Proceedings of Science and Mathematics
Faculty of Science,
Universiti Teknologi Malaysia

Vol. 3, 2021, page 181-191

Forecasting Consumer Price Index using Singular Spectrum Analysis

Venessa Tay Sin Yi [a], Norhaiza Ahmad [b]*
[a,b]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia.
*Corresponding author: norhaiza@utm.my

Abstract
Consumer price index (CPI) measures the price changes of consumer demand within a time period. CPI time series data tend to change over time which may lead to have the noise components with non-stationary data. In this study, the univariate non-parametric forecasting time series model namely Singular Spectrum Analysis (SSA) is used to predict Malaysian CPI based on different training datasets and window length L. Our finding shows that window length $L = 3$ indicate the lowest forecasting errors for 50%, 70% and 75% of training datasets.
Keywords: Consumer price index; singular spectrum analysis; forecasting

1 Introduction

Consumer Price Index (CPI) is a measurement to estimate the average price changes of representative basket of goods and services purchased by consumer in a specific time period. The changes of CPI during the specific period of time might be affected by the customers on their purchasing behavior [1], income security and standard of living. Due to the elevated volatility in CPI that cause of the factors such as cost-push and the demand of buying and services, therefore it is necessary to employ a forecast technique in predicting the CPI in order to give a potential insight to the economic sector.

The method that have been widely developed in the CPI forecasting field known as ARIMA. Box-Jenkins ARIMA is a classical time series analysis that used to predict future data that mostly considered in the forecast research. Seasonal ARIMA (SARIMA) model is also used to analyze the time series data containing seasonal component to develop a forecasting model. However, these approaches are not model-free and have the disadvantage due to its restriction to normality, linearity and stationarity assumption that discussed by Silva [2]. Using this model, data need to be transformed from non-stationary to stationary by method of differencing for further analysis which may lead to loss of information. In order to reduce the forecast error, we need to select a suitable period to determine a true signal by filtering the noise series.

Alternatively, a univariate non-parametric technique namely Singular Spectrum Analysis (SSA) can be used to overcome the limitations imposed by ARIMA model. It consists of two complementary stages known as decomposition and reconstruction. SSA is capable to perform filtering steps and separate the original signals into smoothing trend, seasonality and noise series components based on the spectrum of eigenvectors in the singular valued decomposition of the trajectory matrix to form the reconstructed one dimensional series for forecasting that explained by Osmanzade [3] which this method cannot applied to traditional analysis.

2 Literature Review

2.1 Forecasting Methods for CPI Data

ARIMA is a traditional class of linear model in forecasting time series data. Nyoni [4] have applied Bon-Jenkin ARIMA to forecast annual CPI data of Belgium for the upcoming ten years in controlling the rate of inflation from year 1960 to 2017. The time series data depicts an upward trend for the forecast period and conclude that Belgium should participate more in economics policies which may lead to high inflation rate in order to reflect the demand on CPI forecasting. Similar method also conducted by Borniface and Martin [5] that focus on the monthly data from March 2013 to November 2018 to forecast the CPI for Ghana. Based on the analysis, SARIMA (2, 1, 1) (1, 0, 0) consider as the most fitted time series model. The time series behavior shows increasing trend from December 2018 to August 2019 in forecasting which indicate that there is no intervention of economics on 95% of confidence limits. Another research on ARIMA methodology was introduced by Ahmar et al. [6] to forecast Indonesia's CPI time series data based on the 132 data observations that from January 2005 to December 2015. The measuring performance of prediction accuracy is evaluated by using root mean square error (RMSE) and mean absolute percentage error (MAPE). The findings revealed ARIMA (1, 0, 0) model is the most appropriate to fit the CPI data with the RMSE and MAPE value of 5.695 and 1.625 respectively. Akpanta and Okorie [7] have also conducted a study in applying SARIMA model to analyze CPI average monthly data of Nigeria from January 2014 to December 2015. The ACF and PACF graph show initially the CPI time series data exists non-stationary and further transform non-stationary data to stationary by method of differencing before carry out forecasting. The t- test statistics results reveal SARIMA (1, 2, 1) (0, 0, 1) is the best fit to CPI data because there is no significance different between the observed data and predictive values at 5% of significance level.

Since there are some limitations imposed by ARIMA model, SSA forecasting model can be implemented. Based on the research from Elsner and Tsonis [8] had discussed on SSA is a non-linear approach for signal extraction in an observed time series data. SSA time series approach has been developed by Broomhead and King in 1986, Vautard and Ghil in 1989 followed by Vautard, Yiou and Ghil in 1992. Ruch and Bester [9] illustrate the measure of annual CPI inflation by applying SSA. SSA is capable to generate the spectral forecast which underlying the dynamic pattern of the series. The steps that involved known as embedding, singular value decomposition, grouping and diagonal averaging. This study shows the forecasting performance on core inflation is sensitivity to the change of window length and the use of sample size.

In addition, a research on forecasting UK consumer price inflation have been conducted by Hassani and Silva [10] in applying the extension of SSA namely multivariate SSA (MSSA) based on the historical monthly data from January 2018 to May 2018. Univariate model such as ARIMA, Exponential Smoothing (ETS), Neural Network (NN) and Trigonometric Box-Cox ARMA Trend-Seasonal (TBATS) was used to make comparison in generating the best perform forecast within the multivariate SSA with Auxiliary Information MSSA(AI) framework. By applying Vertical MSSA Recurrent (VMSSA-R) and Vertical MSSA Vector (VMSSA-V), the results show 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.

Apart from that, Caporale and Skare [11] had proposed a non-linear analysis of Gibson's paradox in Nevertheland which include 73 macroeconomics variables and used the datasets from the year period of 1800 to 2012. Univariate SSA and MSSA was performed to analyze the co-movement between the long term and short term of interest rates and CPI. 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. The MSSA shows that it improved the accuracy of one-step ahead forecast compared to the forecast by univariate SSA.

Based on the study introduced by Golyandina et al. [12], select a suitable window length is an initial step in order to determine the number of leading components for spectral estimation. SSA is crucial in filtering the nonlinear time series for smoothing. The window length should be in positive integer L which fall in the range of $1 \leq L \leq (N+1)/2$. This paper demonstrated the application of SSA for adaptive filter of time series data on economy analysis based on the optimal choice of the parameters.

Khan and Poskitt [13] had discussed on the selection of the window length size in SSA that reflected to the relatively short data period. The window length is simply assumed to the range of $2 \leq L \leq N/2$ for standard practice such that $L = (logN)^c, c < \infty$. In order to avoid any missing or changing on forecasting data problems, large window length is important to ensure that the signal-noise is clearly separated in yielding the accuracy performance.

In this study, Malaysian CPI monthly data was taken to carry out SSA based on the different training datasets and window length.

3 Methodology

3.1 Stationarity

According to Stephanie [14], Augmented Dickey-Fuller (ADF) is a test statistic to check the unit root for stationarity in time series analysis. The DF formula can be computed as:

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \tag{1}$$

where $\hat{\gamma}$ is the differencing parameter and $SE(\hat{\gamma})$ is the standard error estimates. Reject null hypothesis when $\gamma = 0$ or p-value is less than 0.05 that the data has unit root and non-stationary.

Autocorrelation function (ACF) also determined to shows the summary of correlation at different period of time that diagnosed using a correlogram. The test statistic is calculated as:

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

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

3.2 Structural Break

Chow test was implemented by Manuje [15] to identify whether there exists structural break of the data. The F statistical test are denoted as:

$$F = \frac{(RSS_p - (RSS_1 + RSS_2))/k}{(RSS_1 + RSS_2)/(N_1 + N_2 - 2k)} \tag{3}$$

where RSS_p is the pooled regression line, RSS_1 and RSS_2 are the regression line before break and after break. Reject null hypothesis when the p-value is less than 1% significance level.

3.3 Singular Spectrum Analysis

Rodrigues et al. [16] describe that SSA is an adaptive methodology which consists of two complementary stages namely decomposition and reconstruction.

3.3.1 Decomposition Stage

Embedding: Choose the appropriate window length L for constructing the covariance matrix is important which gives the distinct structure observed behavior of a dataset in order to ensure the signals and noise components are clearly separated. L must be in integer with the range of $2 \ll L \ll N$. The trajectory matrix, X can be defined as:

$$X = [X_1, X_2, \dots, X_k] = \left(x_{ij}\right)_{i,j=1}^{L,K} \tag{4}$$

The estimated matrix denoted as $\hat{X} = \left[\hat{X}_1, \dots, \hat{X}_k\right] = \sum_{i=1}^{L} U_i\, U_i^T X$ and construct the Hankel matrix $\tilde{X} = H\hat{X} = [\tilde{X}_1, \dots, \tilde{X}_k]$.

Singular Valued Decomposition (SVD): Analyze the accuracy of singular values in Hankel matrix in changing the dimension of matrix on eigenvalues. The SVD of trajectory matrix, X as follows:

$$X_1, X_2, \dots, X_L \tag{5}$$

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 $\left(\sqrt{\lambda_i}, U_i,\ V_i\right)$.

3.3.2 Reconstruction Stage

Eigentriple Grouping: Split the elementary matrices of X_i into a few groups and sum the matrices within each group. 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 correspond 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:

$$X = X_{I1}, \dots, X_{Im} \tag{6}$$

Diagonal Averaging: Transform a matrix into Hankel matrix to select the singular values of parameter r for filtration process. Another expansion of matrix is obtained that can be formed as:

$$X = \widetilde{X_{I1}} + \dots + \widetilde{X_{Im}} \tag{7}$$

where $\widetilde{X_{IJ}}$ is the diagonal version of matrix X_{Ij} and $\widetilde{X_{I1}} = HX$. The resulting series constructed by the elementary grouping will called as elementary reconstructed series.

3.4 Singular Spectrum Analysis Recurrent Forecasting

Golyandina and Korobeynikov [17] states that SSA recurrent forecast at h-step ahead with time series $X_{N+h} = (x_1, \dots, x_{N+h})$ is defined as:

$$x_i = \begin{cases} \tilde{x}_i & for\ i = 1, \dots, N \\ \sum_{j=1}^{L-1} \alpha_j x_{i-j} & for\ i = N+1, \dots, N+h \end{cases} \tag{8}$$

where x_i is the components with noise reduced series and the vector is denoted as follow:

$$A = \frac{1}{1-v^2} \sum_{i=1}^{L} \pi_i U_i \tag{9}$$

where $A = (\alpha_1, \dots, \alpha_{L-1})$ and the last components x_L of any vector $X = (x_1, \dots, x_N)^T$ is a linear combination of the first components $x_L = \alpha_1 x_{L-1} + \cdots + \alpha_{L-1} x_1$.

3.5 Forecasting Performance Measures

Chen et al. [18] illustrate the forecasting performance can be evaluated by using the mean absolute integrated error (MAPE) and root mean square error (RMSE). The error metrics are computed as:

$$MAPE = \frac{1}{n} \sum \left| \frac{y_t - F_t}{y_t} \right| \times 100\% \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n} \sum (y_t - F_t)^2} \tag{11}$$

where y_t is the actual values, F_t is the estimated target values and n is the data observation.

4 Results and Discussion

The CPI data is extracted from the website of MEF which sourced from DOSM official portal. The CPI consists of time series monthly data from January 2005 to October 2020 which taken to analyze the behavior of CPI from Malaysia, Peninsular Malaysia, Sabah and WP Labuan as well as Sarawak. Figure 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 overtime.

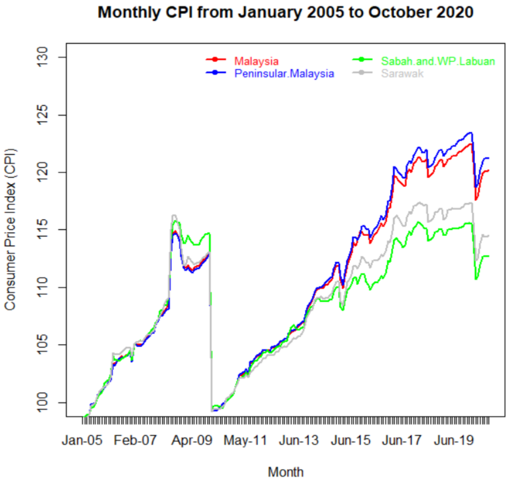

Figure 1: The Time Series Plot of the Consumer Price Index Monthly Data

The rising trend in the monthly CPI data for Peninsular Malaysia may 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. Since, there appears to indicate changes in the overall pattern for all three CPI univariate time series, thus it is important for stationarity checking by computing ADF test. The results conducted by using ADF test shows that all the p-value are greater than $p = 0.05$ implies that the null hypothesis is not rejected. The ACF plots also shows all the original series die down slowly indicate that the data is non-stationary. Hence, this indicate that the CPI data for Malaysia, Peninsular Malaysia, Sabah and WP Labuan and Sarawak have unit roots and are non-stationary.

Since there exists non-stationary behavior therefore, we seek to identify the changes in the period of the CPI time series data by observing any structural break for the CPI which are identified using Chow test. The analysis shows that CPI for Malaysia, Peninsular Malaysia and East Malaysia have the same breakpoints that correspond to the break dates at August 2007, December 2009, April 2012, August 2014 and December 2016. The p-value for all regions are small which is 2.20E-16 imply that the null hypothesis is rejected. This indicate that all regions have similar significant breakpoints that at t=32, 60, 88, 116 and 144.

As there are no significant differences between Malaysia, Peninsular Malaysia, Sabah and WP Labuan and Sarawak, therefore we will use CPI data for Malaysia. Additionally, 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 are used to perform SSA approach in CPI forecasting. The forecasting performance was evaluated by using SSA based on datasets at 50%, 70% and 75% of training data period and different window length $L = 3, L = 4, L = 5, L = 12$. In this study, SSA was applied to 50% training CPI datasets using window length $L = 3$.

4.1 Decomposition and Reconstruction

50% of training data period was taken to apply the decomposition and reconstruction stage of SSA. The initial time series data was used to filter into separating trend and seasonality components for smoothing based on the spectrum of eigenvectors in the singular valued decomposition of the covariance matrix to form the reconstructed series as obtained in Figure 6.

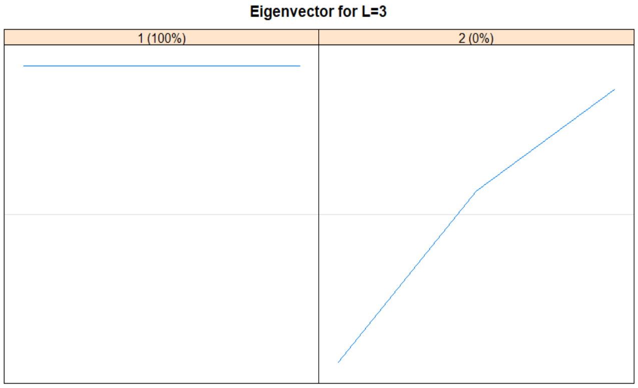

Figure 2: The Plot of Eigenvectors for Window Length $L = 3$ of 50% datasets

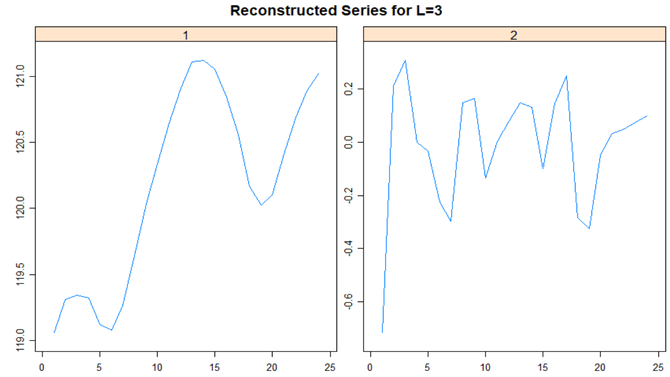

Figure 3: The Plot of Reconstructed Series for Window Length $L = 3$ for 50% datasets

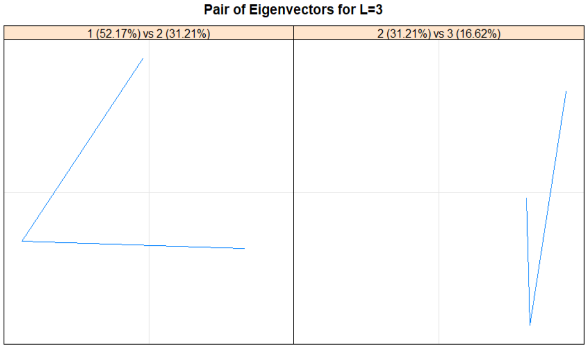

Figure 4: The Pair of Eigenvectors of Elementary Components for Window Length $L = 3$ of 50% datasets

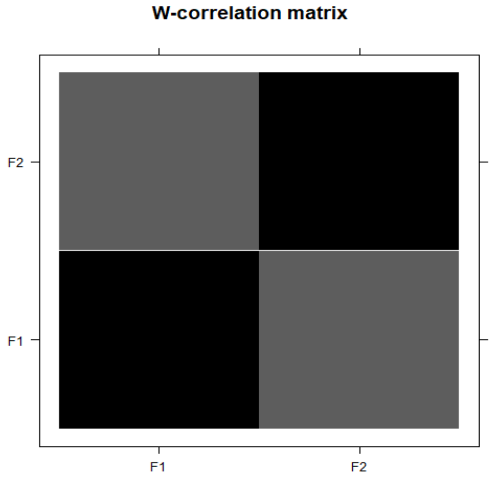

Figure 5: W-correlation Matrix for Window Length $L = 3$ of 50% datasets

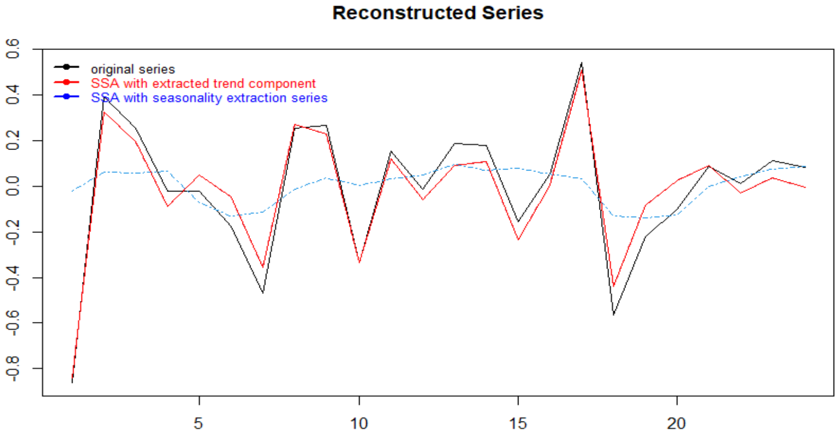

Figure 6: Comparison on Original Data with SSA extracted trend and seasonality components of Window Length $L = 3$ for 50% datasets

4.2 Forecasting Consumer Price Index

Figure 7 illustrate the comparison between the original time series and reconstructed components with forecasting for window length $L = 3$ of 50% datasets. The results of the modelling and forecasting performance of accuracy are evaluated by MAPE and RMSE are shown in Table 1.

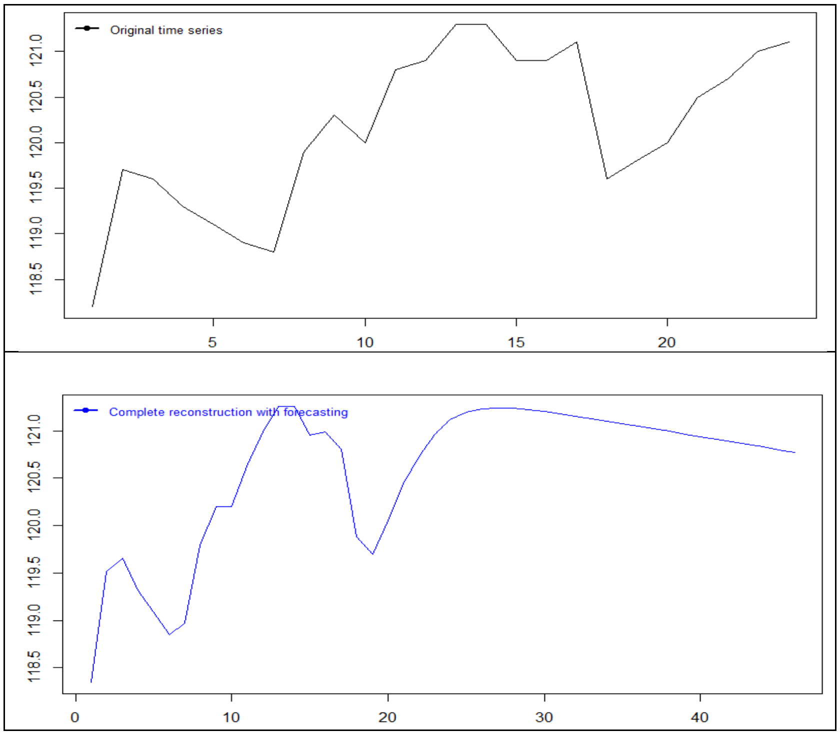

Figure 7: Comparison on Initial Time Series and Reconstructed Components with Forecasting for Window Length $L = 3$ of 50% Training and Testing datasets

Table 1: The Evaluation of Error Measures for Modelling and Forecasting for Window Length $L = 3$ of 50% datasets

50% of datasets	L=3	
	MAPE	RMSE
Modelling	0.1922	0.3084
Forecasting	0.7778	1.2564

4.3 Discussion

According to the analysis, the eigenvector with the highest contribution of the leading eigentriple of 100% with lowest frequency for window length $L = 3$ as shown in Figure 2 was chosen to perform trend smoothing. It is clearly observed that the trend displayed in Figure 2 is coincides with the reconstructed components as shown in Figure 3.

Figure 4 shows each pair of the eigenvectors correspond to the sine wave of the seasonal data behavior in order to detect the fluctuation over the selected CPI data period. The pairs of eigenvectors are highly correlated within each other based on w-correlation matrix as obtained in Figure 5 which indicate that the time series data does not contain any white noise components. By carrying out the decomposition and reconstruction stage of SSA, Figure 6 displayed the form of reconstructed series which taken to generate CPI forecast for the next 14 data points.

Based on Figure 7, the forecasting time series for window length $L = 3$ exhibits decreasing trend with the MAPE value of 0.7778 which is lower compared to RMSE for forecasting as shown in Table 1.

5 Conclusion

Univariate non-parametric SSA is an important powerful tool in analyzing CPI time series data. In this study, we described on the time series analysis on CPI data, the test for stationarity, structural break analysis, post-period structural break selection, the methodology of SSA and the interpretation of results on SSA forecasting performance.

This study analyzed the CPI monthly data in Malaysia from January 2005 to October 2020 that equivalent to 190 months. SSA used initial time series to construct the leading eigenvector in order to obtain the covariance matrix in a singular value decomposition. This technique able to decompose the CPI time series data into trend and seasonality components in order to form the smoothing reconstructed series for forecasting.

Generally, the SSA forecasting performance was evaluated by computing MAPE and RMSE. The comparison was made by using different window length $L = 3, L = 4, L = 5, L = 12$ based on the different proportion at 50%, 70% and 75% monthly datasets of training data period. The results in Table 1 shows the use of $L = 3$ in decomposing the true signal into smoothing reconstructed series has the lowest error of forecasting.

Table 2: The Summary Evaluation of Error Measures for Modelling and Forecasting for Window Length $L = 3, L = 4, L = 5, L = 12$ of 50%, 70, 75% datasets

50% of datasets	L=3		L=4		L=5		L=12	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Modelling	0.1922	0.3084	0.2417	0.3540	0.2882	0.4022	0.4326	0.6284
Forecasting	0.7778	1.2564	0.9371	1.6471	1.8143	2.8320	0.9733	1.4938

70% of datasets	L=3		L=4		L=5		L=12	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Modelling	0.1720	0.2803	0.2063	0.3181	0.2376	0.3550	0.3535	0.5303
Forecasting	1.3402	2.1732	1.4501	2.3243	1.6268	2.6485	1.4881	2.1970
75% of datasets	L=3		L=4		L=5		L=12	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Modelling	0.1591	0.2673	0.1919	0.3037	0.2211	0.3391	0.3302	0.5076

6 References

[1] 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): 7-23.

[2] Silva, A. E. D. S (2016). *Theoretical Advancements and Applications in Singular Spectrum Analysis.* Ph.D. Thesis, Bournemouth University.

[3] Osmanzade, A (2017) *Singular Spectrum Analysis forecasting for financial time series.* Master's Thesis, University of Tartu.

[4] Nyoni, T. (2019) 'Time series modelling and forecasting of consumer price index in Belgium', *MPRA 92414.* University Library of Munich, Germany.

[5] Boniface, A. and Martin, A. (2019) 'Time series modelling and forecasting of consumer price index in Ghana', *Journal of Advances in Mathematics and Computer Science*. 32(2): 1-11.

[6] Ahmar A.S., GS, A.D., Listyorini, T., Sugiato, C.A., Yuniningsih, Y., Rahim, R. and Kurniasih N. (2018) 'Implementation of the ARIMA(p,d,q) method to forecasting CPI Data using forecast package in R Software', *Journal of Physics*. 1028(1): 012189.

[7] Akpanta, A.C. and 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.

[8] Elsner, J. B. and Tsonis, A. A. (1996) '*Singular Spectrum Analysis: A new tool in time series analysis.* Springer Science+Business Media.

[9] Ruch, F. and Bester, D. (2013) 'Towards a measure of core inflation using singular spectrum analysis', *South African Journal of Economics*. 81(3): 307-329.

[10] Hassani, H. and Silva, E. S. (2018) 'Forecasting UK consumer price inflation using inflation forecasts', *Research in Economics*. 72: 367-378.

[11] Caporale, G. M. and Skare, M. (2018) 'A non-linear analysis of gibson's paradox', *Journal of Policy Modelling*. 41(5): 926-942.

[12] Golyandina, N., Pepelyshev, A. and Steland, A. (2012) 'New approaches to nonparametric density estimation and selection of smoothing parameters', *Computational Statistics and Data Analysis.* 56; 2206-2218.

[13] Khan, M. A. R. and Poskitt, D. S. (2011). 'Window Length Selection and Signal-Noise Separation and Reconstruction in Singular Spectrum Analysis', *Econometrics and Business Statistics*. Monash University, Australia.

[14] Stephanie, G. (2016) *ADF-Augmented Dicker Fuller Test.* StatisticsHowTo.com: Elementary Statistics for the rest of us!. https://www.statisticshowto.com/adf-augmented-dickey-fuller-test/

[15] Manuje, J. K. (2018) 'Theory and Practice of Testing for a Single Structural Break in Stata', *EconStor Preprints 172517, ZBW-Leibniz Information Centre for Economics*. https://hdl.handle.net/10419/172517

[16] Rodrigues, P. C., Lourenco, V. and Mahmoudvand, R. (2018) 'A robust approach on singular spectrum analysis', *Quality and Reliability Engineering International.* 34(7): 1437-1447.

[17] Golyandina, N. E. and Korobeynikov, A. (2014). 'Basic Singular Spectrum Analysis and Forecasting with R', *Computational Statistics & Data Analysis*. 71: 934-954.

[18] Chen, C., Twycross, J. and Garibaldi, JM. (2017) 'A new accuracy measure based on bounded relative error for the time series forecasting', *PLOS ONE*. 12(3): e0174202.