-1})^N$, $i = 1, \dots, K$ where $K = N - L + 1$ and the embedding single parameter is the window length, L in integer $2 \ll L \ll N$. The selection of window length for constructing the trajectory matrix is very important because it gives distinct structure observed behaviour of a dataset. Large window length, L must be considered but is not greater than $N/2$ to ensure that the measurement on signals and noise components are clearly separated.

Define the trajectory matrix, X:

$$X = [X_1, X_2, \dots, X_k] = \left(x_{ij}\right)_{i,j=1}^{L,K} \tag{3.6}$$

where X is the Hankel matrix. There are two properties of Hankel matrix which are both its rows and columns of X are subseries of the original time series and all the elements of X along the anti-diagonals are equal to each other. Then, SVD procedure is applied in analysing the accuracy of singular values in Hankel matrix in order to change the dimension of matrix on eigenvalues. Hankel matrix play an important role in orthogonal polynomial theory, stability theory and model reduction. Denote $\lambda_1, \lambda_2, \dots, \lambda_L$ are the eigenvalues of matrix XX^T in non-increasing order of magnitude where $\lambda_1 \geq \dots \lambda_L \geq 0$ and define $U_1, U_2, \dots, U_L$ are the corresponding eigenvectors in orthogonal system. The adjustment of eigenvalues and eigenvectors is required to reduce the effects in separating the signals from noise component. In general, if denote principal components, $V_i = {X_i}^T U_i/\sqrt{\lambda_i}$, then the SVD of trajectory matrix, X written as

$$X_1, X_2, \dots, X_L \tag{3.7}$$

where $X_i = \sqrt{\lambda_i} U_i {V_i}^T$, $\sqrt{\lambda_i}$ is the spectrum of matrix X_i , $P_i = U_i$ of elementary matrices is left singular vector and $Q_i = \sqrt{\lambda_i} V_i$ is right singular vector. The trajectory matrices of X_i have rank X where $d = \max(i, such\ that\ \lambda_i > 0)$. The eigentriple matrix is then called $\left(\sqrt{\lambda_i}, U_i,\ V_i\right)$. The expansion of SVD (3.2) is unique defined when all the eigenvalues have multiplicity of one. The SVD of matrix X, $\sum_{i=1}^{r} X_i$ also create handle in approximating the optimal minimum rank matrix of $\left|\left|X - X^{(r)}\right|\right|$ where rank, $r < d$. Hence, consider the characteristics of the contribution from expansion of SVD (3.2) has a ratio of $\lambda_i / \sum_{i=1}^{d} \lambda_i$ and the characteristics of

the optimal approximation of X_i by the trajectory matrices of rank, r has the sum of the first r ratios, $\sum_{i=1}^{r} \lambda_i \,/\, \sum_{i=1}^{d} \lambda_i$.

(II) Reconstruction stage

After evaluating all the possible combination of rank, r where $(1 \ll r \ll L-1)$ by SVD of Hankel matrix, then the process of eigentriple grouping and diagonal averaging are applied in reconstruction stage. Splitting elementary matrices of X_i into a few groups and summing the matrices within each group. If let $I = i_1, \dots, i_p$ for $p < L$ be a group of indices $i_1, \dots, i_p$, then the matrix of X_I is corresponding to the group I defined as $X_I = X_{i1}, \dots, X_{ip}$. Partition the set of indices $\{1, \dots, L\}$ into diagonal subsets $I_1, \dots, I_m$ and this separation lead to the following decomposition representation:

$$X = X_{I1}, \dots, X_{Im} \qquad (3.8)$$

The procedure when selecting $I_1, \dots, I_m$ called eigentriple grouping. For a given group I, the contribution of the component X_I is measured by the share of eigenvalues $\sum_{i \in I} \lambda_i \,/\, \sum_{i=1}^{d} \lambda_i$. After that, the steps of extraction signal on filtration process is very important in performing diagonal averaging by transforming a matrix to the form of Hankel matrix in order to choose the singular values of parameter r which can then subsequently converted into new time series. If z_{ij} stands for the element of a matrix Z, then the kth term of the resulting series is obtained by averaging z_{ij} for all i, j such that $i + j = k + 1$ and this called as Hankelization of matrix Z. After apply Hankelization procedure to matrix components by performing diagonal averaging, another expansion of matrix $X = \widetilde{X_{I1}} + \cdots + \widetilde{X_{Im}}$ obtained where $\widetilde{X_{Ij}}$ is the diagonal version of matrix X_{Ij} and $\widetilde{X_{I1}} = HX$. This is equivalent to the initial series $x_1, \dots, x_N$ and decomposed to the sum of m series $y_n = \sum_{k=1}^{m} \widetilde{y_n}^{(k)}$, $n = 1, \dots, N$ which corresponding to the matrix $\widetilde{X_{Ij}}$. The resulting series constructed by the elementary grouping will called as elementary reconstructed series.

3.5 Summary

This chapter focussed on SSA procedures in CPI forecasting. SSA consists of two stages which are decomposition stage and reconstruction stage. In decomposition stage, the embedding and singular value decomposition process was used to select the appropriate window length in changing the dimension of the Hankel matrix. Then, SSA carry out grouping and diagonal averaging techniques in reconstruction stage to identify the appropriate parameters and approximating the dimension of rank of trajectory matrix. It also performs filtration process in extracting trend, seasonality and noise components to form reconstructed series. The reconstructed series was then used for forecasting.

CHAPTER 4

ABOUT CPI TIME SERIES DATA

4.1 Introduction

In chapter 4, we will discuss on the time series behaviour, test for stationarity, structural break analysis and post-period structural break selection on CPI data.

4.2 Data Description

In this study, we will describe the plot behaviour analysis on CPI monthly data, check the stationarity of data and carry out structural breaks tests for identification and data selection. The CPI data is obtained from DOSM official portal that taken from the website of Malaysian Employers Federation (MEF). CPI is used to estimate the average price changes of representative basket of goods and services purchased by consumer in a specific time period. It can be categorised as food, clothes, medical care, transportation and education which serves as the economic indicator of inflation in measuring and aggregating the price level and reflect the economics in Malaysia. In this study, the CPI consists of time series monthly data from January 2005 to October 2020 is mainly focus on Malaysia, Peninsular Malaysia, Sabah and WP Labuan as well as Sarawak.

4.2.1 Plot behaviour analysis

Based on the data obtained, the time series plot of the CPI monthly data for Malaysia, Peninsular Malaysia and East Malaysia from January 2005 to October 2020 which are equivalent to 190 months as shown in Figure 4.1. From the graph plotting in Figure 4.1, the red indicate Malaysia, blue indicate Peninsular Malaysia, green represent of Sabah and WP Labuan while grey represent of Sarawak.

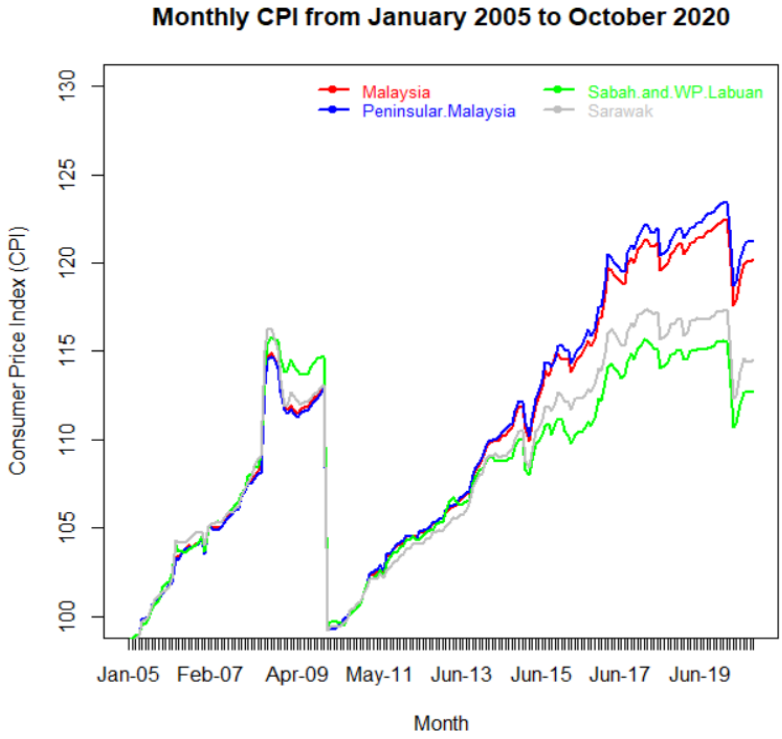


Figure 4.1 The time series plot of the CPI monthly data

Figure 4.1 displays the trend plots of the CPI monthly data for Malaysia, Peninsular Malaysia, Sabah and WP Labuan and Sarawak that exhibit the increasing pattern from January 2005 to December 2009 followed by a sudden drop in year 2010 and then rise up again from January 2011 to October 2020 as the time goes on. The rising in the monthly CPI data for Peninsular Malaysia cause an increase in the overall CPI data of Malaysia as the overall behaviour of the CPI data of East Malaysia always remain lower compare to Peninsular Malaysia.

4.2.2 Stationarity

In order to check the stationarity of the time series data, the test statistic is carried out by performing Augmented Dickey-Fuller (ADF) test. The null hypothesis for this test is the model has a unit root and non-stationary data. The more negative the DF statistic, the higher the chance in rejecting the null hypothesis at the confidence level. In general, the p-value of less than 0.05% indicate reject null hypothesis that the data exists a unit root and is non-

stationary. From Table 4.1, it was displaying the results conducted by using ADF test of the CPI monthly data for Malaysia, Peninsular Malaysia, Sabah and WP Labuan and Sarawak. Based on the table, it shows that all the p-value are greater than $p = 0.05$ implies that the null hypothesis is accepted. Then, the autocorrelation can be diagnosed from the correlogram as shown in Figure 4.2. The ACF plots shows all the original series die down slowly indicate that the data is stationary. Hence, we can conclude that the data have unit roots and are non-stationary.

Data	**Dicker-Fuller (DF)**	**p-value**
Malaysia	-2.0262	0.5652
Peninsular Malaysia	-1.9478	0.5980
Sabah and WP Labuan	-2.4850	0.3732
Sarawak	-2.2362	0.4773

Table 4.1 Augmented Dicker-Fuller (ADF) Test

4.2.3 Structural breaks for data selection

Since this series has non-stationary behaviour therefore, we failed to implement Box-Jenkin ARIMA model for further analysis. This is because ARIMA methodology only capable to use the time series with stationary assumption in predicting the future data. When there is non-stationary data, it required the transformation of data from non-stationary to stationary by method of differencing in order to perform analysis and forecasting. As a result, non-stationary pattern that likely to contain unexpected changes over time may lead to inaccurate forecast.

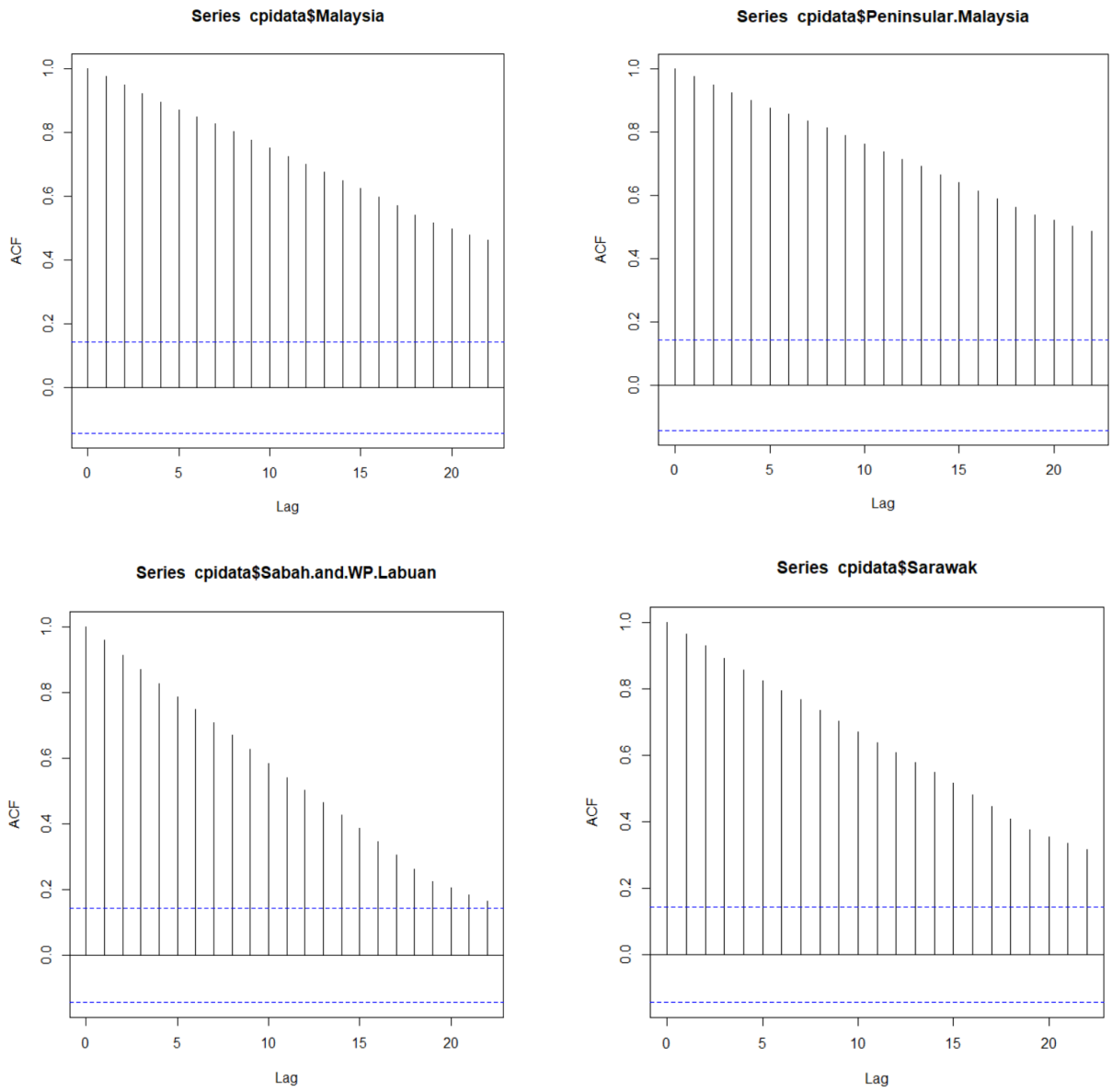


Figure 4.2 The ACF plot for Malaysia, Peninsular Malaysia, Sabah and WP Labuan and Sarawak

In order to reduce forecast error, we have chosen a specific period to analyse CPI data based on the post-structural break period identified via Chow test of the CPI monthly data from January 2005 to October 2020. The null hypothesis of Chow test is that there has no breakpoint in the model. In Chow test, the p-value which less than 0.0005 or if the calculated F statistic is greater compared to the F-critical value will give the evidence of rejecting null hypothesis that the model does not has any breakpoint. Table 4.2 reveals the results that obtained by Chow test. It shows that CPI for Malaysia, Peninsular Malaysia and East Malaysia have same

breakpoints that corresponding to the break dates at August 2007, December 2009, April 2012, August 2014 and December 2016. Since all the p-value of 2.20E-16 are the same which is lower than 0.0005 implies that the null hypothesis is rejected. The resulting breaks at different period as depicted in Figure 4.3 and the vertical dotted lines indicate breakpoints that corresponding to the break dates of the data. Then, the selected post-period structural break on the monthly data period of Malaysia from January 2017 to October 2020 that equivalent to 46 data observations was used to perform SSA approach in CPI forecasting as displayed in Figure 4.4.

Data	F-statistic	Breakpoints	Break dates	p-value
Malaysia	448.4400	t=32, 60, 88, 116, 144	Aug-07, Dec-09, Apr-12, Aug-14, Dec-16	2.20E-16
Peninsular Malaysia	493.1300	t=32, 60, 88, 116, 144	Aug-07, Dec-09, Apr-12, Aug-14, Dec-16	2.20E-16
Sabah and WP Labuan	144.1300	t=32, 60, 88, 116, 144	Aug-07, Dec-09, Apr-12, Aug-14, Dec-16	2.20E-16
Sarawak	250.5000	t=32, 60, 88, 116, 144	Aug-07, Dec-09, Apr-12, Aug-14, Dec-16	2.20E-16

Table 4.2 Chow Test

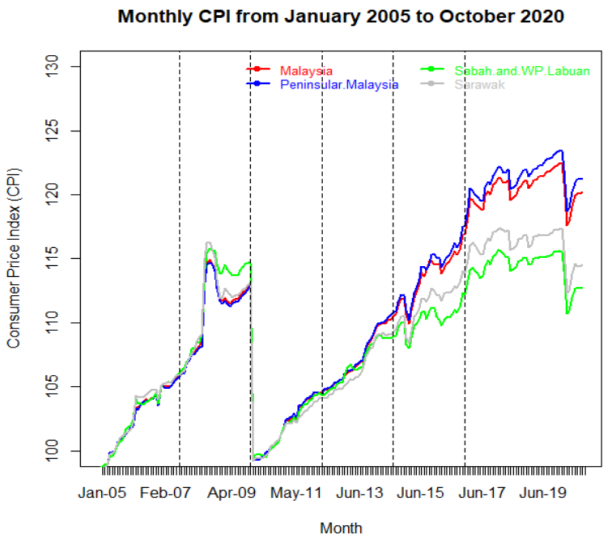


Figure 4.3 The time series plot with breakpoints

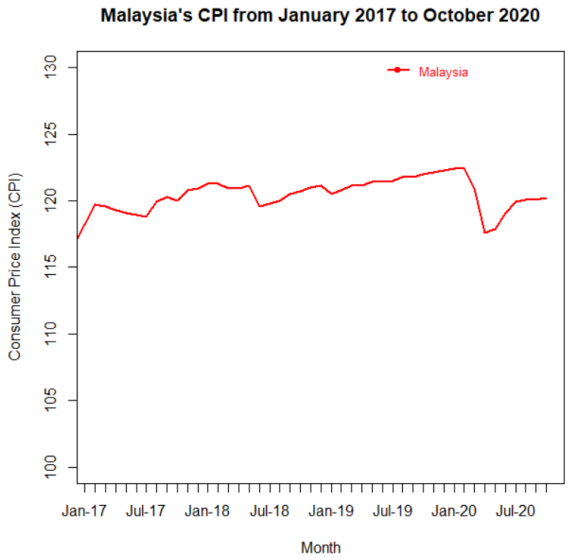


Figure 4.4 The post-period structural break for the CPI monthly data

4.3 Summary

This chapter applied the monthly of CPI data from January 2005 to October 2020. The behavior of CPI time series data exhibit trend plot for Malaysia, Peninsular Malaysia and East Malaysia and the data is not stationary. In order to improve accurate forecast, we compute Chow test for structural change of the model. According to the analysis, we select the post-period structural break on Malaysia's CPI data from January 2017 to October 2020 which used to carry out SSA in CPI forecasting.

CHAPTER 5

EXPECTED RESULTS

5.1 Introduction

This chapter will discuss about Singular Spectrum Analysis (SSA) approach to post-period structural break selection on Malaysia CPI data.

5.2 Expected Findings

In SSA approach, there are two important complementary procedures namely decomposition stage and reconstruction stage. In decomposition stage, the step of embedding is very important in selecting the appropriate number of window length which should be divisible by the period (Golyandina & Korobeynikov, 2014). Large window length is required to ensure that the signals components is clearly separated which cause the method of single value decomposition to be more effective. Since there are 46 data observations for Malaysia CPI data that contain 23 window length, therefore we select the possible window length of 12 for trend extraction as shown in Figure 4.5. In Figure 4.5, it shows that there are 12 eigenvectors and 0 elementary reconstruction series as well as the eigenvalues and eigenvectors of diagnostic plot with 12 window length as depicted in Figure 4.6.

```
Call:
ssa(x = cpidata$Malaysia, L = 12)

Series length: 46,      Window length: 12,      SVD method: eigen
Special triples:  0

Computed:
Eigenvalues: 12,        Eigenvectors: 12,       Factor vectors: 0

Precached: 0 elementary series (0 MiB)

Overall memory consumption (estimate): 0.003246 MiB
```

Figure 4.5 Decomposition stage of Singular Spectrum Analysis

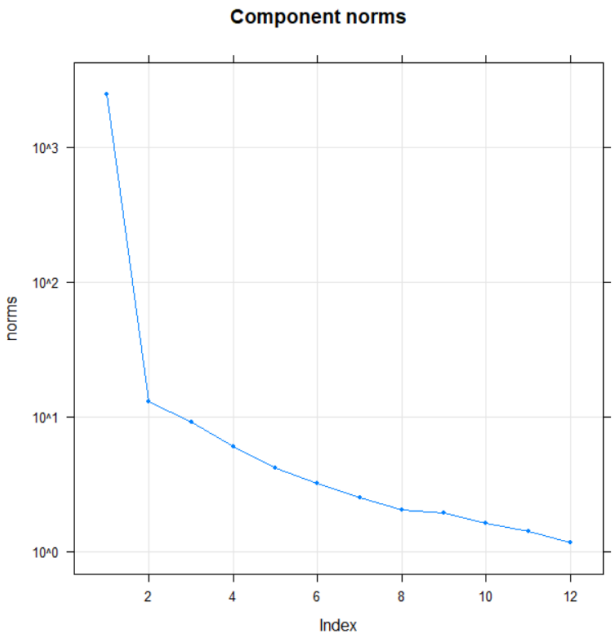


Figure 4.6 Eigenvalues and eigenvectors of diagnostic plot of original data

Then, the one dimensional plot of the eigenvectors and the reconstruction series are displayed in Figure 4.7 and Figure 4.8. Based on Figure 4.7, the selection of the eigenvector in the first graph is necessary to perform trend smoothing since it has the highest contribution of the leading eigentriple which is 99.99% with lowest frequency while the other graphs show only 0% of the leading eigentriple with high frequency which are not suitable for trend extraction. Moreover, we can notice that the trend as depicted in Figure 4.7 is coincides with the reconstructed components as shown in Figure 4.8. It is clearly observed that the combination trend from second, third and fourth graphs of one dimensional eigenvectors in Figure 4.8 is coincides with the second graph of reconstructed series as displayed in Figure 4.7. After that, the graph of comparison on original data and the SSA with extract trend is displayed in Figure 4.9 and the periodogram of the residual trend series is depicted in Figure 4.10. From Figure 4.9, the black line indicates the original data while the red dotted line indicates the SSA with estimated trend components. Thus, this indicate that SSA is capable to decompose the true signals from noise series into trend components for smoothing based on the spectrum of eigenvectors in the singular valued decomposition of the covariance matrix. Next, the diagnostic plot of the number of eigenvalues and eigenvectors with extract trend of 12 window length as revealed in Figure 4.11.

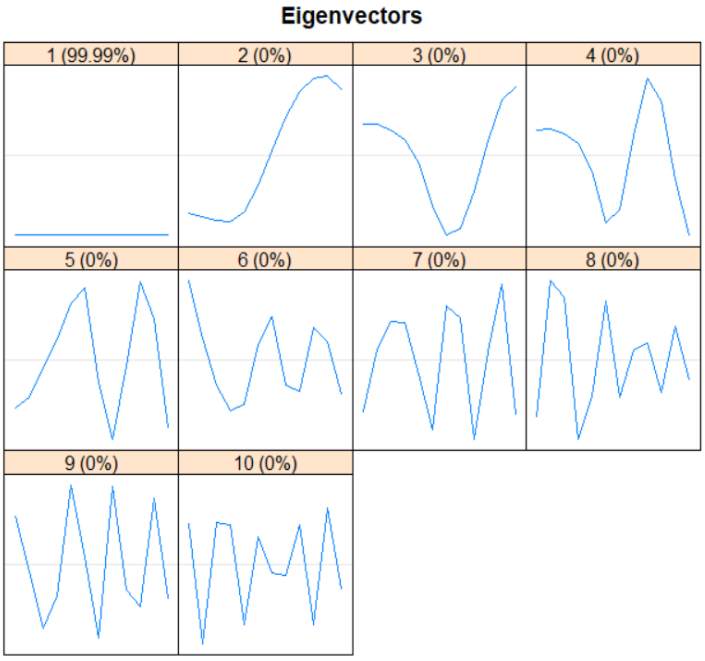


Figure 4.7 Visual Information for eigenvectors

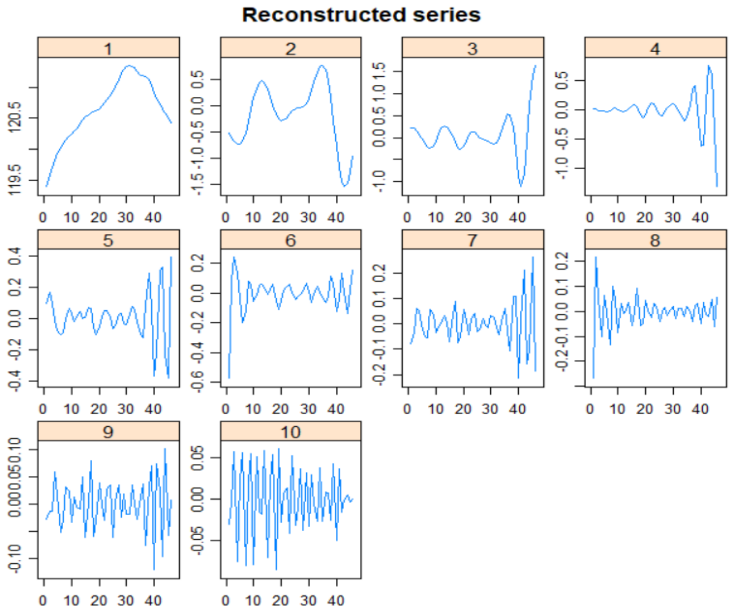


Figure 4.8 Visual Information for reconstructed series

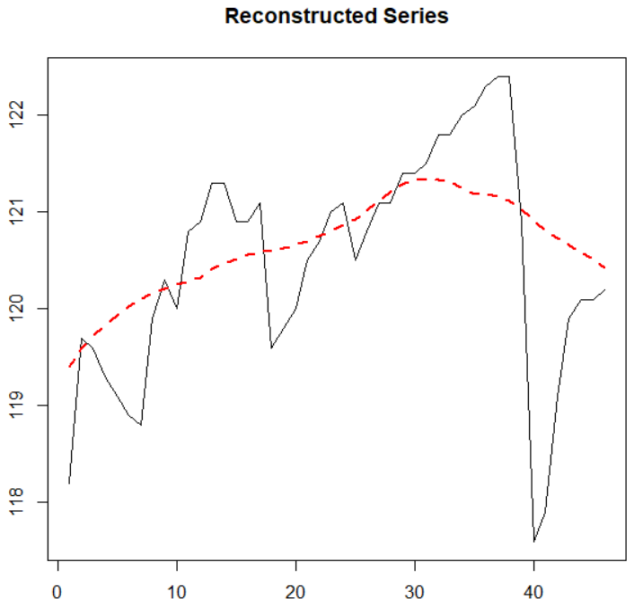


Figure 4.9 Graph of comparison on original data and SSA with trend extraction series

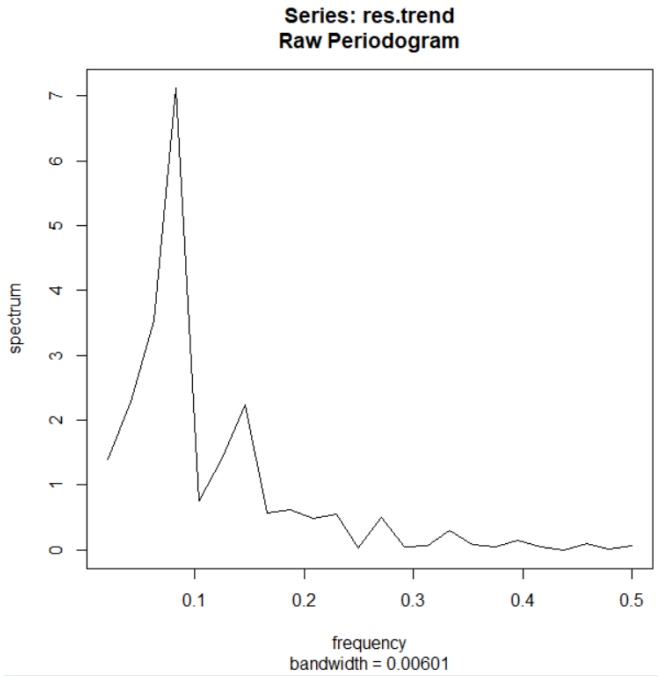


Figure 4.10 Periodogram of residual trend series

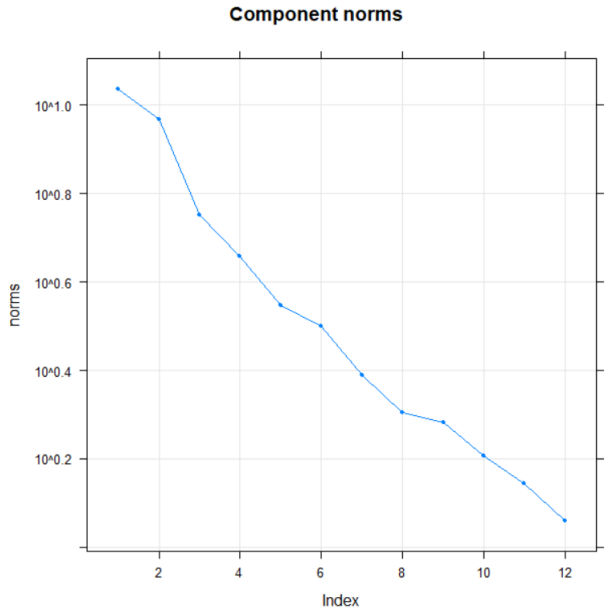


Figure 4.11 Eigenvalues and eigenvectors of diagnostic plot of extracted trend

Subsequently, we carried out the extraction of seasonality from the residual. In order to detect and identify the fluctuation over the selected CPI data period, the graph of scatter plot for the pairs of eigenvectors and w-correlation matrix of the elementary components are used. Each pair of the eigenvectors that shown in Figure 4.12 is corresponding to the sine wave of the seasonal data behaviour. According to Figure 4.12, we can observe that the regular shape of the pairs of eigenvectors can be described by the periodogram of residual trend series as depicted in Figure 4.10. The w-correlation matrix in Figure 4.13 also shows that the pairs of eigenvectors are highly correlated within each other. As a result, we can indicate that the CPI data of Malaysia does not contain any white noise components. The graph of comparison on original data and the SSA with extract seasonal is depicted in Figure 4.14 and the reconstructed series of the combination of the original series as well as SSA with extract trend and seasonality must be plotted as revealed in Figure 4.15. The reconstructed series of the smoothing trend and seasonality components with SSA was then used to perform CPI forecasting.

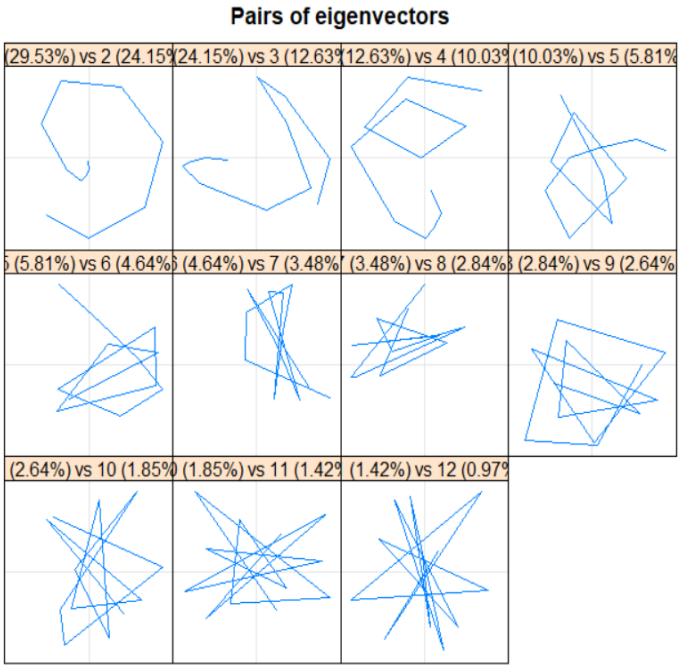


Figure 4.12 Visual information on the pair of eigenvectors of elementary components

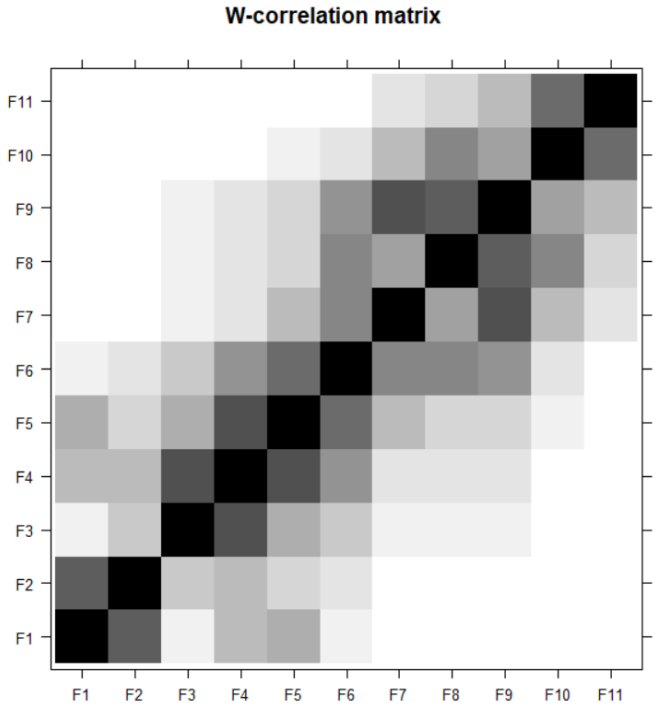


Figure 4.13 W-correlation matrix

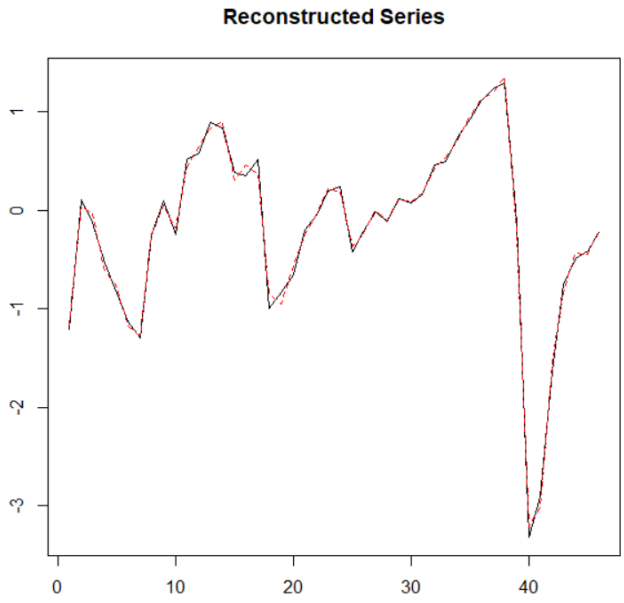


Figure 4.14 Graph of comparison on original data and SSA with seasonality extraction series

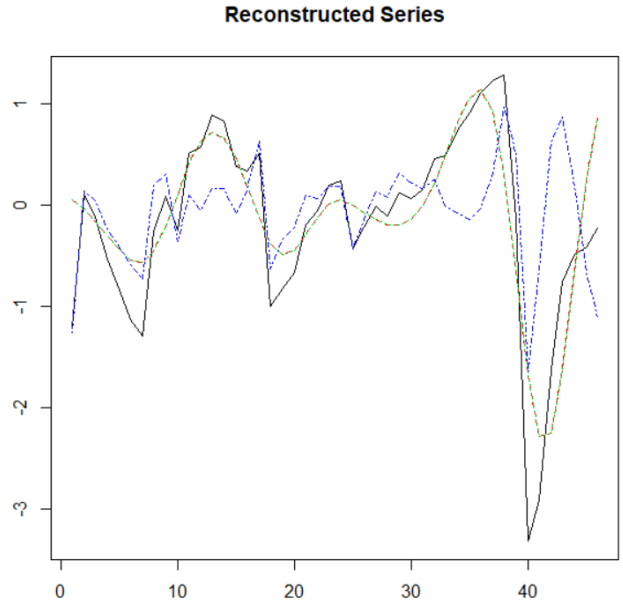


Figure 4.15 Graph of comparison on original data and SSA with trend and seasonality extraction series

5.3 Summary

This chapter discussed on the application of SSA to CPI data of Malaysia from January 2017 to October 2020. SSA is able to perform two stages which include decomposition and reconstruction stages as well as filtering process for trend and seasonality extraction to form a reconstructed series. As a result, we can firm to say that the non-stationarity of data is valid to use in CPI forecasting by univariate non-parametric SSA without required any transformation of data such as ARIMA model.

REFERENCES

Trading Economics. (n.d.). Malaysia Inflation Rate. Retrieved from https://tradingeconomics.com/malaysia/inflation-cpi

Hanaysha, J. R. (2018). An examination of the factors affecting consumer's purchase decision in the Malaysian retail market. *Journal of Management Studies,* 2(1), doi: 10.1108/PRR-08-2017-0034

Silva, A. E. D. S. (2016). Theoretical Advancements and Applications in Singular Spectrum Analysis. [Doctoral dissertation, University of Bournemouth]. ProQuest Dissertations and Thesis database.

Wang, H., & Zhao, W. (2009). ARIMA Model Estimated by Particle Swarm Optimization Algorithm for Consumer Price Index Forecasting. *Artificial Intelligence and Computational Intelligence,* 5855, 48-58. https://doi-org.ezproxy.utm.my/10.1007/978-3-642-05253-8_6

Nyoni, T. (2019). Time series modelling and forecasting of consumer price index in Belgium. MPRA Paper 92414. Retrieved from https://mpra.ub.uni-muenchen.de/92414/

Boniface, A., & Martin, A. (2019). Time series modelling and forecasting of consumer price index in Ghana. *Journal of Advances in Mathematics and Computer Science,* 32(1), 1-11. doi: 10.9734/JAMCS/2019/v32i130134

Ahmar, A.S., GS, A.D., Listyorini, T., Sugiato, C.A., Yuniningsih, Y., Rahim, R., & Kurniasih, N. (2018). Implementation of the ARIMA(p,d,q) method to forecasting CPI Data using forecast package in R Software. *Journal of Physics*, doi: 10.1088/1742-6596/1028/1/012189

Mia, M. S., Nabeen, A. H. M. M. R., & Akter, M. M. (2019). Modelling and Forecasting the Consumer Price Index in Bangladesh through Econometric Models. *Journal for Engineering, Technology and Sciences*, 59(1). Retrieved from: https://asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/5041

Akpanta, A.C., & Okorie, I. E. (2015). On the Time Series Analysis of Consumer Price Index data of Nigeria -1996 to 2013. *American Journal of Economics,* 5(3), 363-369. doi: 10.5923/j.economics.20150503.08

Murdipi, R., & Law, S. H. (2016). Dynamic Linkages between Price Indices and Inflation in Malaysia. *Journal Economics Malaysia,* 50(1), 41-52. http://dx.doi.org./10.17576/JEM-2016-5001-04

Harchaoui, T. M., & Janssen, R. V. (2018). How can big data enhance the timeliness of official statistics? The case of the U.S. consumer price index. *International Journal of Forecasting*, 34, 225-234. https://doi.org/10.1016/j.ijforecast.2017.12.002

Qiu, Y. (2020). Forecasting the Consumer Confidence Index with tree-based MIDAS regressions. *Economic Modelling*, 91, 247-256. https://doi.org/10.1016/j.econmod.2020.06.003

Hassani, H., & Silva, E. S. (2018). Forecasting UK consumer price inflation using inflation forecasts. *Research in Economics*, 72, 367-378. https://doi.org/10.1016/j.rie.2018.07.001

Hassani, H., Soofi, A. S., & Zhigljavsky, A. (2013). Predicting inflation dynamics with singular spectrum analysis. *Journal of the Royal Statistical Society*, 176(3), 743-760. Retrieved from https://ideas.repec.org/a/bla/jorssa/v176y2013i3p743-760.html

Caporale, G. M., & Skare, M. (2018). A non-linear analysis of gibson's paradox. *Engineering Economics*, 29(4), doi: 10.5755/j01.ee.29.4.20403

Ge, Li., & Yin, G. (2012). Application of Process Neural Network on Consumer Price Index Prediction. *Advances in Intelligent and Soft Computing*, 137, 427-432. Retrieved from https://link-springer-com.ezproxy.utm.my/search?query=%22Application+of+Process+Neural+Network+on+Consumer+Price+Index+Prediction+%22

Venkadasalam, S. (2015). The Determinant of Consumer Price Index in Malaysia. *Journal of Economics, Business, and Management*, 3(12), doi: 10.7763/JOEBM.2015.V3.344

Enke, D., & Mehdiyev, N. (2014). A Hybrid Neuro-Fuzzy Model to Forecast Inflation. *Procedia Computer Science*, 36, 254-260. doi: 10.1016/j.procs.2014.09.088

Qiong, L. F., & Zhong, H. S. (2013). The Application of the Genetic Anneal Simulation Support Vector Machine on the Predicting of the Consumer Price Index. *International Conference on Communication, Electronics and Automation Engineering*, 181, 545-550. Retrieved from https://link-springer-com.ezproxy.utm.my/chapter/10.1007/978-3-642-31698-2_77

Alsamara, M., Mrabet, Z., & Hatemi-J, A. (2020). Pass-through of import cost into consumer prices and inflation in GCC countries: Evidence from a nonlinear autoregressive

distributed lags model. *International Review of Economics and Finance*, 70, 89-101. https://doi.org/10.1016/j.iref.2020.07.009

Sun, M. D., & Li, X. Y. (2017). Window length selection of singular spectrum analysis and application to precipitation time series. *Global NEST Journal*, 19(2), 306-317. Retrieved from https://journal.gnest.org/sites/default/files/Submissions/gnest_02117 /gnest_02117_published.pdf

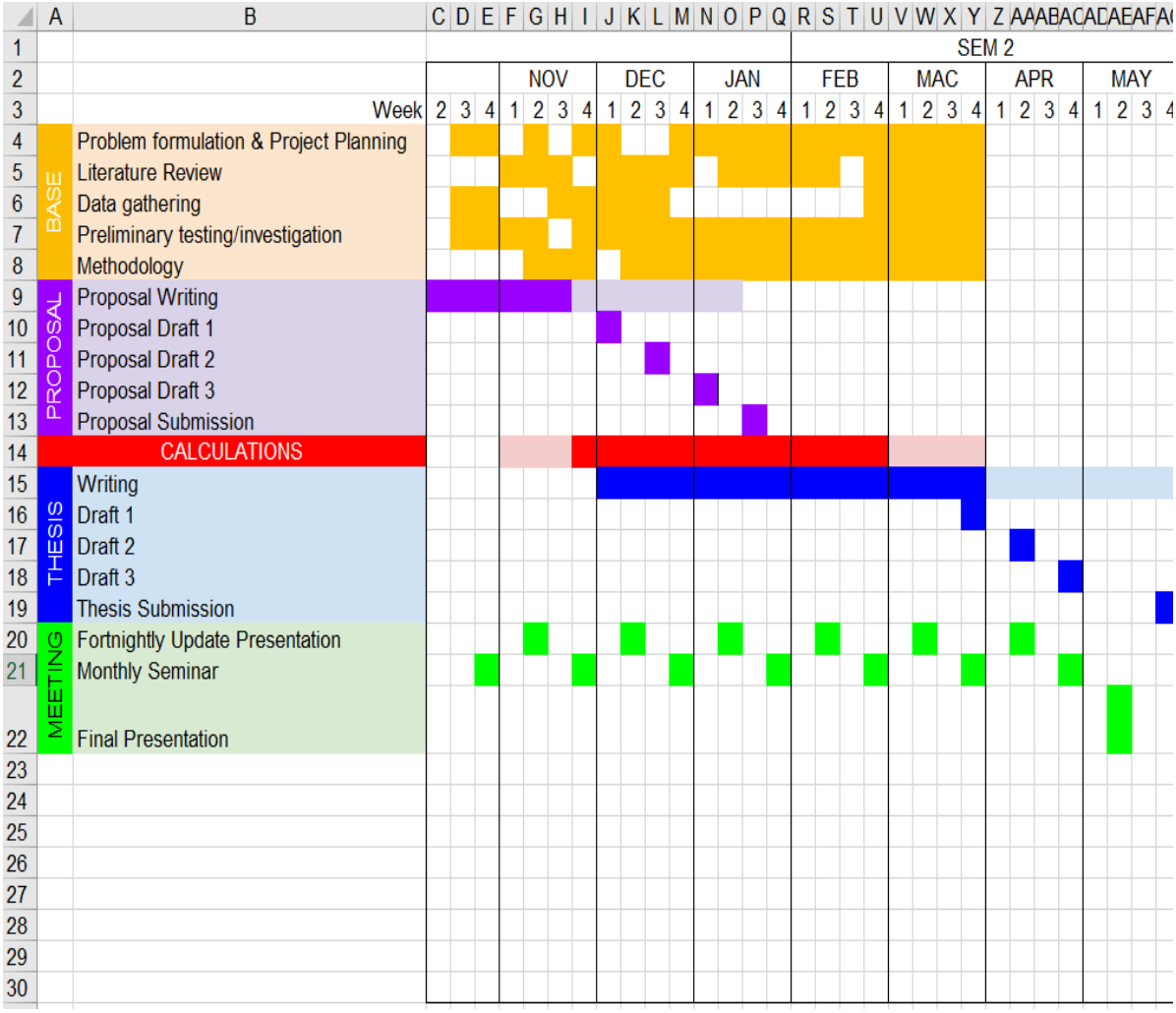
Stephanie. (2016, June 7). *ADF-Augmented Dickey Fuller Test.* Retrieved from https://www.statisticshowto.com/adf-augmented-dickey-fuller-test/

Stephanie. (2016, October 11). *Chow Test: Definition & Examples.* Retrieved from https://www.statisticshowto.com/chow-test/

Osmanzade, A. (2017). Singular Spectrum Analysis forecasting for financial time series. [Master Thesis, University of Tartu]. http://hdl.handle.net/10062/57101

Golyandina, N. E., & Korobeynikov, A. (2014). Basic Singular Spectrum Analysis and Forecasting with R. *Computational Statistics & Data Analysis*, 71, 934-954, doi: 10.1016/j.csda.2013.04.009

Appendix Gantt chart



Appendix R code

```
library(tseries)
#Change the excel workbook to CSV type file before import to R
cpidata <- read.csv(file.choose(), header=T)
cpidata

#unit root test by using Augmented Dicker-Fuller(ADF) Test
#Taking 0.05 as critical value of t-distribution
#Ho: The series has unit root indicate that the series is non-stationary
#H1: The series has no unit root indicate that the series is stationary
#test the stationarity for CPI of Malaysia
adf.test(cpidata$Malaysia)
#test the stationarity for CPI of Peninsular Malaysia
adf.test(cpidata$Peninsular.Malaysia)
#test the stationarity for CPI of Sabah and WP Labuan
adf.test(cpidata$Sabah.and.WP.Labuan)
#test the stationarity for CPI of Sarawak
adf.test(cpidata$Sarawak)

library(stats)
#ACF test to determine whether the data is stationary or not
#test the stationarity for CPI of Malaysia
acf(cpidata$Malaysia)
#test the stationarity for CPI of Peninsular Malaysia
acf(cpidata$Peninsular.Malaysia)
#test the stationarity for CPI of Sabah and WP Labuan
acf(cpidata$Sabah.and.WP.Labuan)
#test the stationarity for CPI of Sarawak
acf(cpidata$Sarawak)

 #Time series plot of CPI data
```

```
#Shows 190 datasets and 4 variables: Malaysia, Peninsular Malaysia, Sabah & WP. Labuan, Sarawak
str(cpidata)

#Malaysia's CPI data
M=cpidata$Malaysia
M
#Peninsular Malaysia's CPI data
PM=cpidata$Peninsular.Malaysia
PM
#Sabah & WP.Labuan's CPI data
SWPL=cpidata$Sabah.and.WP.Labuan
SWPL
#Sarawak's CPI data
S=cpidata$Sarawak
S
Month=cpidata$Month
Month

plot(M,type="l", lwd=2,
xaxt="n", ylim=c(100,130), col="red",
xlab="Month", ylab="Consumer Price Index (CPI)",
main="Monthly CPI from January 2005 to October 2020")
axis(1,at=1:length(Month), labels=Month)
lines(PM, col="blue",type="l", lwd=2)
lines(SWPL, col="green",type="l", lwd=2)
lines(S, col="grey",type="l", lwd=2)

#Add a legend to the plot
legend("topright", legend=c("Malaysia", "Peninsular.Malaysia",
"Sabah.and.WP.Labuan","Sarawak"),
lty=1, lwd=2, pch=21, col=c("red", "blue", "green", "grey"),
ncol=2, bty="n", cex=0.8,
```

```
text.col=c("red", "blue", "green", "grey"),
inset=0.01)

plot(M,type="l", lwd=2,
xaxt="n", ylim=c(100,130), col="red",
xlab="Month", ylab="Consumer Price Index (CPI)",
main="Monthly CPI from January 2005 to October 2020")
axis(1,at=1:length(Month), labels=Month)
lines(PM, col="blue",type="l", lwd=2)
lines(SWPL, col="green",type="l", lwd=2)
lines(S, col="grey",type="l", lwd=2)

#Add a legend to the plot
legend("topright", legend=c("Malaysia", "Peninsular.Malaysia",
"Sabah.and.WP.Labuan","Sarawak"),
lty=1, lwd=2, pch=21, col=c("red", "blue", "green", "grey"),
ncol=2, bty="n", cex=0.8,
text.col=c("red", "blue", "green", "grey"),
inset=0.01)

library(strucchange)
library(tseries)
library(stats)
#Time series plot of CPI data
plot.ts(cpidata)
#Time series plot of CPI data for Malaysia
M <- ts(cpidata$Malaysia, start=c(0:190))
plot(M)
#Test the model null hypothesis that the model has no breakpoint
#Structural break on Chow test
#Find the breakpoints that corresponding to the breakdates
#Compute F statistics
fsM <- Fstats(M ~ 1)
```

```
plot(fsM)
#Perform structural change test based on the CPI data of Malaysia
#Calculate supremum (least upper bound) F statistic and p value
sctest(fsM)
#Shows the optimal 2-segment partition
breakpoints(fsM)
lines(breakpoints(fsM))
#Shows the optimal (m+1)-segment partition
#Shows all the breakpoint and corresponding breakdates
#Calculate RSS and BIC for identification number of breaks
bpM <- breakpoints(M ~ 1)
summary(bpM)
plot(bpM)
#Shows the optimal 6-segment partition
#Shows the breakpoint and corresponding breakdates when m= 5
strucchange::breakpoints(cpidata$Malaysia~1)
fm0 <- lm(M ~ 1)
fm1 <- lm(M ~ breakfactor(bpM, breaks=1))
plot(M)
lines(ts(fitted(fm0), start = 0), col = 3)
lines(ts(fitted(fm1), start = 0), col = 4)
lines(bpM)
#Compute confidence interval for parameter of Malaysia's CPI
ciM <- confint(bpM)
ciM
lines(ciM)

#Time series plot of CPI data for Peninsular Malaysia
PM <- ts(cpidata$Peninsular.Malaysia, start=c(0:190))
plot(PM)
#Test the model null hypothesis that the model has no breakpoint
#Structural break on Chow test
#Find the breakpoints that corresponding to the breakdates
```

```
#Compute F statistics
fsPM <- Fstats(PM ~ 1)
plot(fsPM)
#Perform structural change test based on the CPI data of Peninsular Malaysia
#Calculate supremum (least upper bound) F statistic and p value
sctest(fsPM)
#Shows the optimal 2-segment partition
breakpoints(fsPM)
lines(breakpoints(fsPM))
#Shows the optimal (m+1)-segment partition
#Shows all the breakpoint and corresponding breakdates
#Calculate RSS and BIC for identification number of breaks
bpPM <- breakpoints(PM ~ 1)
summary(bpPM)
plot(bpPM)
#Shows the optimal 6-segment partition
#Shows the breakpoint and corresponding breakdates when m= 5
strucchange::breakpoints(cpidata$Peninsular.Malaysia~1)
fm0 <- lm(PM ~ 1)
fm1 <- lm(PM ~ breakfactor(bpPM, breaks=1))
plot(PM)
lines(ts(fitted(fm0), start = 0), col = 3)
lines(ts(fitted(fm1), start = 0), col = 4)
lines(bpPM)
#Compute confidence interval for parameter of Peninsular Malaysia's CPI
ciPM <- confint(bpPM)
ciPM
lines(ciPM)

#Time series plot of CPI data for Sabah and WP Labuan
SWPL <- ts(cpidata$Sabah.and.WP.Labuan, start=c(0:190))
plot(SWPL)
#Test the model null hypothesis that the model has no breakpoint
```

```
#Structural break on Chow test
#Find the breakpoints that corresponding to the breakdates
#Compute F statistics
fsSWPL <- Fstats(SWPL ~ 1)
plot(fsSWPL)
#Perform structural change test based on the CPI data of Sabah & WP Labuan
#Calculate supremum (least upper bound) F statistic and p value
sctest(fsSWPL)
#Shows the optimal 2-segment partition
breakpoints(fsSWPL)
lines(breakpoints(fsSWPL))
#Shows the optimal (m+1)-segment partition
#Shows all the breakpoint and corresponding breakdates
#Calculate RSS and BIC for identification number of breaks
bpSWPL <- breakpoints(SWPL ~ 1)
summary(bpSWPL)
plot(bpSWPL)
#Shows the optimal 6-segment partition
#Shows the breakpoint and corresponding breakdates when m= 5
strucchange::breakpoints(cpidata$Sabah.and.WP.Labuan~1)
fm0 <- lm(SWPL ~ 1)
fm1 <- lm(SWPL ~ breakfactor(bpSWPL, breaks=1))
plot(SWPL)
lines(ts(fitted(fm0), start = 0), col = 3)
lines(ts(fitted(fm1), start = 0), col = 4)
lines(bpSWPL)
#Compute confidence interval for parameter of Sabah and WP Labuan's CPI
ciSWPL <- confint(bpSWPL)
ciSWPL
lines(ciSWPL)

#Time series plot of CPI data for Sarawak
S <- ts(cpidata$Sarawak, start=c(0:190))
```

```
plot(S)
#Test the model null hypothesis that the model has no breakpoint
#Structural break on Chow test
#Find the breakpoints that corresponding to the breakdates
#Compute F statistics
fsS <- Fstats(S ~ 1)
plot(fsS)
#Perform structural change test based on the CPI data of Sarawak
#Calculate supremum (least upper bound) F statistic and p value
sctest(fsS)
#Shows the optimal 2-segment partition
breakpoints(fsS)
lines(breakpoints(fsS))
#Shows the optimal (m+1)-segment partition
#Shows all the breakpoint and corresponding breakdates
#Calculate RSS and BIC for identification number of breaks
bpS <- breakpoints(S ~ 1)
summary(bpS)
plot(bpS)
#Shows the optimal 6-segment partition
#Shows the breakpoint and corresponding breakdates when m= 5
strucchange::breakpoints(cpidata$Sarawak~1)
fm0 <- lm(S ~ 1)
fm1 <- lm(S ~ breakfactor(bpS, breaks=1))
plot(S)
lines(ts(fitted(fm0), start = 0), col = 3)
lines(ts(fitted(fm1), start = 0), col = 4)
lines(bpS)
#Compute confidence interval for parameter of Sarawak's CPI
ciS <- confint(bpS)
ciS
lines(ciS)
```

```
Month=cpidata$Month
Month
plot(M,type="l", lwd=2,
xaxt="n", ylim=c(100,130), col="red",
xlab="Month", ylab="Consumer Price Index (CPI)",
main="Monthly CPI from January 2005 to October 2020")
axis(1,at=1:length(Month), labels=Month)
lines(PM, col="blue",type="l", lwd=2)
lines(SWPL, col="green",type="l", lwd=2)
lines(S, col="grey",type="l", lwd=2)

#Add a legend to the plot
legend("topright", legend=c("Malaysia", "Peninsular.Malaysia",
"Sabah.and.WP.Labuan","Sarawak"),
lty=1, lwd=2, pch=21, col=c("red", "blue", "green", "grey"),
ncol=2, bty="n", cex=0.8,
text.col=c("red", "blue", "green", "grey"),
inset=0.01)

plot(M,type="l", lwd=2,
xaxt="n", ylim=c(100,130), col="red",
xlab="Month", ylab="Consumer Price Index (CPI)",
main="Monthly CPI from January 2005 to October 2020")
axis(1,at=1:length(Month), labels=Month)
lines(PM, col="blue",type="l", lwd=2)
lines(SWPL, col="green",type="l", lwd=2)
lines(S, col="grey",type="l", lwd=2)

#Add a legend to the plot
legend("topright", legend=c("Malaysia", "Peninsular.Malaysia",
"Sabah.and.WP.Labuan","Sarawak"),
lty=1, lwd=2, pch=21, col=c("red", "blue", "green", "grey"),
ncol=2, bty="n", cex=0.8,
```

```
 text.col=c("red", "blue", "green", "grey"),
 inset=0.01)

 lines(bpM)
 lines(bpPM)
 lines(bpSWPL)
 lines(bpS)

 plot(M,type="l", lwd=2,
 xlim=c(146,190), xaxt="n", ylim=c(100,130), col="red",
 xlab="Month", ylab="Consumer Price Index (CPI)",
 main="Malaysia's CPI from January 2017 to October
 2020")
 axis(1,at=1:length(Month), labels=Month)

 #Add a legend to the plot
 legend("topright", legend=c("Malaysia"),
 lty=1, lwd=2, pch=21, col=c("red", "blue", "green",
 "grey"),
 ncol=2, bty="n", cex=0.8,
 text.col=c("red", "blue", "green", "grey"),
 inset=0.01)


#Singular Spectrum Analysis

library(Rssa)

library(svd)

library(tseries)

library(stats)

#Construction of the time series for decomposition stage

#Perform decomposition stage using maximum window length
```

```
#Trend extraction

#Show various information about decomposition

M<- ssa(cpidata$Malaysia)

summary(M)

M<-ssa(cpidata$Malaysia, L=12)

summary(M)

#Shows eigenvalues diagnostic plot

#Shows the number of calculated eigenvalues and eigenvectors with 12 window length

plot(M)


# Eigenvectors shows 1D graphs to detect trend components
#'idx' argument denotes the indices of vectors
plot(M, type = "vectors", idx=1:10)
# Plot of elementary reconstructed series
# Here the ’groups’ argument specifies the grouping
plot(M, type = "series", groups = as.list(1:10))

#Reconstruction stage
#Extract trend
res1 <- reconstruct(M, groups = list(1))
trend <- res1$F1
plot(res1, add.residuals = FALSE, plot.type = "single",
col = c("black", "red"), lwd = c(1, 2))
#Extract seasonality from the residual obtained
res.trend <- residuals(res1)
spec.pgram(res.trend, detrend = FALSE, log = "no")

#Stage of decomposition and visual information
res.trend <- residuals(res1)
M <- ssa(res.trend, L=12)
```

```
plot(M)
# Scatter plots for the pairs of eigenvectors shows 2D graphs to detect sine waves
#Shows the amplitude and periodic behaviour in complex form
plot(M, type = "paired", idx = 1:11)
# Calculate the w-correlation matrix using the first 30 components
#'groups' argument denotes as grouping
# w-correlation matrix plot of elementary reconstruction components to determine separability points
w <- wcor(M, groups = as.list(1:11))
plot(w)

#Reconstruction stage and plotting of the results
res2 <- reconstruct(M, groups=list(1:10))
seasonality <- res2$F1
res <- residuals(res2)
# Extracted seasonality
plot(res2, add.residuals = FALSE)
#Reconstruction plot, cumulative view
# Decomposition into trend and seasonality
recon <- reconstruct (M, groups = list(trend=c(1,2), seasonality=c(1:2)))
plot(recon)
```