P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : *https://tci-thaijo.org/index.php/SciTechAsia* Science & Technology Asia
Vol. 29 No.4 October - December 2024 Page: [75-87]
Original research article

Forecasting Search Trends for "Silverqueen" Chocolate Keywords using the Singular Spectrum Analysis Method and the Hybrid Singular Spectrum Analysis-ARIMA Model

Gavrila Nathania Rambung*, Sri Wahyuningsih, Andrea Tri Rian Dani

Statistics Program Study, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda 75119, Indonesia

Received 3 June 2024; Received in revised form 29 November 2024
Accepted 3 December 2024; Available online 27 December 2024

ABSTRACT

This research uses a hybrid model of time series in the process. The Singular Spectrum Analysis (SSA) method will be combined with the Autoregressive Integrated Moving Average (ARIMA) method to model noise from SSA. This research applied the hybrid SSA-ARIMA method to Google Trends data, especially to the chocolate keyword "Silverqueen" search trend. This research aims to assess the accuracy and identify the best forecasting method for search trends for the keyword "Silverqueen" chocolate in Indonesia. Based on the results, the accuracy value obtained for the SSA method was 0.54% (MAPE) and 0.04 (RMSE) for in-sample data and 28.93% (MAPE) and 1.49 (RMSE) for out-sample data. The hybrid SSA-ARIMA (5.1.0) method has two outliers with an accuracy value of 0.35% (MAPE) and 0.02 (RMSE) for in-sample data and 31.00% (MAPE) and 1.50 (RMSE) for out-sample data. The results of the SSA forecasting method for the next 17 periods show that the trend will increase, with the highest trend occurring in the second week of February 2024, namely 100 points. Then, the forecast results of the hybrid SSA-ARIMA(5,1,0) method with outliers for the next 17 periods, the trend will increase, with the highest trend occurring in the second week of February 2024, namely around 95 points. The best method for forecasting search trends for the chocolate keyword "Silverqueen" is the SSA method.

Keywords: ARIMA;, Forecasting; Hybrid; Google Trends; SSA

***Corresponding author:** gavrilanathania13@gmail.com

1. Introduction

The Singular Spectrum Analysis (SSA) method uses the decomposition principle in its analysis. The basic principle in the decomposition method is to break down the time series of data into four basic series components and then identify each component of the time series separately. The SSA method algorithm consists of two main parts, namely decomposition and reconstruction. Decomposition consists of two different stages, namely embedding and singular value decomposition (SVD), while the reconstruction stage consists of grouping and diagonal averaging [1].

Researchers who have apply the SSA method in their research include [2] with MAPE forecasting accuracy results of 5%, [3] with forecasting accuracy on seasonal patterns using the SSA method smaller than the SARIMA method, and [4] which shows that the SSA method provides good prediction results on seasonal data patterns.

Currently, there are developments in the field of research, especially time series analysis, where researchers combine several individual methods in the analysis process which are known as hybrid methods. One of the hybrid methods that can be used currently is hybrid SSA and ARIMA.

SSA acts as preprocessing when combined with other methods [1]. The preprocessing application uses the SSA method, where the input data will be analyzed and produce noise components, which are then modeled using the ARIMA model. The SSA-ARIMA hybrid method forecast is obtained by combining the forecasting results from the SSA and ARIMA methods. ARIMA as a time series analysis method has many advantages, including being able to handle data that has non-stationary patterns with a differentiation process; also, it can be expanded into multivariable models such as the seasonal ARIMA (SARIMA) model and the autoregressive integrated moving average exogenous (ARIMAX) model [5, 6]. ARIMA consists of autoregressive and moving average components. Many studies use the ARIMA model for forecasting and obtain quite good results.

Several studies, including those by [7] on consumer price index data, [8] on O_3 concentration data, and [9] on inflation data, have applied the hybrid SSA-ARIMA method. This research shows that SSA-ARIMA can provide good forecasting performance.

The SSA-ARIMA hybrid method has been applied to various data types such as economic data, climatology, number of tourists, and even big data. Google Trends is an example of a platform for utilizing and providing big data that is easy to access and focuses on search trends or searches for keywords on Google pages within a certain period [10]. Google Trends is used in the business world as a medium for market research [11]. One type of information that can be obtained through Google Trends is food search trends, one of which is chocolate.

Chocolate is a food liked by various groups, including children, teenagers, and adults. Chocolate purchases usually increase during certain celebrations, for example, Valentine's Day, Eid al-Fitr, and Christmas. Google Trends data is used to predict the effects of calendar and seasonal variations on sales of chocolate using a classic time series approach [12]. This trend pattern can reveal a business opportunity in marketing activities for chocolate industry companies when planning their product launch strategies. Silverqueen chocolate is a popular chocolate brand that can compete in the Indonesian chocolate industry. This existence is proven by Silverqueen occupying the leading position in the chocolate bar category by having a very high-Top Brand Index (TBI) compared to other chocolate brands for five consecutive years based on the Top Brand Award (2021) website[13].

Based on the description previously presented, researchers will forecast Silverqueen chocolate trends using the

hybrid SSA-ARIMA method. The results of this forecasting can be used to illustrate Silverqueen's position in society, which helps related industries in industrial marketing strategic planning.

2. Materials and Methods

2.1 Periodogram analysis

The periodogram can be interpreted as a function of the power spectrum over its frequency [14]. Periodogram analysis was carried out to determine whether the observation data was influenced by seasonal factors. Periodogram analysis hypothesis testing is carried out using Eq. (2.1).

$$T = \frac{I^{(1)}(\omega_{(1)})}{\sum_{i=1}^{n/2} I(\omega_i)}, \qquad (2.1)$$

where $I^{(1)}(\omega_{(1)})$ is the largest periodogram value, $\omega_{(1)}$ is Fourier frequency in the largest periodogram value, $I(\omega_i)$ is the periodogram value at the *i*-th Fourier frequency, and ω_i is the *i*-th Fourier frequency [15].

2.2 Singular Spectrum Analysis

The basic SSA algorithm consists of two stages, namely decomposition and reconstruction [16].

2.2.1 Decomposition

Decomposition consists of two stages, namely embedding and singular value decomposition (SVD) [17].

1. Embedding
   At the embedding stage, the time series data is transformed into the form of a trajectory matrix $\mathbf{X}$. For example, the time series data with length n is represented by $\mathbf{Z}_t = \{Z_1, Z_2, ..., Z_n\}$. This data is transformed into a matrix of size $L \times K$ with $2 < L < n/2$. The trajectory matrix of the $\mathbf{Z}$ series is shown in Eq. (2.2) [16].

$$\mathbf{X} = [\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_K] = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_K \\ Z_2 & Z_3 & \cdots & Z_{K+1} \\ \vdots & \cdots & \ddots & \vdots \\ Z_L & Z_{L+1} & \cdots & Z_n \end{bmatrix} \qquad (2.2)$$

2. Singular Value Decomposition
   Singular value decomposition (SVD) aims to obtain component separation in decomposing time series data. The SVD of the trajectory matrix is as in Eq. (2.3).

$$\mathbf{X} = \mathbf{u}_1\sqrt{\eta_1}\mathbf{v}_1^T + \mathbf{u}_2\sqrt{\eta_2}\mathbf{v}_2^T + ... + \mathbf{u}_{r^*}\sqrt{\eta_{r^*}}\mathbf{v}_{r^*}^T \qquad (2.3)$$

Matrix $\mathbf{X}$ is formed from eigenvectors $(\mathbf{u}_i)$, singular values $(\sqrt{\eta_i})$, and principal components $\left(\mathbf{v}_i^T\right)$. The three elements that form SVD are called eigentriples [16].

2.2.2 Reconstruction

The reconstruction stage consists of two steps, namely grouping and diagonal averaging.

1. Grouping
   Grouping is the stage of grouping matrix $\mathbf{X}_I$. This grouping aims to separate the eigentriple components obtained at the SVD stage into several subgroups, namely trend, seasonality, and noise. Eigenvectors are the basis for grouping in the grouping process [2].

2. Diagonal Averaging
   Diagonal averaging is the stage of reconstructing grouping results into new time series data. Diagonal averaging transforms the $\mathbf{Y}^{(h)}$ matrix grouping results into a sequential form again using Eq. (2.4) [17].

$$\hat{f}_t^{(h)} = \begin{cases} \frac{1}{t}\sum_{i=1}^{t} y_{i,t-i+1}^{*(h)}, & 1 \le t < L^* \\ \frac{1}{L^*}\sum_{i=1}^{L^*} y_{i,t-i+1}^{*(h)}, & L^* \le t < K^* \\ \frac{1}{n-t+1}\sum_{t=t-K^*+1}^{n-K^*+1} y_{i,t-i+1}^{*(h)}, & K^* \le t < n \end{cases} \qquad (2.4)$$

The initial series will be decomposed into a number of m reconstructed series by using Eq. (2.5).

$$\hat{f}_t = \sum_{h=1}^{m} \hat{f}_t^{(h)}, \quad t = 1,2,...,n, \qquad (2.5)$$

where $h = 1$ is for the trend component and $h = 2$ is for the seasonal component [17].

2.2.3 SSA Forecasting

R-forecasting is related to estimating the Linear Recurrent Formula (LRF). The LRF coefficient of a component can be calculated using Eq. (2.6) [18].

$$\mathbf{r}^{(h)} = (r_{L-1}^{(h)},...,r_1^{(h)}) = \frac{1}{1-v^2}\sum_{i=1}^{l} \sigma_i \mathbf{u}_i^{\nabla}. \qquad (2.6)$$

The prediction and forecasting results time series can be written with Eq. (2.6).

$$\hat{f}_t^{(h)} = \begin{cases} \hat{f}_t^{(h)}, & t = 1,2,...,n \\ \sum_{j=1}^{L-1} r_j^{(h)} \hat{f}_{t-j}^{(h)}, & t = n+1,...,n+M \end{cases} \qquad (2.6)$$

2.3 Autoregressive Integrated Moving Average

Autoregressive Integrated Moving Average (ARIMA) is a development of the ARMA model that uses differencing because the data is not yet stationary. The ARIMA process begins by identifying the model to determine whether the time series data is stationary or not, both in terms of variance and average. If stationarity in the variance is not fulfilled, then a power transformation is carried out using Eq. (2.7) [15].

$$Z_t^* = \begin{cases} \dfrac{Z_t^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(Z_t), & \lambda = 0 \end{cases}. \qquad (2.7)$$

Detecting stationarity in average time series data can be done using a time series data plot, an autocorrelation function (ACF) plot, and an augmented Dickey Fuller (ADF) hypothesis test. If the data is not stationary, differencing is carried out [19].

The next stage in model identification is to form an ACF plot and a partial autocorrelation function (PACF) plot. The autoregressive order and moving average order in the ARIMA temporary model can be determined via the ACF and PACF plot. The ARIMA model (p,d,q) is written as in Eq. (2.8).

$$\phi_p(B)(1-B)^d Z_t = \theta_q(B) e_t, \qquad (2.8)$$

After obtaining the temporary ARIMA model, the ARIMA model parameters were estimated using conditional least squares (CLS) [15]. Then, a diagnostic examination was performed, consisting of testing the significance of parameters using the *t*- test, testing normally distributed residuals using the Kolmogorov-Smirnov test, and testing the independence of residuals using the Ljung-Box test [19].

If the residual normality assumption is not met, outlier detection can be carried out. There are four types of outliers, one of which is an additive outlier (AO). The general model of additive outliers in time series analysis is given in Eq. (2.9) [20].

$$\hat{Z}_t = \begin{cases} Z_t, & t \neq T \\ Z_t + \psi I_t^{(T)}, & t = T \end{cases}, \qquad (2.9)$$

where,

$$I_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}, \qquad (2.10)$$

2.4 Hybrid SSA-ARIMA

Generally, a hybrid structure in a time series combines two or more different forecasting methods. The hybrid model can be written as in Eq. (2.11).

$$\hat{H}_t = \hat{f}_t + \hat{g}_t, \qquad (2.11)$$

where $\hat{H}_t$ is the forecasting result of hybrid model, $\hat{f}_t$ is the forecasting result from the SSA method and $\hat{g}_t$ is the forecasting result

from the ARIMA method with the data used is noise component data from the SSA method [21].

2.5 MAPE

The measures of forecasting accuracy that will be used to select the best model in this research are mean absolute percentage error (MAPE) and root mean square error (RMSE). MAPE is obtained using Eq. (2.12) [19].

$$MAPE = \frac{1}{n}\left[\sum_{t=1}^{n}\left|\frac{Z_t - F_t}{Z_t}\right| \times 100\%\right]. \quad (2.12)$$

RMSE is used to evaluate forecasting results. RMSE is a good measure of accuracy, but it is only used to compare forecasting errors of different models for a particular variable and not between them; the accurate method is the RMSE that produces the smallest value. RMSE is obtained using Eq. (2.13) [9].

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(Z_t - F_t)^2}\ . \quad (2.13)$$

3. Results and Discussion

3.1 Data Description

The data used in this research is weekly data on search trends for "Silverqueen" chocolate keywords for 2019-2023, totaling 261 data items obtained via the official Google Trends website. The data was divided into a proportion of 80:20 in 209 in-sample data and 52 out-sample data. A time series plot created to see the data pattern is displayed in Fig. 1.

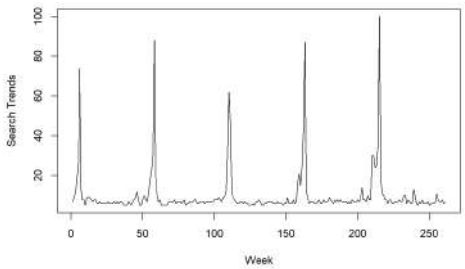


Fig. 1. Time series plot of "Silverqueen" chocolate keywords.

Based on Fig. 1, it can be seen that the weekly data on search trends for the keyword "Silverqueen" chocolate in Indonesia for the period January 2019 to December 2023 does not form a stationary or trending pattern but tends to form a seasonal pattern. This is because the pattern of the data experiences repeated increases and decreases periodically every second week of February. Because the data's seasonal pattern is clearly visible, it can be seen that the seasonal period for the search trend for the keyword "Silverqueen" chocolate in Indonesia is 52 weeks.

3.2 Singular Spectrum Analysis

3.2.1 Decomposition

The initial stage of decomposition is determining the window length (L) parameter through a trial-and-error process for all possible L parameters, namely $3 < L < n/2 = 3 < L < 104$, where $n = 209$. Based on the MAPE criteria, the smallest MAPE is 0.54%, which is found at window length $L = 53$. MAPE is used because it has clear categories and interpretations of its percentage value.

a. Embedding
The embedding stage is carried out by transforming the actual time series data in vector form into a path matrix $\mathbf{X}$ of size $L \times K$ with values $L = 53$ and $K = n - L + 1 = 209 - 53 + 1 = 157$. The results of the embedding process using Eq. (2.2) are as follows:

$$\mathbf{X} = (\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, ..., \mathbf{Z}_{157}) = \begin{pmatrix} 7 & 9 & 12 & \cdots & 7 \\ 9 & 12 & 18 & \cdots & 15 \\ 12 & 18 & 26 & \cdots & 21 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 7 & 12 & 18 & \cdots & 8 \end{pmatrix}_{(53\times157)}$$

b. Singular Value Decomposition
Singular Value Decomposition begins by forming a singular matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^T$.

$$\mathbf{S} = \mathbf{X}\mathbf{X}^T = \begin{pmatrix} 30.35 & 21.63 & 17.36 & \cdots & 30.16 \\ 21.63 & 30.56 & 21.90 & \cdots & 21.18 \\ 17.36 & 21.90 & 30.91 & \cdots & 17.52 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 30.16 & 21.18 & 17.52 & \cdots & 34.68 \end{pmatrix}_{(53\times53)}$$

The SVD process will decompose the path matrix **X** into 53 eigentriples each consisting of 53 eigenvalues, 53 eigenvectors and 53 principal components.

3.2.2 Reconstruction

After decomposing the data, the next stage is reconstruction. The reconstruction process consists of grouping and diagonal averaging.

a. Grouping

The grouping stage begins by grouping the eigentriples obtained in the SVD stage. The plot between the singular value and the window length parameter used to determine the grouping effect (R) value is as follows:.

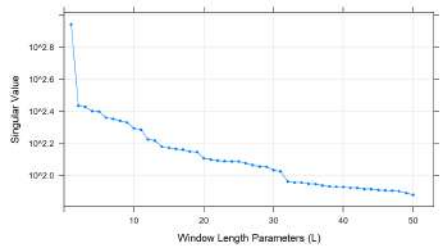


Fig. 2. Singular value plot.

Plot sequences that decrease slowly or gradually from singular values are usually associated with the noise component of the series. The eigentriples that have been separated in Fig. 2 produce a less subjective grouping, so the grouping effect (R) value taken is $R = 53$ where all eigentriples will be rechecked. You can use an eigenvector plot to determine the eigentriple that contains this element. The eigenvector plot of 50 eigentriples is as follows.

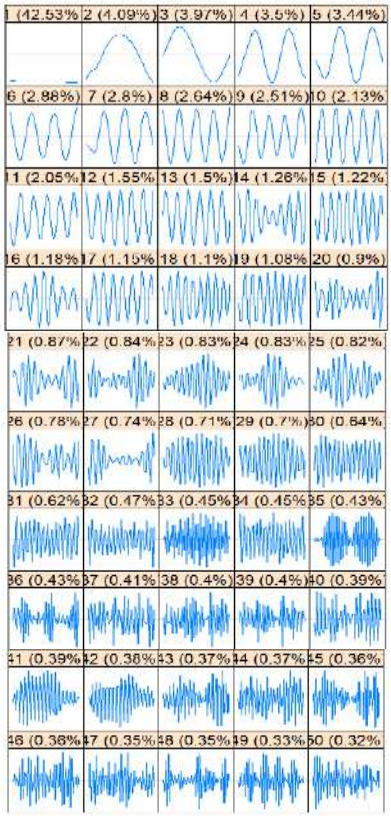


Fig. 3. Eigenvector plot.

Because R software can only display 50 eigenvector plots, the 51st to 53rd eigenvectors are identified through the plot percentages presented in Table 1.

Table 1. Percentage Contribution of 51st to 53rd Eigenvector Plot.

Eigenvector	Percentage
51	0.31%
52	0.30%
53	0.10%

Slowly varying eigenvector graphic patterns should be grouped into trend components. Based on Fig. 3, the plot of eigenvector 1 varies slowly, or in other words, there are no repeated increases and decreases; so eigentriple 1 is grouped into the trend component. Next is the grouping of the eigentriples that contain seasonal elements in eigentriples 2 to 53 using periodogram analysis, and the eigentriple groupings are obtained as in Table 2.

Table 2. Eigentriple Grouping.

Group	Eigentriple
Trend	1
Seasonal	2,3,4,5,6,7,9,10,11, …., 52
Noise	53

b. Diagonal Averaging

The next stage is diagonal averaging where each component will be reconstructed into a new series using each associated eigentriple. The reconstructed series is then formed into a plot for the trend, seasonal, and noise components, respectively, as in Fig. 4.

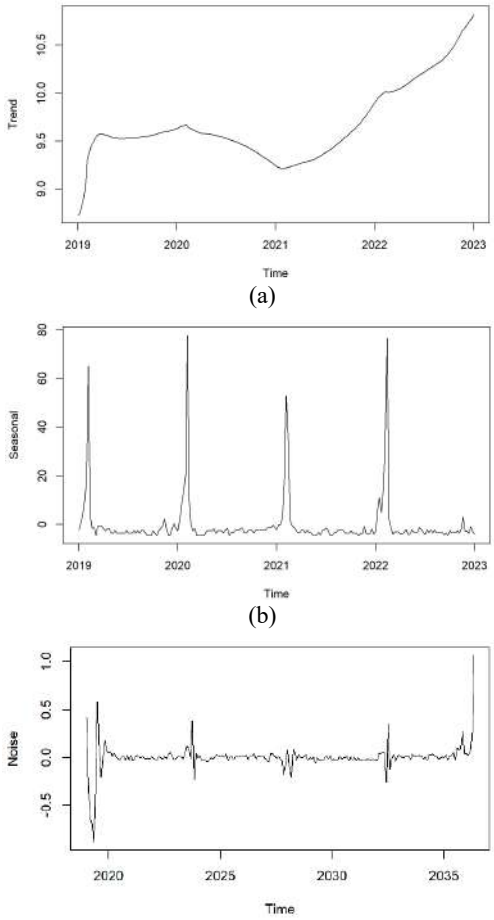


Fig. 4. Reconstructed components.

Diagonal averaging is obtained from the sum of the trend and seasonal reconstruction results. The series of reconstruction results and diagonal averaging can be seen in Table 3.

Table 3. Reconstruction and Diagonal Averaging Results.

t	Reconstruction		Diagonal Averaging
	Trend	Seasonal	
1	8.73	-2.15	6.58
2	8.75	0.46	9.21
3	8.82	3.80	12.62
4	8.91	9.79	18.69
5	9.01	17.88	26.89
6	9.30	65.09	74.39
7	9.39	3.03	12.42
8	9.45	-1.46	7.98
9	9.49	-1.28	8.21
10	9.52	-4.55	4.97
⋮	⋮	⋮	⋮
205	10.70	-2.73	7.97
206	10.72	-3.75	6.97
207	10.76	-0.81	9.95
208	10.78	-3.03	7.75
209	10.82	-3.90	6.93

3.2.3 SSA Forecasting

The SSA forecasting used is the R-forecasting method. The forecasting model for the trend component is as follows:

$$\hat{f}_t^{(1)} = 0.02\hat{f}_{t-1}^{(1)} + 0.02\hat{f}_{t-2}^{(1)} + ... + 0.01\hat{f}_{t-52}^{(1)}.$$

Meanwhile, the forecasting model for the seasonal component is as follows:

$$\hat{f}_t^{(2)} = 0.11\hat{f}_{t-1}^{(2)} + (-0.13)\hat{f}_{t-2}^{(2)} + ... + 0.99\hat{f}_{t-52}^{(2)}.$$

SSA forecasting is carried out by adding the forecast results of the trend and seasonal components using the forecasting model formed for these two components. Forecasting data for the first week of January 2023 to December 2024 using the SSA method is presented in Table 4.

Table 4. Results of Forecasting Search Trends for "Silverqueen" Chocolate with the SSA Model.

t	Forecasting		Diagonal Averaging
	Trend	Seasonal	
1*	10.57	4.56	15.13
2*	10.59	14.36	24.95
3*	10.60	6.92	17.51
4*	10.61	9.32	19.93
5*	10.62	20.82	31.44
⋮	⋮	⋮	⋮
51*	11.21	-2.56	8.65
52*	11.22	-5.61	5.60
53	11.22	1.93	13.16
54	11.24	17.61	28.85

55	11.25	11.90	23.15
56	11.26	6.93	18.20
57	11.28	8.69	19.96
⋮	⋮	⋮	⋮
68	11.42	-1.34	10.07
69	11.43	-7.10	4.33

Note: The period marked (*) is the result of out-sample data forecasting

The prediction and forecasting results that have been obtained are then presented in graphical form in Fig. 5.

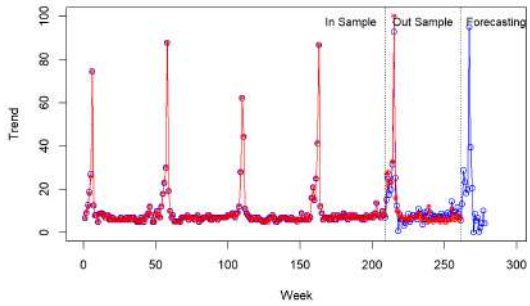


Fig. 5. Plot of actual data and SSA method forecasting.

Based on Fig. 5, the pattern in the plot resulting from forecasting search trends for the chocolate keyword "Silverqueen" using the SSA model almost follows the actual data pattern. The accuracy value obtained was 0.54% (MAPE) and 0.04 for in-sample data and 28.93% (MAPE) and 1.49 (RMSE) for out-sample data.

3.3 Hybrid SSA-ARIMA

Hybrid SSA-ARIMA modeling uses noise component data from the reconstruction results of the SSA method to capture time series patterns in noise that are not modeled in the SSA method. The time series plot of noise component is displayed as follows:

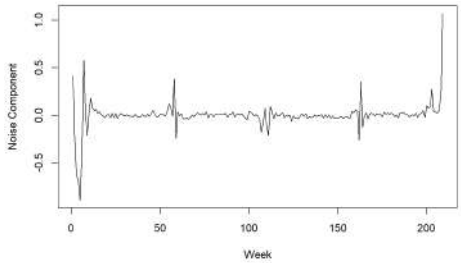
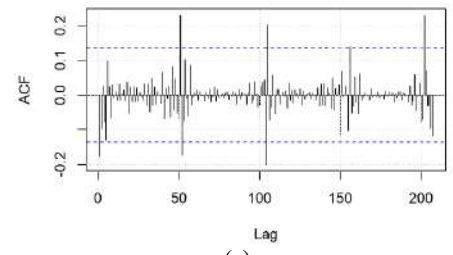


Fig. 6. Time series plot of noise components.

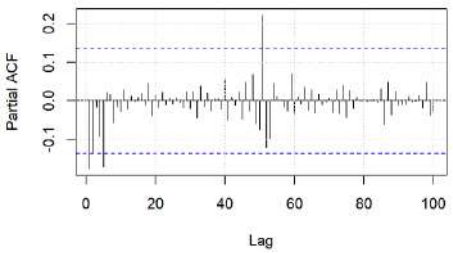
Based on Fig. 6, the noise component data does not form seasonal patterns and trends, but tends to form stationary patterns. This is because the pattern of the data fluctuates around a constant average value line.

In analysis using the ARIMA method, it is necessary to check stationarity in variance with the Box-Cox transformation. Data is considered stationary in terms of variance if the value of λ is close to or has a value of 1. Based on the test results, the value of $\lambda = 1.18$ is still far from 1, which means that the data is not yet stationary in variance and requires a lambda power transformation. After carrying out a power transformation on the noise component data, the value $\lambda = 1.00$ is obtained. Because the value of λ has a value of 1, the conditions for the estimated value have been fulfilled and it can be concluded that the noise component data is stationary in variance.

Next, the stationarity of the average noise component data after transformation is checked using the ADF test. Based on the ADF test, the results obtained are $p - value = 0.98$. This value is bigger than alpha, where $\alpha = 0.05$, which means that the noise component data after transformation is not yet stationary on average, so it is necessary to carry out first order of differencing on the noise component data after transformation. ADF testing after differencing gives a result of $p - value =< 2 \times 10^{-6}$. This value is less than alpha, where $\alpha = 0.05$ which means that the noise component data after transformation and first-order differencing is stationary on average. If the data is stationary in variance and average, an ACF and PACF plot is created, as in Fig. 7.



(a)



(b)

Fig. 7. Plot of ACF(a) and PACF(b) of noise components after transformation and differencing.

Based on Fig. 7(a), the ACF cut off values after lag are 1, 51, 52, 104, 105, 156, and 202. Using the parsimony principle, only lag 1 will be used, so that the order for the moving average (q) is obtained. 1. Then, based on Fig. 7(b), the PACF cut off value after lag 1, 5, 51. Using the parsimony principle, only lags 1 and 5 will be used so that the possible order for autoregressive (p) is 1,2,3, 4, and 5. There are 11 temporary ARIMA models were formed as in Table 5.

Table 5. Temporary ARIMA Model.

No	Model	No	Model
1	ARIMA(0,1,1)	7	ARIMA(3,1,1)
2	ARIMA(1,1,0)	8	ARIMA(4,1,0)
3	ARIMA(1,1,1)	9	ARIMA(4,1,1)
4	ARIMA(2,1,0)	10	ARIMA(5,1,0)
5	ARIMA(2,1,1)	11	ARIMA(5,1,1)
6	ARIMA(3,1,0)		

The best model selection from eleven temporary ARIMA models was conducted with a diagnostic examination, which included parameter significance testing, residual normality assumption testing, and residual independence assumption. The results of parameter estimation and significance testing are presented in Table 6.

Based on parameter significance testing in Table 6, it can be concluded that the models that have significant parameters are the ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(2,1,0), ARIMA(4,1,0), ARIMA(4,1,1), and ARIMA(5,1,0) because each parameter in the model has a p-value smaller than the significance level $\alpha = 0.05$.

The results of testing the residual normality assumption using the Kolmogorov-Smirnov test are presented in Table 7.

Table 6. Parameter Estimation and Significance Testing.

Model	Parameter Estimation	*p-value*	Model	Parameter Estimation	*p-value*
ARIMA(0,1,1)	$\hat{\theta}_1 = -0.3394$	0.0009	ARIMA(4,1,1)	$\hat{\phi}_1 = -0.3183$	0.0194
ARIMA(1,1,0)	$\hat{\phi}_1 = -0.2113$	0.0033		$\hat{\phi}_2 = -0.4222$	2.377×10^{-5}
ARIMA(1,1,1)	$\hat{\phi}_1 = 0.1903$	0.0044		$\hat{\phi}_3 = -0.2528$	0.0042
	$\hat{\theta}_1 = -0.7751$	$<2.2\times10^{-16}$		$\hat{\phi}_4 = -0.2759$	0.0001
ARIMA(2,1,0)	$\hat{\phi}_1 = -0.3602$	2.045×10^{-6}		$\hat{\theta}_1 = -0.2800$	0.0243
	$\hat{\phi}_2 = -0.2109$	0.0030	ARIMA(5,1,0)	$\hat{\phi}_1 = -0.6593$	9.380×10^{-15}
ARIMA(2,1,1)	$\hat{\phi}_1 = 0.1099$	0.2736		$\hat{\phi}_2 = -0.6358$	1.599×10^{-11}
	$\hat{\phi}_2 = -0.2034$	0.0036		$\hat{\phi}_3 = -0.4902$	2.963×10^{-11}
	$\hat{\theta}_1 = -0.6564$	2.397×10^{-15}		$\hat{\phi}_4 = -0.3930$	2.564×10^{-6}
ARIMA(3,1,0)	$\hat{\phi}_1 = -0.4557$	2.287×10^{-8}		$\hat{\phi}_5 = -0.3546$	2.761×10^{-7}

	$\hat{\phi}_2$	$= -0.3342$	4.091×10^{-5}	ARIMA(5,1,1)	$\hat{\phi}_1$	$= 0.2524$	0.0183
	$\hat{\phi}_3$	$= -0,1087$	0.1282		$\hat{\phi}_2$	$= -0,0806$	0.3434
ARIMA(3,1,1)	$\hat{\phi}_1$	$= 0.0156$	0.9034		$\hat{\phi}_3$	$= 0.1093$	0.2025
	$\hat{\phi}_2$	$= -0.2028$	0.0213		$\hat{\phi}_4$	$= 0.0094$	0.9066
	$\hat{\phi}_3$	$= -0.1555$	0.0291		$\hat{\phi}_5$	$= -0.1414$	0.0562
	$\hat{\theta}_1$	$= -0.5840$	6.092×10^{-8}		$\hat{\theta}_1$	$= -0.8892$	$<2.2\times10^{-16}$
ARIMA(4,1,0)	$\hat{\phi}_1$	$= -0.5536$	1.633×10^{-11}				
	$\hat{\phi}_2$	$= -0.5202$	6.236×10^{-9}				
	$\hat{\phi}_3$	$= -0.3184$	0.0001				
	$\hat{\phi}_4$	$= -0.2199$	0.0015				

Table 7. Residual Normality Test Results.

Model	*p-value*
ARIMA(0,1,1)	$<2.2\times10^{-16}$
ARIMA(1,1,0)	$<2.2\times10^{-16}$
ARIMA(1,1,1)	$<2.2\times10^{-16}$
ARIMA(2,1,0)	$<2.2\times10^{-16}$
ARIMA(4,1,0)	5.55×10^{-16}
ARIMA(4,1,1)	1.19×10^{-14}
ARIMA(5,1,0)	4.53×10^{-14}

Based on Table 7, it can be concluded that all ARIMA models formed do not meet the assumption of residual normality because each model has a *p*-value smaller than the significance level $\alpha = 0.05$.

Residual independence testing was carried out using the Ljung-Box test. The ARIMA model is said to meet the assumption of residual independence if all *p*-values for each lag in each model are greater than the significance level $\alpha = 0.05$. Based on the Ljung-Box test, it was found that all ARIMA models did not meet the assumption of residual independence or autocorrelation occurred because there was a lag that had a *p*-value smaller than the significance level $\alpha = 0.05$.

Because no model meets the assumptions of residual normality and residual independence, outlier detection is carried out. It is suspected that outliers in the data are causing the assumptions to not be met. Outlier detection was carried out on the seven ARIMA models that had significant parameters with the additive outlier (AO) type.

Outlier detection shows that only the ARIMA(5,1,0) model with added outliers meets the residual independence assumption, but it does not meet the residual normality assumption. There were 14 outliers detected in the ARIMA(5,1,0) model, namely in data 4, 5, 6, 7, 9, 10, 11, 59, 205, 206, 207, 208, and 209. The results of the significance test of the parameters, as well as the assumptions of residual normality and residual independence, are presented in Table 8.

Table 8. Recapitulation of Significance Test of ARIMA Model Parameters and Assumptions with Outliers.

ARIMA (5,1,0) & Outlier Data	Parameter Significance	Independence	Normality
5	√	√	x
5,7	√	√	x
5,7,9	x	√	x
5,7,9,59	x	√	x
5,7,9,59,205	√	√	x
5,7,9,59,205,206	x	x	x

5,7,9,59,205,206,207	x	x	x
5,7,9,59,205,206,207,208	x	x	x
5,7,9,59,205,206,207,208,209	x	x	x
5,7,9,59,205,206,207,208,209,4	x	x	x
5,7,9,59,205,206,207,208,209,4,6	x	x	x
5,7,9,59,205,206,207,208,209,4,6,8	x	x	x
5,7,9,59,205,206,207,208,209,4,6,8, 10	x	x	x
5,7,9,59,205,206,207,208,209,4,6,8, 10,11	x	x	x

Table 8 shows that with AO type outlier detection, three ARIMA(5,1,0) models with outliers meet the assumption of residual independence, but no model meets the assumption of residual normality. The best ARIMA(5,1,0) model with normality assumptions deemed fulfilled is the ARIMA(5,1,0) model with the 5th and 7th outlier data, which has the smallest MAPE of 193.78%. So the model used to forecast the noise component data is the ARIMA(5,1,0) model with the 5th and 7th outlier data, which mathematically can be written as follows:

$$\hat{Z}_t = Z_{t-1} - 0.86Z_{t-1} + 0.86Z_{t-2} - 0.62Z_{t-1} + 0.64Z_{t-3} - 0.43Z_{t-1} + 0.43Z_{t-4} - 0.38Z_{t-1} + 0.38Z_{t-5} - 0.23Z_{t-1} + 0.23Z_{t-6} + e_t - 2.22I_t^{(5)} + 0.41I_t^{(7)}$$

where,

$$I_t^{(5)} = \begin{cases} 1, & t = 5 \\ 0, & t \neq 5 \end{cases} \quad \text{and} \quad I_t^{(7)} = \begin{cases} 1, & t = 7 \\ 0, & t \neq 7 \end{cases}.$$

Based on this model, the noise component forecasting results obtained are presented in Fig. 8.

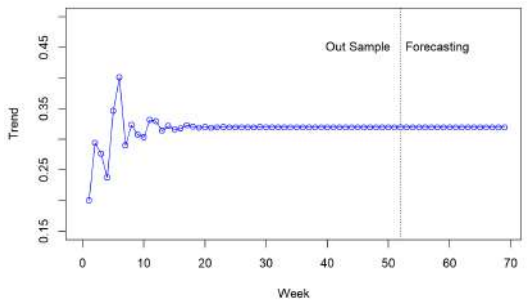


Fig. 8. ARIMA forecasting of noise components.

SSA-ARIMA hybrid forecasting is obtained by adding up the SSA and ARIMA models' forecasting results. The results of predicting and forecasting search trends for "Silverqueen" chocolate keywords in Indonesia for the period January 2019 to April 2024 using the hybrid SSA-ARIMA(5,1,0) model with the 5th and 7th data outliers are presented in graphical form in Fig. 9.

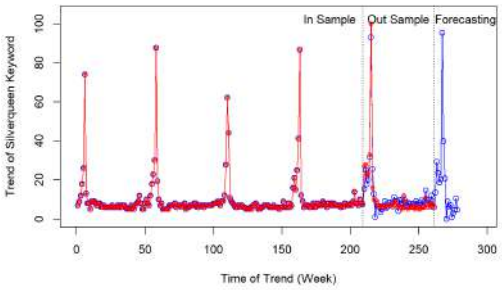


Fig. 9. Plot of actual data and forecasting results of the hybrid SSA-ARIMA(5,1,0) model with outliers.

Based on Fig. 9, the pattern in the plot resulting from forecasting search trends for the chocolate keyword "Silverqueen" using the hybrid SSA-ARIMA(5,1,0) model with the 5th and 7th data outliers tends to follow the actual data pattern. The accuracy value obtained was 0.35% (MAPE) and 0.02 (RMSE) for in-sample data and 31.00% (MAPE) and 1.50 (RMSE) for out-sample data. This condition is caused by the prediction results on the in-sample data that are outside the range of the Google trend value, which is 0-100. Predictions using the SSA method take into account several previous time series periods so that they have an impact on the prediction results of the out-

sample data which have a accuracy value greater than the MAPE of the in-sample data.

4. Conclusion

Based on the results of data analysis and discussion, it is concluded that, for accuracy in forecasting search trends for the chocolate keyword "Silverqueen" based on Google Trends data with a data proportion of 80:20 using the SSA method, the accuracy value is 0.54% (MAPE) and 0.04 (RMSE) for in-sample data and 28.93% (MAPE) and 1.49 (RMSE) for out-sample data. If the hybrid SSA-ARIMA(5,1,0) method is used with the 5^{th} and 7^{th} data outliers where the assumption of residual normality is deemed to be met, the accuracy value obtained is 0.35% (MAPE) and 0.02 (RMSE) for in-sample data and 31.00% (MAPE) and 1.50 (RMSE) for out-sample data.

The best method that can be used in forecasting search trends for the chocolate keyword "Silverqueen" based on Google Trends data in Indonesia, is the SSA method which has a MAPE value of 0.54% for in-sample data. However, the SSA method has limitations, namely that it can only be used on data containing seasonal patterns because it has the basic principle of decomposing data patterns into trends, seasonality, and noise.

This research can be a guideline for related industries in determining marketing strategy policies. Silverqueen is one of Indonesia's local chocolate products that has been exported to various countries; through this research it can be estimated related to the potential income for related industries and collaborating parties and the country's export value.

Reference

[1] Goyaldina N, Nekrutkin V, Zhigljavsky A. Analysis of Time Series Structure: SSA and Related Techniques. United States of America: Chapman&Hall/CRC; 2001.

[2] Ischak R, Asrof A, Darmawan G. Peramalan Rata-Rata Harga Beras di Tingkat Penggilingan Menggunakan Model Singular Spectrum Analysis (SSA). Prosiding Seminar Nasional Matematika dan Pendidikan Matematika. Yogyakarta; 2018.

[3] Puspita W, Rustiana S, Suparman, Purwandari T. Perbandingan Hasil Peramalan Curah Hujan Bulanan Kota Bogor Dengan Seasonal Autoregressive Integrated Moving Average (SARIMA) dan Singular Spectrum Analysis (SSA). Prosiding SENDIKA. 2019;5(2):206-17.

[4] Hidayat K, Wahyuningsih S, Nasution Y. Pemodelan Jumlah Titik Panas di Provinsi Kalimantan Timur Dengan Metode Singular Spectrum Analysis. Jambura Journal of Probability and Statistics. 2020;1(2):78-88.

[5] Lai Y, Dzombak D, Use of the Autoregressive Integrated Moving Average (ARIMA) Model to Forecast Near-Term Regional Temperature and Precipitation. American Meteorological Society Journals. 2020;35(3):959-76.

[6] Kusumaningrum N, Purnamasari I, Siringoringo M. Peramalan Menggunakan Model Hybrid ARIMAX-NN untuk Total Transaksi Pembayaran Nontunai. Journal Of Statistics and Its Application on Teaching and Research. 2023;5(1):1-14.

[7] Ilahi E, Zukhronah E, Susanti Y. Model Hibrida Singular Spectrum Analysis (SSA) dan Autoregressive Integrated Moving Average (ARIMA) untuk Peramalan Indeks Harga Konsumen. Jurnal Seminar Nasional Pendidikan Matematika Ahmad Dahlan. 2023;7(11):72-81.

[8] Kumar U. An Integrated SSA-ARIMA Approach to Make Multiple Day Ahead Forecast for the Daily Maximum Ambient O3 Concentration. Aerosol and Air Quality Research. 2015;15(1): 208-19.

[9] Arumsari M, Wahyuningsih S, Siringoringo M. Peramalan Inflasi Provinsi Kalimantan Timur Menggunakan Model Hybrid Singular Spectrum

Analysis Autoregressive Integrated Moving Average. Jurnal Matematika Sains dan Komputasi. 2021;18(1):78-92.

[10] Chumnumpan P, Shi X, Understanding New Products' Market Performance Using Google Trends. Australian Marketing Journal. 2019;27(2):91-103.

[11] Boone T, Ganeshan R, Jain A, Sanders N. Forecasting Sales In The Supply Chain: Consumer Analytics In The Big Data Era. International Journal of Forecasting. 2019;35(1):170-80.

[12] Diksa I. Forecasting the Existence of Chocolate with Variation and Seasonal Calendar Effects Using the Classic Time Series Approach. Jurnal Matematika, Statistika & Komputasi. 2022;18(2):237-50.

[13] Fitri A, Anindita R, Nugroho C. Analisis Tingkat Kepuasan Konsumen dan Loyalitas Merek Cokelat Silverqueen di Kabupaten Pekalongan. Jurnal Ekonomi Pertanian dan Agribisnis. 2023;7(1):127-45.

[14] Husnita F, Wahyuningsih S, Nohe D. Analisis Spektral dan Model ARIMA untuk Peramalan Jumlah Wisatawan di Dunia Fantasi Taman Impian Jaya Ancol. Jurnal Eksponensial. 2015;6(1): 21-9.

[15] Wei W. Time Series Analysis: Univariate and Multivariate Methods. New York: Pearson Education; 2006.

[16] Khaeri H, Yulian E, Darmawan G. Penerapan Metode Singular Spectrum Analysis (SSA) Pada Peramalan Jumlah Penumpang Kereta Api Di Indonesia Tahun 2017. Jurnal Euclid. 2018;5(1):8-20.

[17] Wulandari G, Wahyuningsih S, Siringoringo M, Sergio A. Inflation Forecasting for Samarinda City Using Hybrid Singular Spectrum Analysis-Neural Network Model. The 4th International Conference on Mathematics and Sciences: AIP Conferences Proceedings. 2024;3095(1).

[18] Hassani H, Mohamoudvand R. Singular Spectrum Analysis With R. Iran: Palgrave Advanced Texts in Econometrics; 2018.

[19] Aswi , Sukarna. Analisis Runtun Waktu dan Teori. Makassar: Andira Publisher; 2006.

[20] Suparti, Sa'adah A. Analisis Data Inflasi Indonesia Menggunakan Model Autoragressive Integrated Moving Average (ARIMA) Dengan Penambahan Outlier. Media Statistika. 2015;8(1):1-11.

[21] Waeto S, Chuarkham K, Intarasit A. Forecasting Time Series Movement Direction with Hybrid Methodology. Journal of Probability and Statistics. 2017; 2-3.