

Singular Spectrum Analysis: An Application to Kenya's Industrial Inputs Price Index

Kimutai K. Emmanuel, Adolphus Wagala, Dennis K., Muriithi

¹Faculty of Science, Engineering and Technology, Chuka University, Kenya,
emmanuelkoech858@gmail.com)

²Department of Mathematics and Computer Science, Bomet University, Kenya,
wagala@buc.ac.ke)

³Faculty of Science, Engineering and Technology, Chuka University, Kenya,
kamuriithi2011@gmail.com).

Abstract

Time series modeling and forecasting techniques serve as gauging tools to understand the time-related properties of a given time series and its future course. Most financial and economic time series data do not meet the restrictive assumptions of normality, linearity, and stationarity of the observed data, limiting the application of classical models without data transformation. As non-parametric methods, Singular Spectrum Analysis (SSA) is data-adaptive; hence do not necessarily consider these restrictive assumptions as in classical methods. The current study employed a longitudinal research design to evaluate how SSA fits Kenya's monthly industrial inputs price index from January 1992 to April 2022. Since 2018, reducing the costs of industrial inputs has been one of Kenya's manufacturing agendas to level the playing field and foster Kenya's manufacturing sector. It was expected that Kenya's Manufacturing Value Added hit a tune of 22% by 2022. The study results showed that the SSA ($L = 12$, $r = 7$) (MAPE = 0.707%) provides more reliable forecasts. The 24-period forecasts showed that the industrial inputs price index remains high above the index in 2017 before the post-industrial agenda targeting a reduction in the cost of industrial inputs. Thus, the industrial input prices should be reduced to a sustainable level.

Keywords: Industrial inputs price index, singular spectrum analysis, singular value decomposition.

I. Introduction

Industrialization is the transformation of an economy from primarily agricultural to one based on manufacturing goods. Industrialization leads to sustainable economic growth and development since it creates employment opportunities, enhances international trade, and ensures maximum resource utilization in an economy. Empirical studies have revealed that the manufacturing sector has the largest employment multiplier effect than other sectors [1], [2]. Therefore, developing

economies can create significant employment in the non-tradable sectors by shifting towards skill-intensive production [2]. Industrialized countries like the United States, Japan, France, Singapore, and Germany have the highest manufacturing value added (MVA) proxied by the percentage share to the gross domestic product of about 11%, 21%, 10%, 20%, and 21% respectively in 2018 [3]. Among the newly industrialized countries like China, South Africa (SA), Brazil, and Mexico, China leads with 28% MVA while SA has 12% in 2018 [3].

In Kenya, the service sector is the leading share in the GDP, followed by agriculture, forestry and fishing, industry (construction), and manufacturing. In 2016, the four-sector had MVA of 44.8% [4], 31.1% [5],

and 17.9% [6] and 9.3% [7] respectively. In 2019 the respective MVA were 43.2%, 34.1%, 16.1%, and 7.5%. The MVA in the manufacturing sector records a higher percentage decrease from 2016 to 2019, representing about – 19.36 % compared to construction (-10.056%) and the service sector (-3.57%). While Africa is the cost-effective location for manufacturers due to its cheap source of labor [8], most manufacturing companies in Kenya face challenges of the increased cost of inputs in their production processes, high taxes, and cost of electricity [8]. Kenya's manufacturing sector majorly relies on imported raw materials, approximated to be at least 80 percent. Hence, the cost of imported industrial inputs is a crucial determinant of Kenya's industrial competitiveness. According to the KAM priority agenda [8], MVA will likely increase from 8.4% in 2017 to 15% by 2022. The first pillar under enhancing manufacturing is competitiveness and leveling the playing field, with one of its agendas being to lower the cost of imported industrial inputs [8]. Evaluation of the evolution of a series using time series analysis techniques is a common practice in policy formulation and appraisal.

Analytical models can be broadly classified into: parametric and non-parametric. The choice of the two broad classifications of models depends on the nature of the data. In parametric methods, the data should meet the assumptions of normality or homogeneity of variance. If the data violates the assumption, two directions can be taken depending on the study's statistical test. One can either adopt the non-parametric alternatives or make the data meet the prior restrictive assumptions by applying linear (log) or non-linear (square root) transformation or removing outlier data where appropriate. In time series analysis, the most observable concern is the distribution and stationarity of the data. Classical time series models such as the Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) capitalize on the normality or stationarity data. Yet, most financial and economic time series data do not meet these assumptions. Another notable feature in time series data is the trend component attributed to time-related factors. Besides, level shifts become problematic in the specification and estimation of ARMA class of models [9]. Violation of the stationarity assumptions, using classical time series techniques, calls for data transformations using methods such as differencing. According to Hassani and Mahmoudvand [10], such data transformation results in a loss of time-related information of a given series. Thus, parametric time series models may fail to accurately predict economic and financial time series

characterized by structural breaks, deviation from normality, and non-stationarity caused by exogenous shocks such as technological advances and changes in government policy changes and consumer preferences [11]. Others, especially in the stock market, are the news, announcements, and insider information that cause sudden unexpected stock price swings.

The parametric assumptions relating to normality, stationarity, and linearity do not apply when using non-parametric models such as the singular spectral analysis (SSA). The SSA is an extension of principal components analysis for time series data. SSA applies matrix algebra to disintegrate a trajectory matrix into independent components. The components are grouped and reconstructed based on their periodicities to mimic the trend, oscillatory, and noise components in a given time series [12]. Unlike the classical spectral methods, SSA signal decomposition and reconstruction employ completely data-adaptive basis functions (the eigenmodes of the trajectory matrix), rather than relying on a prior defined basis function like the Fourier approach (based on sine and cosine functions) or mother wavelets [13]-[15]. Besides, SSA has an advantage over other non-parametric approaches since it requires only two parameters (window length (L) and the number of components (r)) to model a given time series.

II. LITERATURE REVIEW

Several empirical studies have demonstrated the efficacy of SSA in modeling and forecasting. Due to its practical advantages over classical models, SSA has been successfully used in various disciplines such as meteorology [16], medicine [17], image processing [18], economics [19], finance [11], and paleoclimatology [20]. Most empirical studies have demonstrated that the SSA technique has a better forecasting ability than classical models. For instance, [21] compared SSA, additive Holt-Winters exponential smoothing (HWES), and Seasonal ARIMA (SARIMA) model using monthly electric load demand (in MW) in a Venezuelan area supplied by wind power. The authors employed a cross-validation approach where the data ($N = 240$) was split into the training set ($n_1 = 228$) used to fit the model parameters, and the test set ($n_2 = 12$) was used to assess the model performance. The study obtained a window length

(L) of 108 using the formula $L = (T-s)/2$ where $N = 240$, and s is the periodicity of the series (i.e., $s = 12$) to ensure that the chosen window length is a multiple of 12 for better separability of the components. As a result, the study identified seven significant eigentriples used in the reconstruction stage. Based on the minimization of in-sample prediction errors, the results showed that SSA (Mean Absolute Percentage Error (MAPE) = 1.06%) was the best fit model to forecast monthly electric load demand than Additive HWTES model (MAPE = 1.25%) and ARIMA (2,1,0)(2,0,2)!" model (MAPE = 1.15%). The out-sample evaluation was also consistent with the in-sample evaluation period, where SSA (MAPE = 1.26%) outperformed than Additive HWES model (MAPE = 2.28%) and ARIMA (2,1,0)(2,0,2)!" model (MAPE = 1.86%) in making 12-step ahead forecasts.

Reference [22] evaluated the sensitivity of the SSA methodology, ARIMA, Exponential Smoothing (ES), and Neural Network (NN) models to the great economic recession. The authors used the U.S. recession dates, data visualization, and the Bai and Perron (2003) test for structural breaking points to divide the imports and exports data into three periods. The imports data was divided into the period before (January 1989 to July 2008), during (August 2008 - April 2009), and after the global economic recession (May 2009 to December 2011). The exports data spanned the period before (January 1989 to June 2008), during (July 2008 to January 2009), and after the recession (February 2009 - December 2011). The authors employed cross-validation before and after the recession, where two-thirds of the data formed the training set. A third was used to evaluate the forecasting accuracy over 1, 3, 6, and 12-months steps ahead horizon. Based on the minimization of the Root Mean Square Error (RMSE), the direction of change (DC) criteria, the modified Diebold-Mariano test, and the t-test, the authors concluded that SSA is less sensitive to U.S. trade recessions than optimized ARIMA, ES, and NN models. The results from the recession showed that SSA forecasts were the most accurate predictions for the U.S. imports and exports than the ARIMA, ES, and

NN. For instance, the average Relative RMSE (RRMSE) for U.S. imports showed that the SSA forecast was on average (across all step ahead horizons) 11%, 9%, and 51% better than from ARIMA, ES, and NN models, respectively, before the 2008 recession. During the recession, the SSA forecasts that the 2008 recession period was 40% better than the ARIMA and 6% better than HW's model. The post-recession analysis period shows the resilience of SSA, as its forecasts were 34% better than those from the ARIMA model, 42% better than forecasts from the ES model, and 27% better than those from the NN. Intuitively, the ES and NN failed to fit an adequate model due to the relatively small sample sizes used in each period. The authors' contribution is that SSA is a good noise filtering technique and recommends using the methodology for more precise forecasts over a long-term horizon.

The research work of [23] compared SSA forecasting ability of SSA, ARIMA, ES, and feed-forward Neural Network model (NN) using monthly U.S. Tourist arrivals from 1996 to 2012. The authors employed cross-validation where two-thirds of the sample data ($n_1 = 355$) formed the training set (used to tune the model parameters), and a third ($n_2 = 356$ observations) formed the validation set (used to test the forecast accuracy over a given horizon). The model used different L and r, which minimized prediction errors due to changing forecasting horizons. The results showed that SSA outperformed the three competing models in all h- (1, 3, 6, 12, 24, and 36) step-ahead forecasts. For instance, SSA ($L = 15$, $r = 6$, $MAPE = 9\%$) outperformed the SARIMA (2,0,1) (1,1,2) ($MAPE = 14\%$), ES ($\alpha = 0.86$, $\sigma = 0.1231$) with a multiplicative seasonality ($MAPE = 17\%$), and NNAR (2,1,1) ($MAPE = 23\%$).

The study of [24] compared the performance of SSA, ARIMA, and HWES methods in fitting and forecasting the daily ambulance demand for the Emergency Medical Services provided by the Welsh Ambulance Service Trust (WAST) in Wales. The data covered the period between April

1, 2005, to December 31, 2009. In a training set of the first 51 months of data ($N = 1,552$), SSA ($L = 581$, $R = 14$) was the best fit for the ambulance demand series. The results showed SSA outperformed HWES method in minimizing the prediction errors in the in-sample forecasts (RMSE SSA = 6.19, RMSEHWES = 6.37) and across the h-step ahead forecasts in terms of minimizing the RMSE (7-day: SSA = 42.20, HWES = 45.46; 14-day: SSA = 42.86, HWES = 47.47; 21-day: SSA = 43.87, HWES = 48.32; 28-day: SSA = 45.46, HWES = 51.14). The study findings imply that SSA is consistent over a longer forecast horizon and recommended for use in forecasting Welsh ambulance demand.

The reviewed empirical studies covering several fields, economies, time frames, and frequency have demonstrated that the SSA methodology has a better forecasting ability than classical models such as ARIMA class of models. Thus, the current study adds to the existing literature on SSA in its application to Kenya's industrial inputs price index.

III. METHODOLOGY

A. Research Design

The current study employed a longitudinal research design. In time series analysis, longitudinal analysis entails evaluating spatial and temporal changes in a given variable by looking at a repeated collection of consistent historical data over time [25]. Of interest in the current study is to describe the time-related properties present in Kenya's industrial inputs price index. Generally, a time series can portray seasonal, trend, cyclical, or error components.

B. Data

No sampling technique was employed since the series under study is univariate time series data. The study used sample data of Kenya's monthly industrial inputs price index data over 26 years, from January 1992 to April 2021, as the available dataset range from the time of data extraction from the IMF website. The period has 352 data points considered appropriate since the SSA applies to even smaller sample sizes below 120 [12]. The index aggregate both agricultural raw materials and base metals price indices.

C. Data Analysis

Data analysis was done using the R Statistical Software. Modeling using SSA was done in a series of stages: embedding the series, decomposing the series using SVD, grouping the components, reconstructing the additive components of the time series, and eventually forecasting the series [12]. The steps are discussed below.

Step 1: Embedding

Let the series be $X\# = [x\$, x!, x'', \dots, X\%]$. Embedding entails taking lagged values of $X\#$ to obtain a trajectory matrix (Y) of dimension $L \times (N - L + 1)$ from the time-series $\{X\#\}$ where N is the series length, and L is the window length. The window length is the crucial parameter, which dictates the separability of different Eigenmodes. The limit for window length is $2 \leq L \leq N/2$

[12]. Upon selecting the optimal L , embedding maps the series X_t into a trajectory matrix Y of dimension $L \times K$, represented in (1).

$$Y_{L \times K} = T(X_t) = [Y_1, Y_2, Y_3, \dots, Y_K]^T \quad (1)$$

Where;

$$Y_i = [x_{i-1}, x_i, x_{i+1}, \dots, x_{i+L}]^T \quad (2)$$

Step 2: Decomposition Stage

The SVD is a crucial process of SSA, which helps identify the periodicities of different Eigen/principal modes ranked by their significance or share of variance explained. The SVD theorem states:

$$A_{t \times y} = U_{t \times t} S_{t \times t} V_{t \times y} \quad (3)$$

Where columns of the matrix $U_{t \times t}$ are the left singular vectors (row-component correlation); S is a diagonal matrix of singular values, the rows of matrix V are the right singular vectors (column-component correlation) [26]. The $U_{t \times t}$ consists of the eigenvectors of AA^T whereas the eigenvectors of $A^T A$ yields the columns of $V_{t \times y}$. The eigenvalues and eigenvectors of AA^T and $A^T A$ are obtained using the eigen decomposition [27]. In the current study, the resultant trajectory matrix Y is decomposed using SVD into three parts as expressed in (4).

$$Y = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T \quad (4)$$

The above notation can be expressed as additive components in (5):

$$Y = Y_1 + Y_2 + \dots + Y_d \quad (5)$$

Where, λ_i represents the singular values of matrix Y arranged in decreasing order of magnitude corresponding to the i th eigenvector, U_i are the left singular vectors, V_i are the right singular vectors. In univariate series, of interest is to get eigentriples (λ_i, U_i, V_i) of the SVD.

Step 3: Grouping

Each eigentriple is a vector, which, before grouping, can be seen as a possible component. The periodicity and contribution of each Eigen component (eigenvector) depend on the size of their respective eigenvalues. Major/principal components carry a larger weight, whereas more unwanted signals (noise) have low variance and come last [12]. Hence, it is possible to identify the Eigen components corresponding to noise. The contribution of the eigenvector can be expressed in terms of the proportion of variance it explains. The share of each eigenvector ϕ_i is obtained using (6).

$$\phi_i = \frac{\sqrt{\lambda_i}}{\sum_{i=1} \sqrt{\lambda_i}} \quad (6)$$

Where $\sqrt{\lambda_i}$ is the square root of the eigenvalue λ_i associated with eigenvector V_i .

Understanding the proportion of variance is vital because meaningful vectors that explain a substantial variation in the original series are used to reconstruct the time series components, whereas insignificant vectors are treated as noise. The study retained all the eigenvectors that yield a cumulative variance explained of 100% to ensure no vital information from the data is lost.

The components are grouped as pairs if they have the same periodicity [12]. Insignificant Eigen triplets that correspond to low eigenvalues as they resemble the unwanted noise are dropped. If, for example, there exist two such groups given by $G_1 = \{ Y_1, Y_2, Y_3 \}$, $G_2 = \{ Y_4, Y_5 \}$ among d triplets, then; $Y_{678} = G_1 + G_2 + G_6$ where $G_6 = \{ Y_9, Y_{10}, \dots, Y_0 \}$ is the residual signal constituting the noise. The grouping $G_1 + G_2 + \dots + G_6$ yields a reconstructed trajectory matrix (T_6), as expressed in (7).

$$Y_{rec} = G_1 + G_2 + \dots + G_r = \sum_{i=1}^r \sqrt{\lambda_i} U_i V_i^T \quad (7)$$

The separability of the principal components can be determined using the Weighted correlation (W_8) computed using (7) [28].

$$W_c = \frac{(Pc^1, Pc^1)_w}{\|Pc^1\|_w \cdot \|Pc^1\|_w} \quad (8)$$

Where $\|Pc^1\|_w = \sqrt{(Pc^1, Pc^1)_w}$:

$$(Pc^1, Pc^1)_w = \sum_{k=1}^N W_k \cdot Pc^1_k Pc^1_k, W_k, \text{ and}$$

The weighted correlation quantifies the degree of separability among two reconstructed components. Let X_1 and X_2 be two reconstructed time series. The two series are said to be separable if w -correlation (X_1, X_2) is zero. Contrarily, two or more components are grouped if the resultant w -correlation (X_1, X_2) is relatively high or significant.

Step 4: Reconstruction Stage

The next step is the reconstruction of the deterministic component of the time series of length N by applying diagonal averaging on the Trajectory matrix T_6 . The SVD process is reversed through a newly reconstructed trajectory matrix to a deterministic univariate series. Diagonal Averaging provides an optimal way to return to a single time-series data.

Step 5: Forecasting

The forecasts are obtained from the reconstructed series. The h -step ahead predictor model in SSA is defined using (8) [20].

$$X_{N+h} = x_1, \dots, x_{N+h} \quad (9)$$

Forecasting should satisfy recurrent linear formulae in (10):

$$x_{i+d} = \sum_{k=1}^d \alpha_k x_{i+d-k}, 1 \leq i \leq N - d \quad (10)$$

Where d is the dimension of the series and α_k are the coefficients. Notably, d is the rank of Y , which is equal to the rank of YY^* .

D. Models Evaluation Criterion

The current study employed the Mean Absolute Percentage Error (MAPE) to compare the performance of SSA and the two HWTES models, namely, the additive and multiplicative. The formula in (11) gives the MAPE.

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{X_t - F_t}{X_t} \right| \times 100 \quad (11)$$

Where; X_t is the actual data, F_t are fitted data using a given forecasting model.

E. Post-Diagnostic Tests.

The post-diagnostic tests entailed the examination of the distribution of the residuals. In essence, the residuals of an adequate model should be white noise (random), normally distributed, and have no autocorrelation. A visual plot to test for its randomness. Here, the model residuals are expected to have no clear pattern and mimic a purely white noise process. The second check is that the models' in-sample residuals should be normally distributed examined in the current study using a histogram. The third test is the assumption of no residual autocorrelation. The assumptions assume no autoregressive dependence among the residuals was tested using the ACF plot.

IV. RESULTS AND DISCUSSION

A. Preliminary Analysis

The choice of a time series model relies on a clear understanding of the evolution and distribution of data. Thus, the current study carried out data visualization, descriptive statistics, and decomposition of the series as preliminary analysis. The time series plot in Fig. 1 presents an evolution of Kenya's monthly industrial inputs price index data from January 1992 to April 2021. The price index in the 1990s was fairly stable until 2002/03, which saw a steady rise to 2007. The impact of the 2008/2009 economic recession attributed to the banking crisis is evident due to a sharp fall within this period. The period between 2011 to 2016 saw a decline attributed to the smooth transition of the government after the 2012 general elections. Ironically, the industrial inputs price index continued to soar even after Kenya's industrial agenda of reducing the cost of industrial inputs in 2017 [8]. The trend can partly explain that Kenya's MVA declined from 9.3% in 2016 to 7.5% in 2019 [7]. The recurrent fluctuation in the series shows that the industrial inputs price index is elastic to global economic shocks (2008/09 financial crisis and the recent Coronavirus Disease 2019 (COVID-19) and political upheavals. In a statistical sense, such

exogenous shocks make the series noisier or induce the seasonality component in a time series, making the series non-stationary. The Augmented Dickey-Fuller (ADF) test ascertained that the data suffer from a unit root problem, that is, the series is non-stationary (ADF = -2.1324, Lag order = 7, p-value = 0.5209).

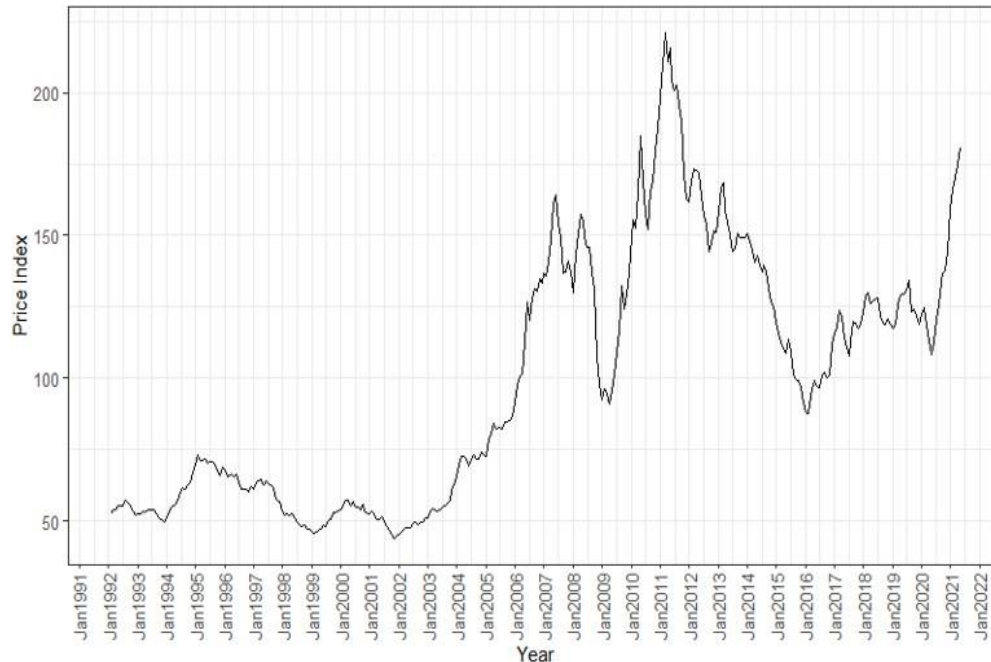


Fig. 1: Monthly industrial inputs price index data (1992M1–2021M4). Note: The base year for the series is 2016

The industrial inputs price index over the study period averaged 99.29 (SD = 44.09) (Table I). The measures of dispersion are close to zero, indicating that the data is nearly symmetrical to the mean (Skewness = 0.490, Kurtosis = - 0.085). The Shapiro-Wilk (S-W) test statistically ascertained that the industrial inputs index data does not follow a normal distribution ($W = 0.9116$, $p = 0.000$). The negative Kurtosis test implies that the distribution of the series is platykurtic; that is, the data is flatter than that of normally distributed data, has a broader peak, and observations tend to spread around the mean.

TABLE I: DESCRIPTIVE STATISTICS

	N	Min	Mean	SD	Max	Skewness	Kurtosis
IIP	352	43.78	99.29	44.09	221.18	0.490	-0.085

Note. IIP: Industrial Inputs Price Index.

The non-stationarity and non-normality aspect of the data substantiates the appropriateness of the selected non-parametric models. In general, the series depicts seasonality and trend patterns. Therefore, smoothing techniques such as SSA can provide better results since they filter or extract the trend and seasonality components from irregular variation [29].

B. The SSA Methodology

The stages of SSA are discussed empirically using the industrial inputs price index series as follows:

Step 1: Embedding

In a one-dimensional series such as the industrial inputs price index series, the univariate series must be transformed into a matrix as a preliminary stage to SVD known as embedding. The step involves taking lagged values as a new series and coalescing them into a trajectory matrix of dimension $L \times K$ (so-called Hankelization process). The most crucial task in this stage is determining the optimal window length (L). The size of L influences the separability of the resultant eigenmodes. Yet, no optimal window length can be generalized to all series since each may portray unique time series components at different periods. Besides, the frequency of the time series varies in each study. Therefore, it is in the researcher's interest to determine an optimal window length.

As suggested by Golyandina et al. [12], $2 \leq L \leq N/2$. Recently, [30] found that the reconstruction of the time series using SSA with Hadamard transform surpasses the standard SSA and proposed to improve the time series analysis, particularly when choosing L in the embedding stage of SSA. Thus, L should be selected from the maximum order Hadamard transform (H_2) satisfying $22 < N$. In the current study, the sample points (352) lie within the interval (256, 512] or (2⁸, 2⁹]. Thus, $h = 8$ and the resultant maximum size of time series embedding is $L^* = 2^{h+1} = 2^9 = 512$. Since window lengths divisible by the period give a better separability [31], a sequence of multiples of 12 between 12 and 180 (NB: $N/2 = 176$) was augmented with 128 as possible window lengths. As a forward-looking approach, the prediction errors of each window length (L), considering the optimal number of significant eigenvalues (r), were obtained (Fig. 2).

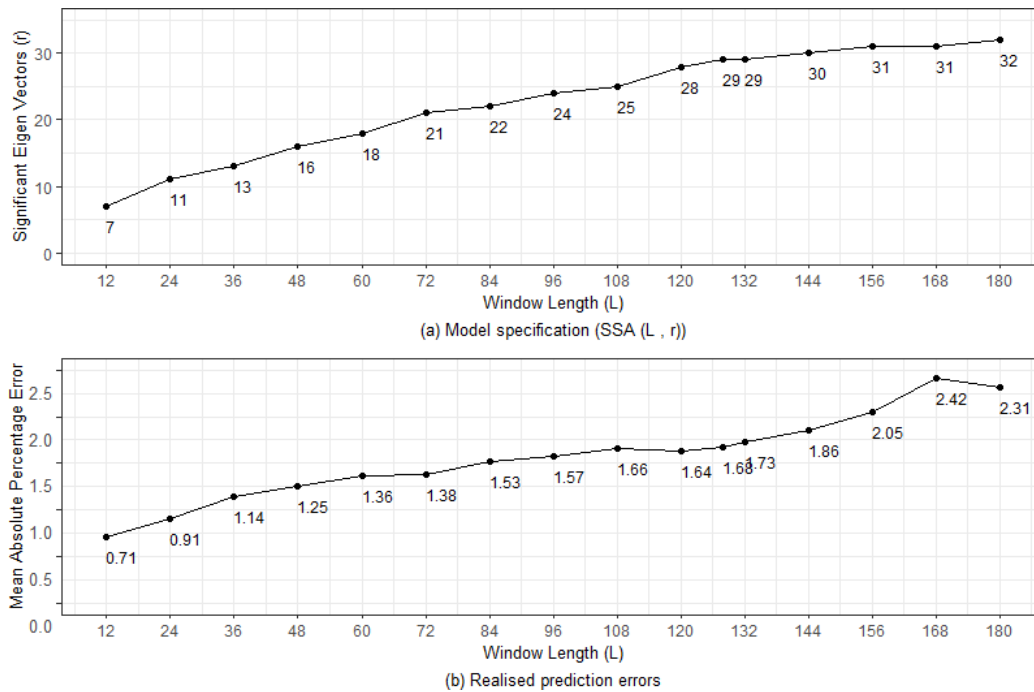


Fig. 2: Window length selection.

Measured by the MAPE, the prediction errors increase as window length increases. An optimal window length for Kenya's industrial input price index was found to be 12 – the frequency of the series – with optimal $r = 7$ (MAPE = 0.71%). Thus, the SSA model for the IPI can be specified as SSA ($L = 12, r = 7$). The results are coherent with empirical studies such as those of [11], who found a gradual increase in RRMSE with an increase in L in reconstructing a simulated sinusoidal noisy series. Using a sample of 355 observations of monthly US Tourist arrivals data as the training set, [23] found that a window length of 15 with associated 6 eigenvectors was the best-fit window length minimization of the prediction errors (MAPE = 9%). Studies that have used a relatively larger window length have also found few significant eigentriples to reconstruct the series, implying that a higher dimensional matrix is unnecessary as it reflects data redundancy – a possible cause of overfitting data. For instance, [21] used $L = 108$ (close to the conventional formula of $L = (N - 12)/2$) in 228 sampled training data of monthly electric load demand (in MW) in a Venezuelan region supplied by wind power found 7 significant eigentriples used in the reconstruction stage. The preferred relatively small L of 12 also reflects the nature of the series. The selection of the small window length is in line with the expectation of [12] that a smaller window length is appropriate in extracting the trend component, which explains a larger variation in Kenya's industrial inputs price index series under study.

Step 2: Singular value decomposition

The main target of SVD is to factorize the trajectory matrix into subcomponents that can carry vital information of the matrix. Embedding using the selected window length results into a trajectory matrix; $Y_{340 \times 12}$ which is then decomposed using SVD as expressed in (12).

$$Y_{340 \times 12} = U_{340 \times 12} \sqrt{\lambda_{12 \times 12}} V_{12 \times 12} \quad (12)$$

The (U_+) and the (V_+) are the and left and right singular vectors (V_+) whereas the singular values (scalars) ($\sqrt{\lambda_+}$) denote the scaling. Of interest is to get the collection $B\lambda_+U_+V_+$ known as the eigentriples, which are donated by F_+ . The SVD decomposition of the embedded matrix results in 12 leading eigentriples (F_+) (Fig. 3) Notably, the eigenvectors are usually arranged in descending order based on the magnitude of their singular values.

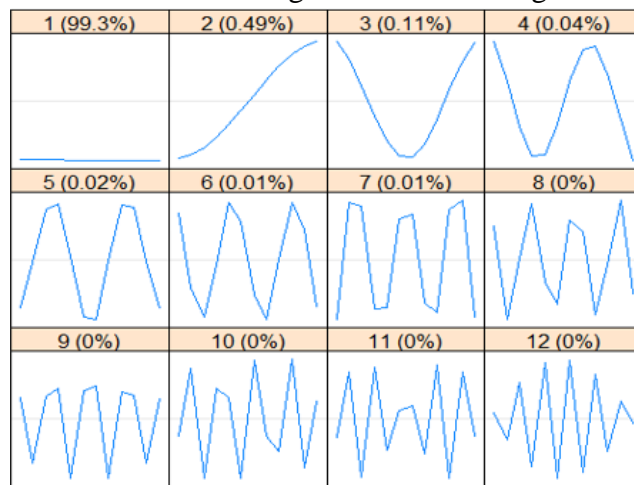


Fig. 3: Leading eigenvectors from SVD decomposition.

The form of an eigenvector replicates the kind of the time series component that produced the eigenvector [31]. The eigentriple (F_+), F_1 , and F_2 are associated with slow varying components by visual impression. That is, F_1 can be seen as the constant level, whereas F_2 is an increasing trend. In contrast, the remaining are associated with high varying or oscillating components (cyclical or seasonality) and possibly noise. Cumulatively, the first seven eigenvectors explain 100% of the industrial inputs price index series variance. Thus, they form significant Eigen triples to reconstruct signals of interest. The remaining five eigenvectors form residual components or noise. The scree plot has also been widely used to determine a suitable number of components in all forms of analysis that involve PCA. The scree plot in Figure 4 reflects the norms against their respective 12 leading eigenvalues in descending order.

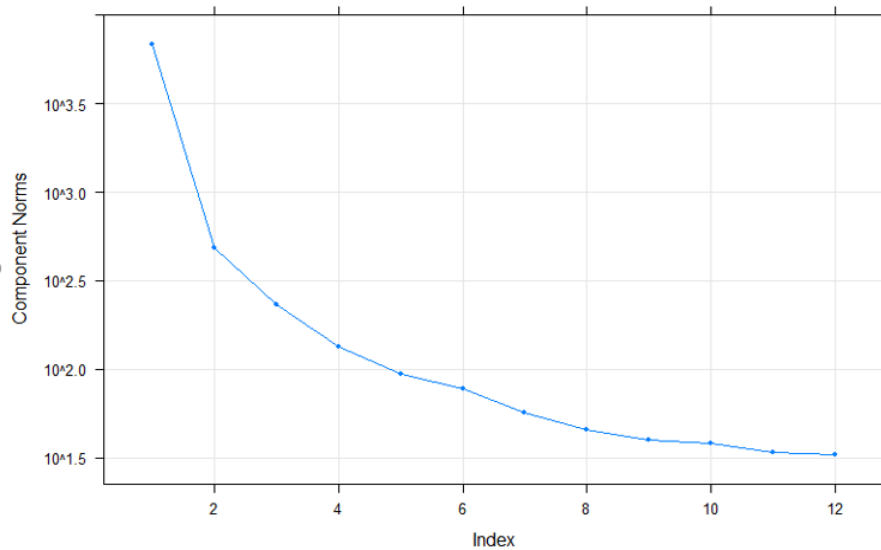


Fig. 4: Scree-plot of component norms against the eigen triples. Note. $Norm_i = \|Y u_i\| = \sqrt{Y u_i^T Y u_i} = \sqrt{\lambda_i} = \text{singular value } i$

The curve levels off after the 7th eigenvector suggesting that the first seven eigentriples can adequately explain the time series components of the industrial inputs price index series. The selection is consistent with the visualization of the variance explained by the resultant eigenvectors in Fig. 5, which added up to 100%. Thus, the reconstruction of the signal of interest was done using eigentriples 1 to 7 while the rest constitute the noise.

Step 3 and 4: Grouping and reconstruction

The last two steps of the SSA methodology helps in the identification of signal of the time series properties of industrial inputs price index series. While the seven eigentriples can be treated as individual components, it is crucial to group them based on their correlation. The periodicity of their respective eigenvectors determines the grouping of the Eigen triplets [12]. The intuition that singular values of the two eigentriples in a harmonic series are close to each other guides the separability and identification of harmonic components. An analysis of the pairwise scatterplots of the singular vectors aid in the identification of eigentriples corresponding to the harmonic components of the series (Fig. 5).

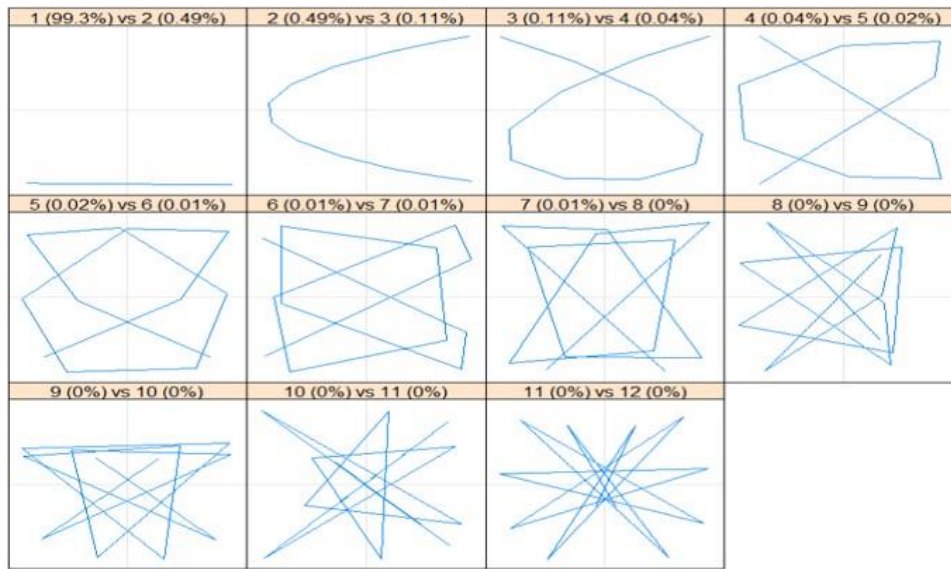


Fig. 5: Pairwise correlation of eigentriples from SVD.

The geometric figures help visualize the degree of correlation between any two close eigenvectors. The first two vectors have nearly no correlation, whereas F2, F3, and F4 are somewhat correlated. The other geometric shapes denoting the pairwise correlation between the subsequent eigenvectors have no apparent pattern. Grouping using a scatter plot is subject to visual bias. As such, the separability of the twelve possible eigentriples can be quantified using a w-correlation (w-cor.) matrix. The matrix in Fig. 6 depicts weighted correlations between 12 eigenvectors as possible components to reconstruct the time series. Each shaded cell (F_i , F_j) symbolizes the correlation between eigenvectors i and j , coded from black (w-cor. = 1) to white (w-cor. = 0). Similar components have a strong correlation (darker shading) and are grouped. Conversely, components with a lighter to the white background are linearly independent hence separable. For instance, the first eigenvector (F1) (now component 1) is separable from the rest and makes up the first reconstructed principal component (RC1). In the same logic, F2 to F3 constitute RC2, F4 to F7 make up RC3. The three components correspond to the time series components, namely level (RC1), trend (RC2), and seasonality (RC3). As initially indicated that the first seven eigenvectors explain 100% of the industrial inputs price index variance. Thus, the remaining eigentriples (F8 to F12) form the residual signal or noise.

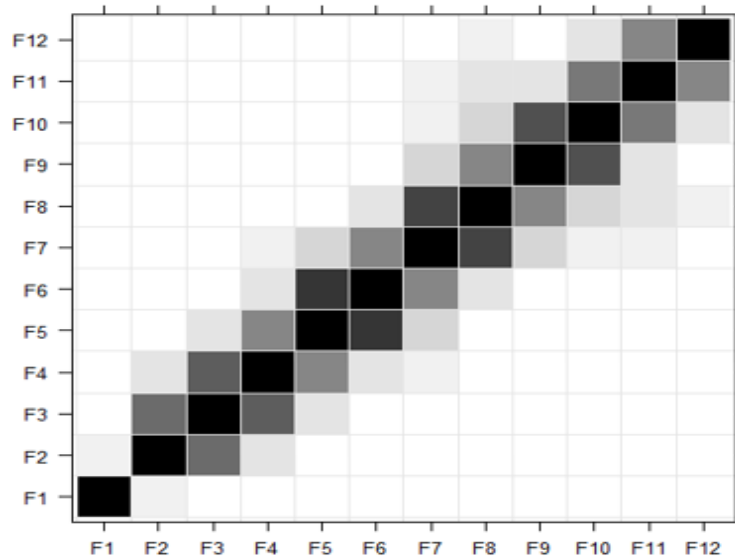


Fig. 6: W-correlation matrix.

The three reconstructed components are visualized in Fig. 7. Since SSA takes the additive form, the original series is obtained by summing RC1, RC2, and RC. The RC1 represents the underlying nominal shifts in the level or periodic averages common in time-series data [9], [32]. The shifts are consistent with the long-term movement in time series, which can eventually form the so-called cyclical component of a series. The level prior to 2003 was somehow constant and low. The trend component oscillates around zero during this period indicating stable prices. The drastic rise in the level between 2003 and 2006 reflects the rising cost of agricultural products partly driven by the rising cost of material and service inputs in the agricultural sector as indicated by [33]. The rise in these commodities, which are inputs in the production

sector, can be associated with the rising cost of material inputs, such as fertilizer, fuel, and seeds in the manufacturing sector.

Regarding the imported raw materials, import duties, Railway development levy, and Import declaration fee (IDF) at a rate of 1.5% and 2%, respectively [34], each play a significant role in determining the cost of imported raw materials such as rubber, Copper, Aluminium, and Iron Ore. The charges are always relatively constant hence do not have a substantial impact that shifts the level of a series. Compiled together, these drivers have been associated with the upward shifts and increasing trend of the cost of industrial inputs index. The RC2 conforms to the trend, which frequently changes direction and magnitude over time. The constantly changing trend component after 2005 shows that the industrial inputs price index is unstable and susceptible to changes in sudden shocks such as the 2007 post-election violence that show a sharp fall in the cost of inputs which coincided with the 2007/08 financial crisis both of which lowers producers' incentive due

to uncertainty in the consumption levels. The reduced demand for raw materials pushes the cost of inputs down.

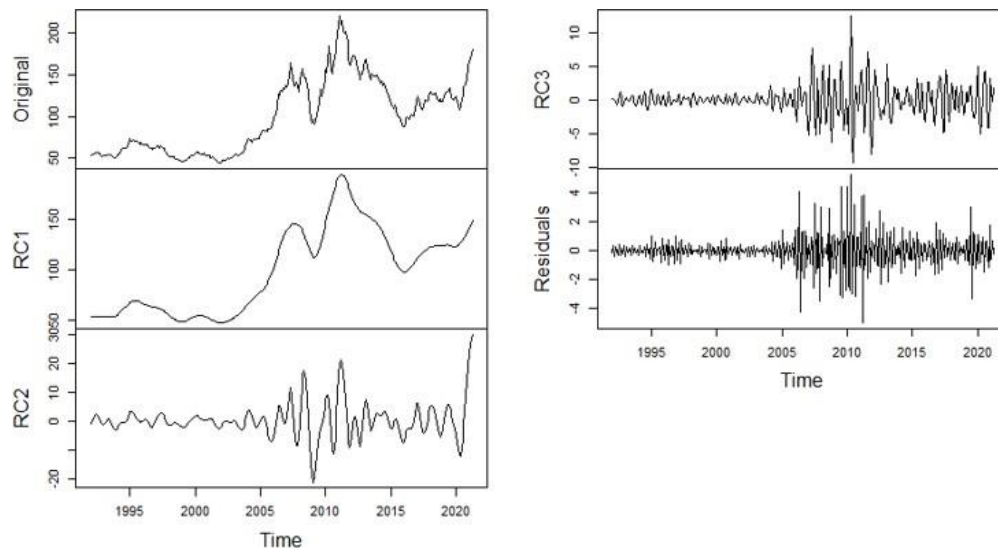


Fig. 7: Reconstructed series.

The recent Covid-19 pandemic whose impact on the cost of inputs was evident in mid of 2020 and spilled over to 2021. The recorded sharp and the extended increasing trend shows that the pandemic had the greatest impact compared to other shocks. COVID-19 disrupted nearly all the countries disrupting the supply chains of raw materials domestically and through imports. The restrictions enacted during the COVID-19 pandemic included the closure of regional and international borders, limiting the flow of raw materials. The seasonal component is a good instrument to identify possible shocks that repeat after some fixed quarterly, semi-annually or monthly interval. A visual examination of RC3 shows no clear pattern in the seasonality of the series. The seasonal component occurs in response to sudden shocks such as the 2007 post-election violence, 2008/09 financial crisis, and the recent COVID-19 pandemic. Overall, Kenya's industrial inputs price index is unstable, with an increasing trend over time. Like the reviewed studies, the current study results depict the suitability of SSA in extracting the time series components.

C. Residual diagnostics

The post-diagnostic tests performed to check for the adequacy of the proposed methodology entailed examining the distribution of the residuals. Fig. 8 shows that the residuals resemble white noise and are approximately normally distributed as indicated by the histogram. Besides, the Autocorrelation Function (ACF) depicts the partial autocorrelation coefficients of the residuals up to lag 36. A few autocorrelation coefficients exceed the 95% significance band; hence autocorrelation is not problematic.

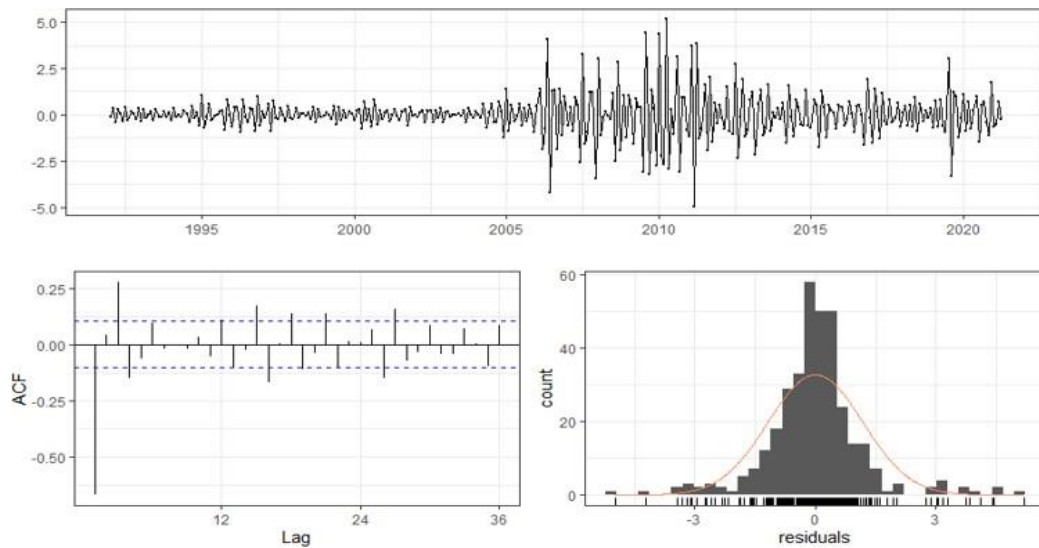


Fig. 8: SSA model's residuals.

Step 5: Forecasting

The ultimate goal of the reverse process from diagonal averaging is to obtain a univariate time series of length N . The sum of the four reconstructed components yields the fitted time series or in-sample forecasts (Fig. 9). The in-sample forecasts mimic the original series with minor deviations hence a good fit for the data. One of the most interesting findings is that the 24-step ahead estimates cannot be affected by an extended trend component in the last period of the estimation data. The model's forecasted curve elbows back after the second step-ahead predictions and mimicked the typical behavior of a time series.



Fig. 9: SSA ($L = 12, r = 7$) forecasts.

Table II provides the 24-step ahead predictions from the fitted SSA ($L = 12, r = 7$) model. The values show a decreasing trend from June 193.65 in 2021 to 178.43 in June 2022 by about 5.64%. The sharp fall is likely to be followed an increasing trend starting July 2022 178.79 to 184.71 in April 2023, representing about a 3.52% increase.

TABLE II: SSA FORECASTS

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2021					189.1	193.7	192.5	190.3	190.6	192.4	191.9	188.2
2022	184.6	183.9	184.7	183.8	180.8	178.4	178.8	180.6	181.2	180.3	180	181.6
2023	183.9	184.9	184.6	184.7								

V. CONCLUSION

Time series analysis and forecasting are essential in economic policymaking and evaluation; hence indispensable to policymakers. Forecasting requires models that produce a better fit for the data and accurately predict the future course of a given series. With the realization of the restrictive assumptions of parametric methods such as non-stationarity of the data, which does not always hold in most economic and financial time series, non-parametric models such as the SSA methodology are often preferable. Besides, SSA is a novel model compared to classical models since it is a forecasting model and a data filtering technique. The current study demonstrated that SSA methodology clearly distinguishes the time-series properties of Kenya's industrial inputs price index (level, trend, and seasonality) in coherence with external and internal shocks such as the 2008/09 financial crisis, recurrent political instability around the election period, and the recent COVID-19. The shocks have contributed to a sharp rise in the prices of industrial inputs. The fitted model provided realistic forecasts. The 24-period forecasts using the SSA ($L = 12, r = 7$) methodology showed that the industrial inputs price index is likely to rebound in the last quarter of 2021, followed by another increasing trend from the third quarter of 2022. The index is still high relative to 2017 before the post-industrial agenda targeting a reduction in the cost of industrial inputs. The high cost of industrial inputs is a disincentive to the investors in the industrial sector since it contributes to the overall increase in the cost of production, affecting the industrial sector's growth. Secondly, cost-push inflation is likely to occur as investors in the industry sector seek to protect their profit margins. Thus, there should be a concerted effort to reduce the cost-competitiveness disadvantage among Kenya manufacturers by reducing and stabilizing the industrial input prices to a sustainable level.

ACKNOWLEDGMENT

I thank God for giving me the strength and ability to complete my undergraduate and post-graduate studies at Chuka University.

Thanks to Chuka University for the opportunity to undertake my post-graduate studies. Special thanks go to staff from the Department of Physical Sciences who taught us throughout our studies at Chuka University.

My sincere gratitude goes to my supervisors, Adolphus Wagala (Department of Mathematics and Computer Science, Bomet University, Kenya) and Dennis K. Muriithi (Faculty of Science, Engineering and Technology, Chuka University, Kenya), for their steadfast guidance and support throughout this study.

FUNDING

The research received no external funding.

CONFLICT OF INTEREST

The authors declared no conflicts of interest concerning the authorship or publication of this article.

References

- [1] Bivens J. (2003). Updated employment Multipliers for the US Economy. Report, Economic Policy Institute, August.
- [2] Berger T, Chen C, Frey CB. Cities, Industrialization and Job Creation: Evidence from Emerging Countries. [Internet] mimeo, Oxford Martin School. 2017. Available from: https://www.oxfordmartin.ox.ac.uk/downloads/JEG_Manuscript.pdf.
- [3] World Bank. Manufacturing, value added (% of GDP) | Data [Internet]. Data.worldbank.org. 2021 [cited 19 September 2020]. Available from: <https://data.worldbank.org/indicator/NV.IND.MANF.ZS>
- [4] World Bank. Services, value added (% of GDP) - Kenya | Data [Internet]. Data.worldbank.org. 2021 [cited 14 September 2020]. Available from: <https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS?locations=KE>
- [5] World Bank. Agriculture, forestry, and fishing, value added (% of GDP) - Kenya | Data [Internet]. Data.worldbank.org. 2021 [cited 14 September 2020]. Available from: <https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?locations=K>
- [6] World Bank. Industry (including construction), value added (% of GDP) - Kenya | Data [Internet]. Data.worldbank.org. 2021 [cited 14 September 2020]. Available from: <https://data.worldbank.org/indicator/NV.IND.TOTL.ZS?locations=KE>
- [7] World Bank. Manufacturing, value added (% of GDP) - Kenya | Data [Internet]. Data.worldbank.org. 2021 [cited 14 September 2020]. Available from: <https://data.worldbank.org/indicator/NV.IND.MANF.ZS?locations=KE>
- [8] Kenya Association of Manufacturers. 2018-Manufacturing-Priority-Agenda: Sparking Kenya's Industrial Transformation for Job Creation [Internet]. 2018 [cited 14 September 2020]. Available from: <https://kam.co.ke/kam/wp-content/uploads/2018/02/2018-Manufacturing-Priority-Agenda.pdf>
- [9] Balke, NS. Detecting level shifts in time series. Journal of Business & Economic Statistics. 1993; 11(1): 81-92.
- [10] Hassani H, Mahmoudvand R. More on Filtering and Forecasting by SSA. In Singular Spectrum Analysis. 2018: 103-115.

- [11] Hassani H, Thomakos D. A review on singular spectrum analysis for economic and financial time series. *Statistics and Its Interface*. 2010; 3(3): 377-397.
- [12] Golyandina N, Nekrutkin V, Zhigljavsky AA. *Analysis of time series structure: SSA and related techniques*. CRC Press. 2001 Jan 23.
- [13] Bansal A, Dimri V. Depth determination from a non-stationary magnetic profile for scaling geology. *Geophysical Prospecting*. 2005; 53(3): 399-410.
- [14] Ghil M, Allen M, Dettinger M, Ide K, Kondrashov D, Mann M et al. *Advanced Spectral Methods for Climatic Time Series*. *Reviews of Geophysics*. 2002; 40(1): 31-41.
- [15] Rekapalli R, Tiwari R, Dhanam K, Seshunarayana T. T-x frequency filtering of high resolution seismic reflection data using singular spectral analysis. *Journal of Applied Geophysics*. 2014; 105: 180-184.
- [16] Ghil M, Vautard R. Interdecadal oscillations and the warming trend in global temperature time series. *Nature*. 1991; 350(6316): 324-327.
- [17] Sanei S, Ghodsi M, Hassani H. An adaptive singular spectrum analysis approach to murmur detection from heart sounds. *Medical Engineering & Physics*. 2011; 33(3): 362-367.
- [18] Rodríguez-Aragón L, Zhigljavsky A. Singular spectrum analysis for image processing. *Statistics and Its Interface*. 2010; 3(3): 419-426.
- [19] de Carvalho M, Rodrigues PC, Rua A. Tracking the US business cycle with a singular spectrum analysis. *Economics Letters*. 2012; 114(1): 32-5.
- [20] Vautard R, Yiou P, Ghil M. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Non-linear Phenomena*. 1992; 58(1-4): 95-126.
- [21] Briceno H, Rocco C, Zio E. Singular spectrum analysis for forecasting of electric load demand. *Chemical Engineering Transactions*. 2013; 33: 919-24.
- [22] Silva ES, Hassani H. On the use of singular spectrum analysis for forecasting US trade before, during and after the 2008 recession. *International Economics*. 2015; 141: 34-49.
- [23] Hassani H, Webster A, Silva ES, Heravi S. Forecasting US tourist arrivals using optimal singular spectrum analysis. *Tourism Management*. 2015; 46: 322-35.
- [24] Vile JL, Gillard JW, Harper PR, Knight VA. Predicting ambulance demand using singular spectrum analysis. *Journal of the Operational Research Society*. 2012; 63(11): 1556-65.
- [25] Menard S, editor. *Handbook of longitudinal research: Design, measurement, and analysis*. Elsevier; 2007 Oct 11.
- [26] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*. 2000; 97(18): 10101-6.
- [27] Datta B. *Numerical methods for linear control systems*. Academic Press; 2004.
- [28] Ghil M, Taricco C. *Advanced spectral-analysis methods. In Past and present variability of the solar-terrestrial system: measurement, data analysis and theoretical models*. IOS Press. 1997: 137-159.

- [29] Silva ES, Hassani H, Heravi S. Modeling European industrial production with multivariate singular spectrum analysis: A cross- industry analysis. *Journal of Forecasting*. 2018; 37(3): 371-84.
- [30] Syukur A, Marjuni A. Stock Price Forecasting Using Univariate Singular Spectral Analysis through Hadamard Transform. *International Journal of Intelligent Engineering and Systems*. 2020; 13(2): 96-107.
- [31] Golyandina N, Korobeynikov A. Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis*. 2014; 71: 934-54.
- [32] Box GE, Tiao GC. A change in level of a non-stationary time series. *Biometrika*. 1965; 52(1/2): 181-92.
- [33] Kenya National Bureau of Statistics. *Statistical Abstract 2010*. [Internet] 2010. [cited 2021 August 5] Available from: <https://www.knbs.or.ke/?wptdmpro=statistical-abstract-2010>
- [34] Kenya Revenue Authority. *Importing & exporting* [Internet]. KRA. 2020 [cited 2021 July 7]. Available from: <https://kra.go.ke/en/business/companies-partnerships/companies-partnerships-pin-taxes/company-partnership-imports-exemptions>