JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 23, No. 1, September 2026, pp. 59-75
DOI: 10.20956/kxgq2516

A Study of Multivariate Time Series Clustering Based on Vector Autoregressive with Exogenous Models for Forecasting Cooking Oil Prices in Indonesia

Mawar Cahyanindri[1], Embay Rohaeti[2*], Maya Widyastiti[3]
[1,2,3]Program Studi Matematika, FMIPA, Universitas Pakuan, Bogor
***Email**: [1]cahyanindrim@gmail.com, [2]embay.rohaeti@unpak.ac.id, [3]maya.widyastiti@gmail.com*
**Corresponding Author*
Received: 6 April 2026, revised: 27 July 2026, accepted: 3 August 2026

Abstract

The fluctuations in cooking oil prices in 77 regencies and cities included in the Strategic Food Price Information Center data show varying levels of volatility over time. This phenomenon presents a challenge for forecasting cooking oil prices in Indonesia in the future. Consequently, a method capable of estimating these prices is required. One such approach involves Multivariate Time Series Clustering (MTSCLUST) method based on the Vector Autoregressive with Exogenous (VARX) model. The MTSClust method is employed to map spatial time series pattern interdependencies across regions, while the VARX model is utilized for its ability to capture the dynamics of intervariable relationships influenced by external factors. This study aims to examine the application of the MTSClust method based on the VARX model in forecasting cooking oil prices in Indonesia in the future. Cluster formation results with K-Means with the Chebyshev distance metric indicate that the 77 regencies and cities in Indonesia can be grouped into three clusters, achieving excellent forecasting accuracy with MAPE values below 10%. Cluster 3 became the best cluster using the VARX (24,1) model, demonstrating optimal performance with a MAPE value of 0.228%. The findings of this study confirm that the MTSClust model with VARX integration is effective for forecasting future cooking oil prices.

Keywords: VARX, MTSClust, K-Means Chebyshev, Forecasting, Cooking Oil

Abstrak

Fluktuasi harga minyak goreng di 77 Kabupaten dan Kota yang tergolong dalam data Pusat Informasi Harga Pangan Strategis menunjukkan tingkat volatilitas yang beragam dari masa ke masa. Fenomena tersebut menjadi tantangan dalam proses peramalan harga minyak goreng di Indonesia pada masa mendatang. Oleh karena itu, diperlukan sebuah metode yang dapat memperkirakan harga minyak goreng. Salah satu upaya yang dilakukan yaitu dengan pendekatan metode Multivariate Time Series Clustering (MTSClust) berbasis model Vector Autoregressive with Exogenous (VARX). Metode MTSClust digunakan untuk memetakan keterkaitan pola deret waktu spasial antarwilayah, sedangkan model VARX digunakan karena mampu menangkap dinamika hubungan antarvariabel yang dipengaruhi faktor eksternal. Adapun penelitian ini bertujuan untuk mengkaji penerapan metode MTSClust berbasis model VARX dalam meramalkan harga minyak goreng di Indonesia di masa mendatang. Hasil pengklasteran menggunakan metode K-Means dengan ukuran jarak Chebyshev menunjukkan bahwa 77

Kabupaten dan Kota di Indonesia dapat dikelompokkan ke dalam tiga klaster dengan tingkat akurasi peramalan yang sangat baik berdasarkan kriteria MAPE yang berada di bawah 10%. Klaster 3 menjadi klaster terbaik dengan menggunakan model VARX (24,1) serta menunjukkan kinerja paling optimal dengan nilai MAPE sebesar 0,228%. Hasil kajian pada penelitian ini menegaskan bahwa pendekatan MTSClust berbasis model VARX efektif dalam meramalkan harga minyak goreng di masa mendatang.

Kata kunci: VARX, MTSClust, *K-Means* Chebyshev, *Forecasting*, Minyak Goreng

1. INTRODUCTION

Cooking oil is included as a strategic commodity in the basic needs of the Indonesian people, listed as part of the nine essential goods according to the Decree of the Minister of Industry and Trade Number 115 of 1998 [6]. Based on referring to Strategic Food Price Information Center figures, the market price of cooking oil shows significant fluctuations across regions in Indonesia [9]. These fluctuations not only affect the purchasing power of the community but also have the potential to drive food inflation, while also threatening the long-term stability of national food security [5][8]. As a strategic commodity in national food security, the resilience of cooking oil prices against fluctuations is crucial to maintaining the purchasing power of the community and supporting price control policies. Therefore, an accurate forecasting method is needed to predict the dynamics of cooking oil prices in the future.

The vast coverage of Indonesia's territory causes the characteristics of cooking oil prices between regions to be heterogeneous [8]. This heterogeneity is further complicated by spatial dependencies, where prices shocks in central market hubs can propragate and trigger a spillover effect of price volatility across other regions in Indonesia [17]. This condition makes the price forecasting process in each region less efficient. Therefore, an approach capable of grouping regions based on similar price movement patterns is needed, namely Multivariate Time Series Clustering (MTSClust) [10]. MTSClust is a clustering based forecasting method utilizing models designed for multivarited time series data, particulary the Vector Autoregressive (VAR) model [10]. Through this approach, regions exhibiting similar price patterns can be analyzed within one cluster, making the forecasting process more efficient and structured [10]. Furthermore, MTSClust plays a vital role in dimensionality reduction by grouping 77 regencies and cities into homogenous clusters, this reduces computational complexity and mitigates the risk of overfitting that might arise when from modeling each regions individually within a large multivariate system [10]. Specifically, the clustering process within the MTSClust in this study was executed using a non-hierarchical algorithm, namely K-Means with Chebyshev distance [10].

In the content of cooking oil prices forecasting, previous studies have incorporated Crude Palm Oil (CPO) prices as an exogenous variable using the ARIMAX models [5]. While these studies successfully captured the influence of CPO price volatility on cooking oil prices, their scope was limited to a univariate model, rendering them unable to model inter regional price dependencies and spatial spillover effects. Conversely, previous research has demonstrated that MTSClust method integrated with Vector Autoregressive (VAR) models is effective in generating high accuracy early warning systems for the prices of various food commodities [13]. However, conventional VAR models consider only relationships between endogenous variables and do not explicitly account for external market drivers [13]. Empirical evidence indicates that cooking oil prices fluctuations are heavily influenced by changes in CPO prices [19]. This relationship is crucial because CPO price volatility directly affects impacts price expectations for derivative products, including cooking oil

[16]. These limitations render the VAR models suboptimal for fully capturing the complexities of cooking oil price movements.

To address these limitations, this study advances existing research by proposing a MTSclust method integrated with the Vector Autoregressive with Exogenous Variables (VARX) model, thereby allowing external factor to be incorporated into the system [12]. Existing studies have generally limited to univariate ARIMAX models that include external variables or MTSClust and VAR model that focus solely on endogenous relationships. Therefore, the proposed MTSClust and VARX model bridges this gap by simultaneously capturing cross regional price dependencies and external market drivers, particularly CPO price volatility, in forecasting cooking oil prices. Specifically, this study aims to examine the application of the MTSClust based on VARX models in forecasting cooking oil prices in Indonesia.

2. METHODOLOGY

All information used in this research consists of cooking oil prices across 77 regencies and cities spread across 34 provinces in Indonesia are available in the Strategic Food Price Information Center. The sample covers each regency and city during the period of January 4, 2021, to September 26, 2025. The data consists of daily data covering three types of cooking oil, namely bulk cooking oil (BCO), packaged cooking oil brand 1 (PCO1), and packaged cooking oil brand 2 (PCO2). The cooking oil price data is used as an endogenous variable obtained through the Strategic Food Price Information Center [9]. Additionally, the prices of Crude Palm Oil (CPO), both domestic (CPO1) and international (CPO2), are used as exogenous variables sourced from the Indonesian Palm Oil Entrepreneurs Association [3].

2.1 DATA PRE-POCESSING

Data pre-processing is the initial stage aimed at ensuring data quality so that the resulting forecasting model becomes more accurate. A key challenge arises from incomplete observations, a critical concern in temporal data analysis since every data point is inherently linked over time [14]. The treatment of such gaps depends on the specific patterns and properties of missing values within each variable. For endogenous variables, missing entries are estimated using the Last Observation Carried Forward (LOCF) technique [21]. The LOCF method replaces missing data with the last available observation [21]. LOCF was chosen because cooking oil prices data shows a similar pattern and do not fluctuate highly. In contrast, exogenous Crude Palm Oil (CPO) prices data shows a highly fluctuating pattern. Therefore, the Vector Autoregressive Imputation Method Moving Average (VARIMMA) is applied, combining Exponential Moving Average (EMA) smoothing with Vector Autoregressive (VAR) modeling optimized through the Expectation-Maximization (EM) algorithm [14]. The EMA equation is written as follows:

$$y_t = \frac{\sum_{i=1}^{k}(1-\alpha)^i y_{t-i} + \sum_{i=1}^{k}(1-\alpha)^i y_{t+i}}{2\times(\sum_{i=1}^{k}(1-\alpha)^i)} \tag{2.1}$$

with y_{t-1} and y_{t+1} espectively representing the observation values before and after time t, k is the window length, and α is the smoothing constant [14]. The reconstructed values are then employed in a VAR-oriented iterative approach, which is stated as follows:

$$y_t = A_0 + \sum_{i=1}^{p} A_i y_{t-i} + u_t \tag{2.2}$$

With A_0 being the intercept vector, i being the time lag index, , A_i being the $k \times k$ coefficient matrix that shows the relationship between variables at lag i, y_t being the variable vector at time t, and u_t being the error vector at time t [10]. The stages in the VARIMMA method are presented in Figure 2.1.

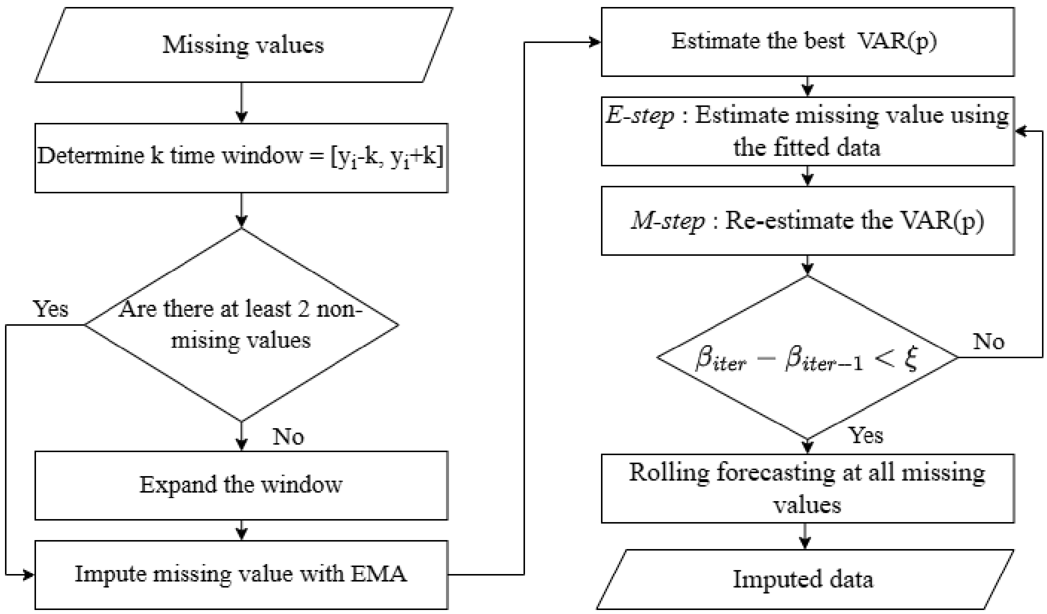

Figure 2.1 Flowchart method VARIMMA

The algorithm of VARIMA shown in Figure 2.1 is detailed as follows:

1. Define time window around the observation, where k is positive integer (k observations before and k observations after).
2. Imputation using Exponential Moving Average (EMA) if at least two observed values are present. Otherwise, dynamically expand the time window.
3. Fit the initially imputed data into a VAR(p) model to obtain the initial coefficient matrix $A^{(0)}$ and fitted values.
4. Expectation Step: Re-impute the missing data using the predicted values from the current VAR(p) model.
5. Maximization Step: Re-estimate the VAR(p) model parameters ($A^{(m)}$ using the update dataset from E-Step.
6. Repeat EM Step until the convergence criterion $\beta_{iter} - \beta_{iter-1} < \xi$ is satisfied.
7. Apply a rolling forecasting strategy using the convergent VAR(p) model to predict all consecutive missing values, producing the final imputed dataset.

2.2 VARX MODELING

A. STATIONARY TEST

Time series data must meet the stationarity assumption before the modeling process is carried out. A sequence of observations over time is classified as stationary when its central tendency and variability exhibit temporal invariance [7]. This property is commonly assessed through the Augmented Dickey-Fuller (ADF) procedure, designed to examine the existence of stochastic trends by testing for unit root behavior within the dataset. The ADF model is written as follows:

$$\Delta Y_t = \beta_0 + \phi Y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta Y_{t-i} + e_t \quad (2.3)$$

The ADF test statistic is expressed as follows:

$$ADF = \frac{\hat{\phi}}{SE(\hat{\phi})} \quad (2.4)$$

where ΔY_t is the first difference of the variable at time t, β_0 being model constant, ϕY_{t-1} being lag coefficient of Y_{t-1}, α_i being lag component of ΔY_t, p being lag length, t being observed time periode, e_t being error value, $\hat{\phi}$ least squares estimates of ϕ, and $SE(\hat{\phi})$ being standart error of ϕ. With the hypothesis $H_0 : \phi = 0$ (non-stationary data) and $H_1 : \phi < 0$ (data stationer). The null hypothesis is rejected Stationarity is concluded when the calculated ADF value lies beneath the Dickey–Fuller critical benchmark.

B. LAG

Lag represents the number of past periods used to predict the value of a variable in the current period [20]. Determination of the lag length is carried out using the Akaike Information Criterion (AIC), the most suitable model is determined based on the lowest AIC measure obtained [10]. The mathematical expression of AIC is given as follows:

$$AIC(\rho) = \ln|\sum(\rho)| + \frac{2}{T}\rho s^2 \quad (2.5)$$

with $\sum(\rho)$ denotes the covariance matrix associated with the model's residuals, whereas s signifies the number of endogenous components variables in the model, ρ is the lag length and T represents the total of obervations.

C. VARX MODELLING

The Vector Autoregressive with Exogenous Variables (VARX) model is a development of the Vector Autoregressive (VAR) model by including external variables into the system of equations. Exogenous variables come from outside the system and are not influenced by endogenous variables, but can influence the dynamics of the modeled endogenous variables [20]. The VARX the model takes the following form as follows:

$$y_t = \theta_0 + \sum_{i=1}^{p} \Phi_\mathrm{i}\ \mathrm{y}_{\mathrm{t}-1} + \sum_{j=0}^{q} \Theta_j\, x_{t-j} + u_t \quad (2.6)$$

with y_t states the endogenous variable, x_t represents the exogenous variable, p and q indicate the lag length of each variable, and u_t is the error vector.

E. STABILTY TEST

Stability testing is carried out to ensure that the VARX model that has been formed meets the stability assumptions and is suitable for use in the forecasting process [2]. Model stability is evaluated based on the roots of the characteristic polynomial which is stated as follows:

$$\det\left(I_k - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p\right) = 0 \quad (2.7)$$

Where I_k is the $k \times k$ identity matriks, Φ_p is the autoregressive coefficient matrix at lag p, p is the lag length, and z is the characteristic root. A model is considered stable if all characteristic roots satisfy $|z| \geq 1$ or, equivalently $|\lambda| \leq 1$ [1][2].

2.3 CLUSTERING STAGE

A. CLUSTERING

The parameter matrix yielded by the VARX specification modeling is used as the basis for the regional clustering process. In this study, clustering is crucial for identifying homogeneous spatial patterns and accounting for regional price dynamics. The clustering process is carried out using the Chebyshev K-Means method to group regions with similar cooking oil price dynamics. The K-Means method was chosen due to its efficiency in partioning objects based on data similarity [10]. Meanwhile, the Chebyshev distance was applied due to ist computational efficiency and robustness to extreme values in high dimensional datasets [22]. The Chebyshev distance is written as follows:

$$d_C(x, y) = \max_{i=1,\dots,p} |x_i - y_i| \tag{2.8}$$

where $d_C(x, y)$ is the Chebyshev distance between object x and object y, x_i is the value of objext x in the i-th variable, y_i is the value of object y in the i-th variable, and p is the number of variables.

The adequacy of the clustering solution is evaluated through the Silhouette Coefficient (SC), which is a measure that shows the level of suitability of an object to the cluster in which it is located compared to other clusters [4][15].

$$SC = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{2.9}$$

$$s(i) = \frac{b(i) - a(i)}{mak(a(i), b(i))} \tag{2.10}$$

$$a(i) = \frac{i}{|U| - 1} \sum_{m \epsilon U, m \neq i} d_{i,m} \tag{2.11}$$

$$b(i) = \min_{v \neq U} \frac{1}{|V|} \sum_{z \epsilon V} d_{i,z} \tag{2.12}$$

where $s(i)$ is the silhouette value of the i-th object, n is the number of objects in all clusters, $a(i)$ is the average distance between object i and all other objects in the same cluster, $b(i)$ is the average distance between object i and the nearest object in the other cluster, $d_{i,m}$ is the average distance between object i and object m, U is the cluster in which object i is located, V is the other cluster being compared, $|U|$ is the number of members in U, $|V|$ is the number of members in V.

B. PROFILING

Cluster profiling is carried out to describe the price movement patterns and characteristics of each cluster, which aims to provide a general overview of the data dynamics in each cluster [13]. This process is carried out using the Equation (2.13).

$$\bar{x}_{k,t} = \frac{1}{n_k} \sum_{i \epsilon C_k} x_{i,t} \tag{2.13}$$

where $x_{i,t}$ is the average value for cluster k at time – t, C_k is the set of members of cluster k, and n_k is the number of members in cluster k.

2.4 FORECASTING AND EVALUATION

Forecasting is performed at the cluster level, based on groups of regions with similar price movement patterns. This approach was chosen because cluster-level forecasting is considered more stable and representative than individual-level forecasting [13].

The evaluation of forecasting results aims to determine the accuracy level of a model in generating predicted values relative to actual data [13]. Model performance is evaluated using two error metrics, namely, Root Mean Squared Erorr (RMSE) and Mean Absolute Percentage Error (MAPE). The accuracy of predictions is assessed through RMSE and MAPE, in which reduced values signify better model performance [13][18]. These two metrics were chosen to provide a comprehensive evaluation, RMSE is sensitive to large forecasting error (extreme errors), while MAPE provides a relative measure in the form of a percentage that is independent of the data scale, thus allowing easy interpretation across variables [13].

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} \tag{2.14}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \bar{y}_t}{y_t}\right| \tag{2.15}$$

In the formulation, y stands for the true value at time t, $\bar{y}$ is the predicted counterpart, n denotes the number of observations, and t specifies the time step.

3. MAIN RESULTS

Prior to model estimation, the dataset underwent complete preprocessing, including the missing data handling specified in Section 2. After the data pre-processing stage, the data is divided using a sequential data validation method using a gradually expanding window scheme. This approach allows the model to be trained and tested repeatedly on several time-based data subsets, resulting in a more stable evaluation that effectively minimizes prediction bias and preserves the chronological consistency of time series data [11].

As a result of this scheme, the dataset was partitioned into 25 folds. In the first fold, the training data consists of the initial 352 observations to obtain a stable model. In the subsequent folds, the training data is gradually expanded by adding 55 observations from the next period, while the last 55 observations are used as test data. Thus, the size of the training data gradually increases until it reaches 1,672 observations in the 25th fold. The data partitioning scheme through a dynamic cross-validation process with an expanding temporal
window is presented in Figure 2.2.

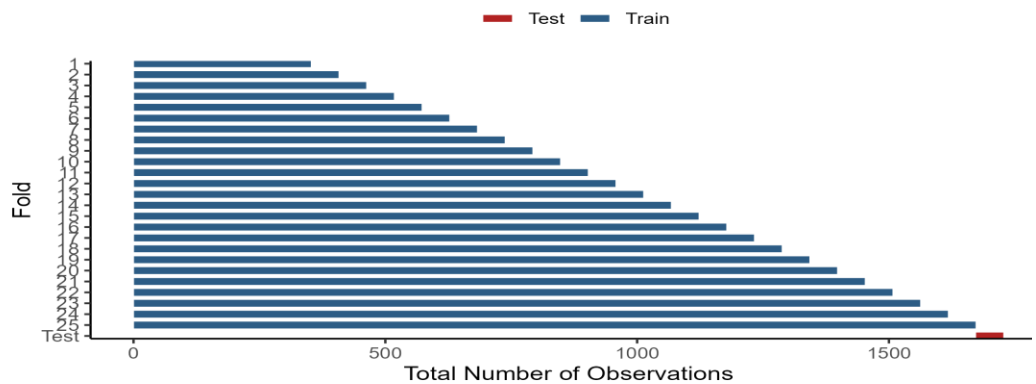

Figure 3.1 Scheme Time Series Cross Validation with Expanding Window

As presented in Figure 3.1, in the first fold, the training data starts from the first observation to the 352nd observation. In the second fold, the training data is increased by 55 observations, so the training data starts from the first observation to the 407th observation. Up to the 25th fold, the training data starts from the 1st observation to the 1,672nd observation. The test data is set at the last 55 observations and is not used in the cross-validation process.

3.1 MODELING STAGES

The stages of the Vector Autoregressive with Exogenous Variables (VARX) include stationarity tests, optimal lag selection, model estimation, and stability tests.

A. STATIONARITY TEST

Stationarity testing is conducted on each variable and fold with a significance level of $\alpha = 0{,}05$. For example, the unit root test for the price series of bulk cooking oil in Cirebon City on the first fold produces the following adf:

$$ADF = \frac{0{,}028}{0.05} = 0{,}56$$

The initial test results indicate that the data still contain a unit root, thus not meeting the stationarity assumption, a condition commonly found in fluctuating daily price data. Therefore, first order differencing ($d = 1$), was applied to remove the trend and make the data stationary. The results of the retesting show that all variables in each fold have met the stationarity assumption with a $p-value < \alpha$, as presented in Figure 2.3.

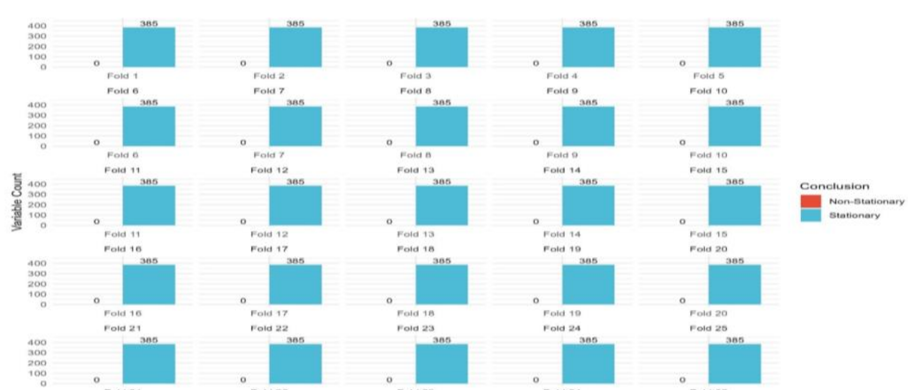

Figure 3.2 Stationary test result

Figure 3.2 shows that across all 25 cross-validation folds, all variables achieved stationarity after the first differentiation transformation. No non-stationary variables remained. This consistent result confirms that the first differentiation approach successfully removed unit roots across all splitting iterations, ensuring that all time series inputs were suitable for subsequent forecasting modeling.

B. DETERMINATION OF OPTIMAL LAG

Equation (2.5) is used to calculate the AIC value for each region and each fold. The evaluation results for the optimal lag choice are displayed in the figure 2.4.

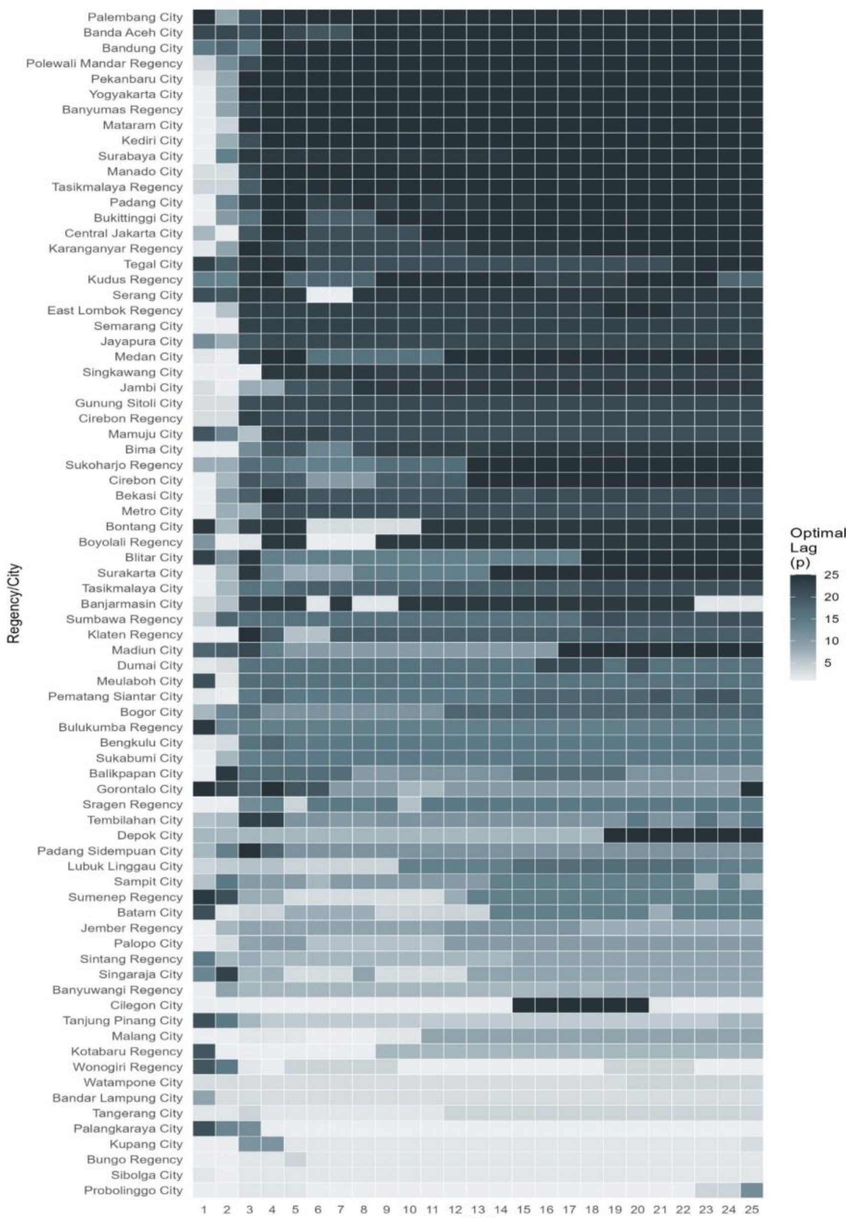

Figure 3.3 Optimal Lag for each fold and region

Figure 3.3 shows the optimal lag selection using AIC for various districts/cities and 25-fold cross-validation. The heatmap highlights spatial heterogeneity and dynamic temporal shifts in lag selection. Specifically, regions located at the top of the grid (such as Palembang City and Banda Aceh City) consistently require higher lag lengths, represented by darker colors. Conversely, regions at the bottom (such as Probolinggo City and Sibolga City) generally require shorter lag lengths, represented by lighter colors. This spatial variation underscores the importance of using region-specific VAR lag selection rather than applying a single, uniform lag length across all geographic nodes.

C. VARX MODELING

Modeling was carried out in 77 districts/cities with three endogenous variables and two exogenous variables. At each fold, the modeling process produces 231 VARX models. The VARX(1,1) modeling results for Cirebon City on the first fold are presented as a representative illustrative example due to its parsimonious lag structure, which provides a clear and highly interpretable demonstration of the model coefficients and dynamic interactions. The model is written as follows:

$$BCO = 16{,}7062 + 0{,}0043BCO_{t-1} + 0{,}0345PCO1_{t-1} - 0{,}0445PCO2_{t-1}$$
$$-0{,}0300CPO1_{t-1} + 0{,}0101CPO2_{t-1}$$

$$PCO1 = 11{,}9044 + 0{,}2503BCO_{t-1} + 0{,}1319PCO1_{t-1} - 0{,}2439PCO2_{t-1}$$
$$-0{,}0178CPO1_{t-1} - 0{,}0075CPO2_{t-1}$$

$$PCO2 = 12{,}7951 + 0{,}1718BCO_{t-1} + 0{,}1655PCO1_{t-1} - 0{,}1919PCO2_{t-1}$$
$$-0{,}0169CPO1_{t-1} + 0{,}0090CPO2_{t-1}$$

Overall, 5,775 VARX models were estimated, generating 286,470 coefficients that serve as the feature set for regional clustering. Substantively, these parameters capture both the time-lag dynamics of local price adjustments and the sensitivity of regional markets to exogenous CPO price shocks, thereby providing a rich quantitative foundation to group regions with similar dynamic behavior.

D. STABILITY TEST

Based on calculations using equation (2.7) for the VARX model in Cirebon City for the first fold, the characteristic root values obtained are $|\lambda_1| = 0{,}115$, $|\lambda_2| = 0{,}115$, $|\lambda_3| = 0{,}042$. With a maximum modulus of lambda $|\lambda_{max}| = 0{,}115$ is strictly less than 1, all roots lie inside the unit circle, indicating that the VARX model meets the stability conditions.

Using the same calculations, the stability test was applied to the entire cross-validation scheme, evaluating a total of 5,775 VARX models. Because each model generates multiple eigenvalues, reporting all lambda values would result in an excessively large number of parameters. Therefore, Table 3.1 presents a summary displaying the highest maximum modulus recorded across all 25 folds for each regency/city.

Table 3.1 Stability Test Result

Fold	Number	Regencies and Cities	Model	$\|\lambda_{max}\|$	Stability Criteria
	1	Banda Aceh City	VARX (22,1)	0,991	Stable
	2	Meulaboh City	VARX (21,1)	0,972	Stable
1	3	Medan City	VARX (2,1)	0,440	Stable
	4	Sibolga City	VARX (2,1)	0,446	Stable
	5	Pematang Siantar City	VARX (2,1)	0,474	Stable

⋮	⋮	⋮	⋮	⋮	⋮
	5.770	Palopo City	VARX (10,1)	0,829	Stable
	5.771	Watampone City	VARX (4,1)	0,644	Stable
	5.772	Manado City	VARX (25,1)	0,939	Stable
25	5.773	Mamuju City	VARX (21,1)	0,947	Stable
	5.774	Polewali Mandar Regency	VARX (25,1)	0,975	Stable
	5.775	Jayapura City	VARX (22,1)	0,956	Stable

Table 3.1 summarizes the optimal VARX model specifications and their corresponding maximum characteristic roots (λ_{max}) for all 77 regencies/cities. All estimated models satisfy the stability condition with $|\lambda_{max}| < 1$, confirming dynamic system stability across all regions. The results reveal notable spatial heterogeneity, ranging from low-order parsimonious models to higher-order lag structures to capture longer price persistence.

3.2 MTSCLUST STAGES

The stages in Multivariate Time Series Clustering (MTSClust) begin with cluster formation, followed by cluster profiling, forecasting and evaluation of forecasting results.

A. CLUSTERING

The desired number of clusters is obtained by considering the maximum score of the Silhouette Coefficient (SC) value calculated using Equation (2.9) based on the K-Means algorithm with the Chebyshev distance metric in Equation (2.8). The calculation of the average SC value is as follows:

$$SC = \frac{1}{77}\sum_{i=1}^{77}(0{,}4888 + 0{,}2975 + \cdots + 0{,}2339 + 0{,}1769)$$

$$SC = \frac{24{,}7614}{77} = 0{,}3215$$

Based on this calculation, the SC value is 0,3215, a result also shown in Figure 3.4.

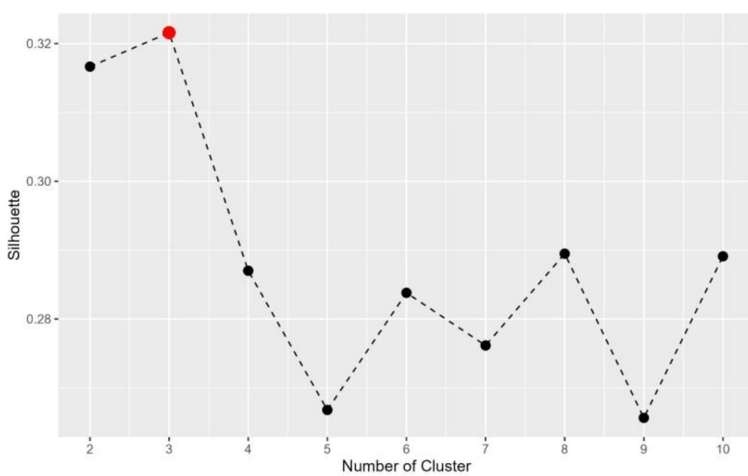

Figure 3.4 Silhouette Optimal Cluster Plot

In Figure 3.4, the red point shows the highest Silhouette Coefficient value obtained in the K-Means method with the Chebyshev distance measure. This point reaches its peak at $k = 3$. The clustering results produced 3 clusters with different characteristics of cooking oil price movement patterns. Cluster 1 contains 28 regencies/cities, followed by 23 in Cluster 2 and 26 in Cluster 3. Examples of cluster 1 members are Banyumas Regency, Bogor City, and Manado City, cluster 2 includes Cirebon Regency, Surabaya City, and Yogyakarta City, while cluster 3 includes Balikpapan City, Malang City, and Palembang City. Regional grouping shows differences in cooking oil price movement patterns that align with exogenous Crude Palm Oil (CPO) price dynamics. The clustering results map is presented in Figure 3.5.

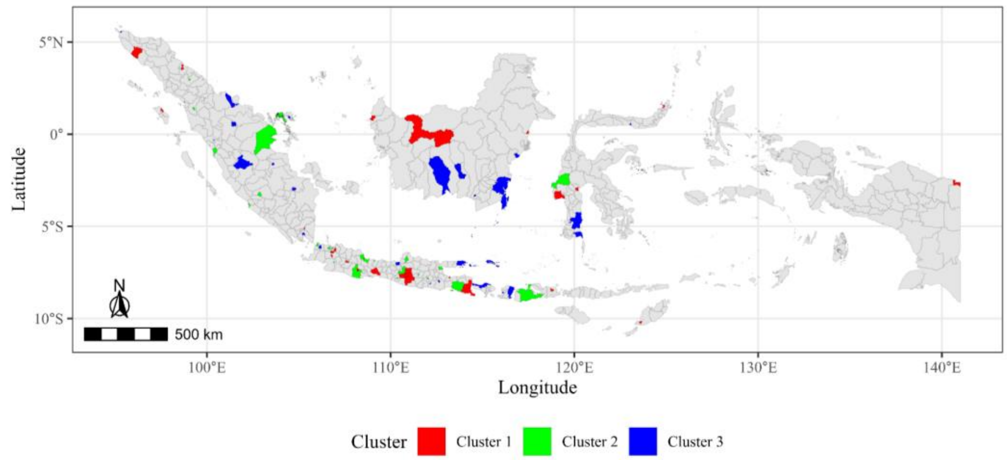

Figure 3.5 Cluster Result Map

Figure 3.5 shows that each cluster is distributed across various islands, indicating differences in regional price sensitivity to CPO price dynamics.

B. PROFILING

This cluster profiling process is carried out by estimating the mean value of each variable across clusters for each period in accordance with Equation (2.13), thereby obtaining the cluster characteristics shown in Figure 3.6.

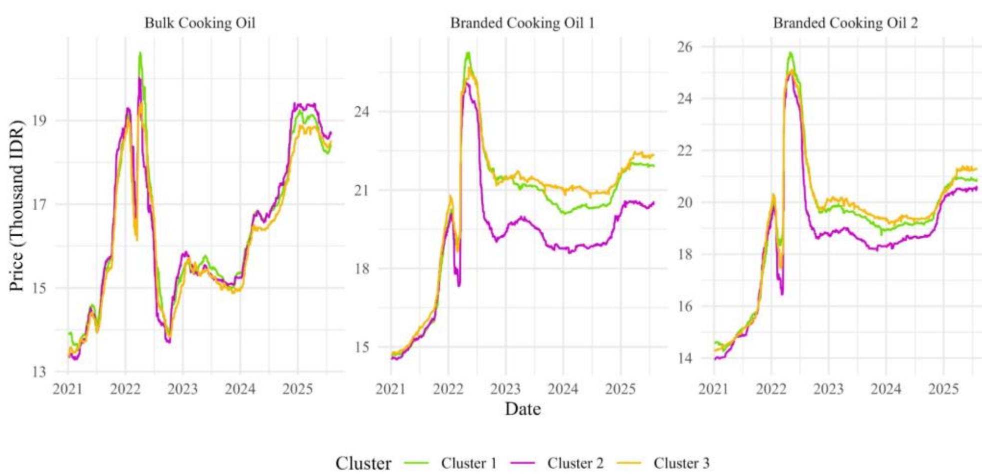

Figure 3.6 Cluster Profiling

The cluster profiles show that the pattern of cooking oil price movements across clusters is relatively uniform, characterized by a sharp spike during the 2021–2022 period, a decline in 2022–2023, and a gradual increase toward the end of the observation period. However, different trajectory patterns emerge in certain product categories. For packaged cooking oil, Brand 1 in Clusters 1 and 3 displays a highly synchronized upward trend, while Cluster 2 maintains a much lower base price level. For Brand 2, Cluster 2 shows higher price volatility during the peak surge period, while Cluster 3 shows a faster price stabilization after 2022. Meanwhile, bulk cooking oil in Cluster 3 shows relatively higher prices toward the end of the period. This suggests that cluster differences lie not only in the base price level, but also in the differential response to exogenous CPO price dynamics within the VARX model.

C. FORECASTING

After the clustering phase, each cluster was re-estimated using the VARX model. The optimal model for each cluster was selected based on the lowest Akaike Information Criterion (AIC) value. This process resulted in three representative VARX models for each cluster presented in Table 3.2.

Table 3.2 Model VARX formed

Cluster	Model Formed
1	VARX(25,1)
2	VARX(24,1)
3	VARX(24,1)

As presented in Table 3.2, while Cluster 1 exhibits a different lag structure, Clusters 2 and 3 have the same lag order. Despite sharing the same lag specification, Clusters 2 and 3 differ significantly in their internal market dynamics. These differences are reflected in variations in coefficient magnitude, statistical significance, and transmission response to CPO price dynamics.

D. EVALUATION OF FORECASTING RESULTS

Equations (2.14) and (2.15) are used to measure the accuracy of the forecast results for each cluster. The evaluation results are presented in Table 3.3.

Table 3.3 Evalution of forecasting result

Cluster	Model Formed	Forecast Accuracy	
		RMSE	MAPE
1	VARX (25,1)	362,5368	1,5%
2	VARX (24,1)	1.119,5706	4,43%
3	VARX (24,1)	65,2610	0,228%

Overall, the forecasting results track the actual data movement patterns exceptionally well across all three cooking oil types, supported by low MAPE values below 10% for all clusters. Cluster 3 with the VARX (24,1) model showed the best output with the smallest RMSE also MAPE values, indicating the highest level of forecasting accuracy compared to the other clusters. The comparison plots between the actual data and the forecasting finding for each cluster are presented in Figure 3.7, Figure 3.8, and Figure 3.9.

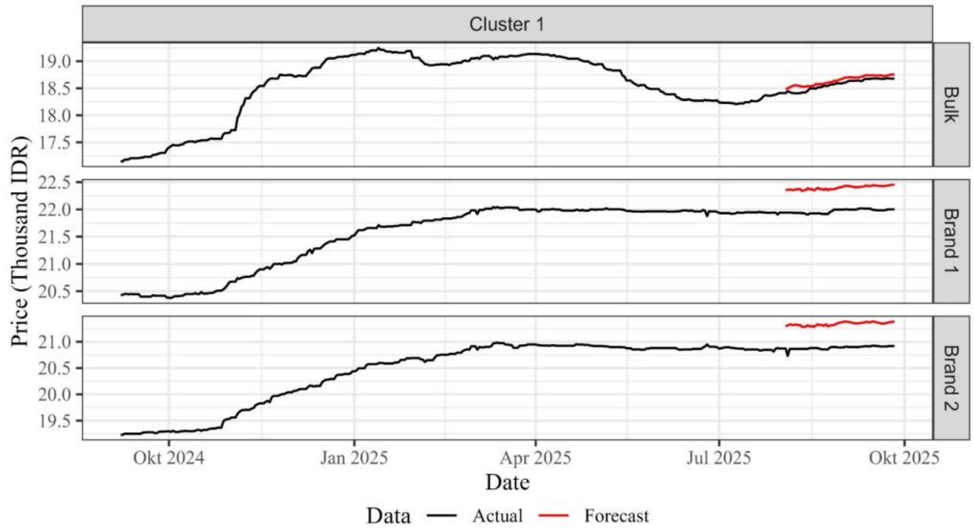

Figure 3.7 Comparison of Actual Data and Cluster Forecast Results 1

Figure 3.7 shows that the predicted values for Cluster 1 consistently follow the trend of the actual data. The VARX(25,1) model demonstrates high accuracy, supported by low error metrics, namely RMSE of 362.54 and MAPE of 1.5%, so its performance is very accurate.

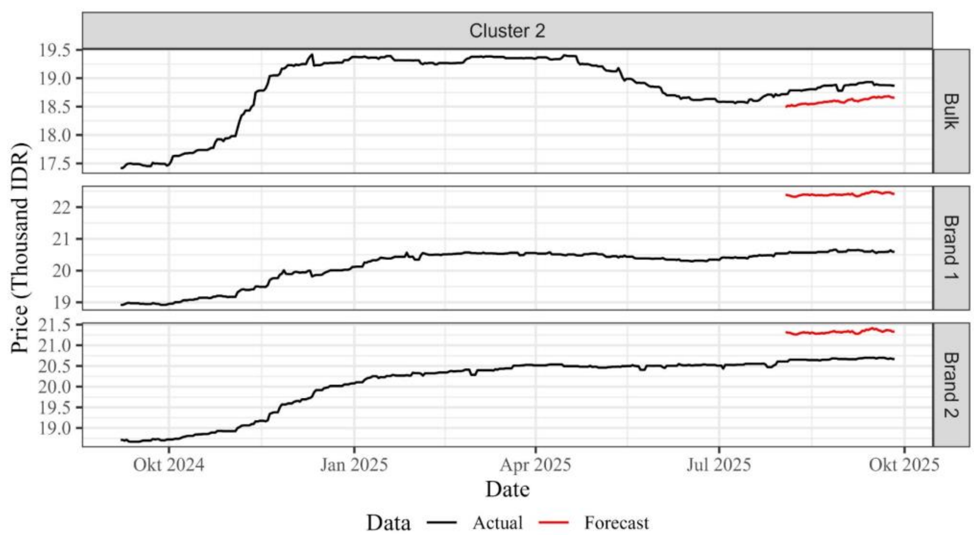

Figure 3.8 Comparison of Actual Data and Cluster Forecast Results 2

Figure 3.8 shows that the predicted values for Cluster 2 closely align with the actual data movement patterns. The VARX(24,1) model achieves highly accurate performance, as evidenced by the RMSE value of 1,119.57 and MAPE of 4.43%.

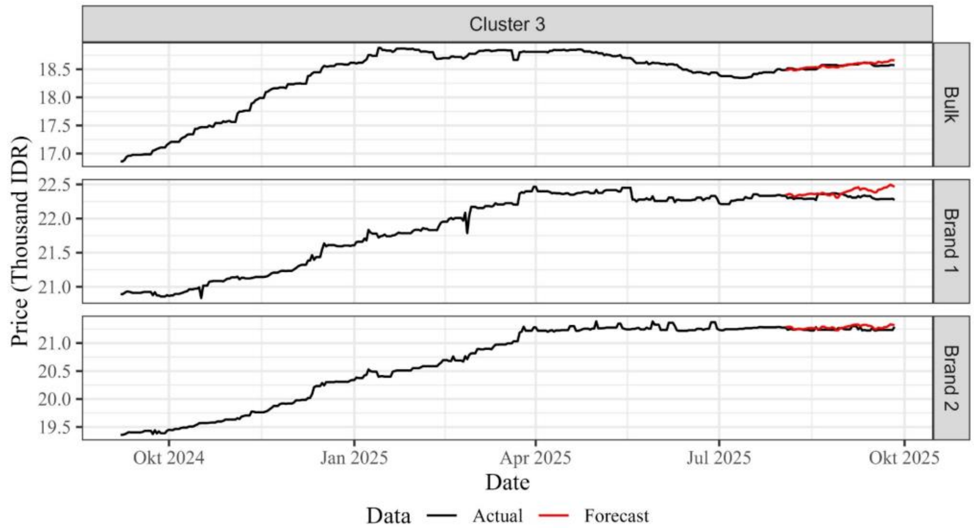

Figure 3.9 Comparison of Actual Data and Cluster Forecast Results 3

Figure 3.9 shows that the forecasting results for Cluster 3 closely match the actual data pattern. Compared to other clusters, the VARX(24,1) model for this cluster recorded the highest performance, with the lowest RMSE of 65.26 and MAPE of 0.228%, indicating that this cluster is very accurate.

In general, the forecasting results closely followed the actual data movement patterns across all clusters. The closeness between the actual and forecasted values indicates that the model performs adequately in predicting future cooking oil prices.

4. CONCLUSION

The results of this research study indicate that based on Multivariate Time Series Clustering (MTSClust) using K-Means algorithm with Chebyshev ditance of 77 Regencies and Cities, 3 clusters were formed, with the best cluster being cluster 3 which uses the VARX (24,1) model. Cluster 3 shows the highest level of forecasting accuracy with a MAPE value of 0,228%. In Cluster 3, the VARX model accurately mapped the price dynamics of all types of cooking oil. This was demonstrated by a relatively higher price increase for bulk cooking oil at the end of the observation period, as well as a stable price movement pattern for Packaged Cooking Oil Brands 1 and 2. The evidence indicates that the VARX model embedded in the MTSClust method approach is able to represent the pattern of cooking oil price movements well and produce accurate forecasts.

BIBLIOGRAPHY

1. Aidil, & Sulistijanti, W., 2025. Penerapan Model Vector Autoregressive (VAR) Untuk Peramalan Produksi Kakao di Provinsi Sulawesi Tengah. *JUSTEK : Jurnal Sains Dan Teknologi*, Vol. 8, No. 2, 181–191.
2. Amelia, R., Wahyuni, D., & Julianti, E., 2024. Implementation of Vector Autoregressive (VAR) and Vector Error Correction Model (VECM) Method in Pneumonia Patients with Weather Elements in Pangkalpinang City. *Barekeng: Journal of Mathematics and Its Applications*, Vol. 18, No. 3, 1447–1458.

3. Gabungan Pengusaha Kelapa Sawit Indonesia (GAPKI), 2025. *Posisi Harga Komoditas*. https://gapki.id/posisi-harga-komoditas/ [10 September 2025]
4. Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E., 2025. Cluster validity indices for automatic clustering: A comprehensive review. *Heliyon*, Vol. 11, No 2.
5. Ilmani, E. A., Amatullah, F. F., Notodiputro, K. A., Angraini, Y., & Mualifah, L. N. A., 2025. The Performace of The Arimax Model on Cooking Oil Price Data. *Barekeng: Journal of Mathematics and Its Applications*, Vol. 19, No. 2, 819–828.
6. Keputusan Menteri Perindustrian dan Perdagangan Nomor 115/MPP/Kep/2/1998 tentang Jenis Barang Kebutuhan Pokok Masyarakat.
7. Lestari, L., Sulistianingsih, E., & Perdana, H., 2024. Vector Autoregressive with Outlier Detection on Rainfall and Wind Speed Data. *Barekeng : Journal of Mathematics and Its Applications*, Vol. 18, No. 1, 117–128.
8. Nendissa, D. R., Pellokila, M. R., Lerik, M. D. C., Sa'dyah, A. A., Kana, Y. R., Nahas, A. E., Harini, T. S., Kapa, M. J., & Abidin, Z., 2025. Cooking Oil Price Volatility in the Consumer Market and Wholesalers Market in Indonesia. *Journal of Global Innovations in Agricultural Sciences*, Vol. 13, No. 1, 61–70.
9. Pusat Informasi Harga Pangan Strategis (PIHPS), (2025). *Informasi Harga Pangan Antar Daerah*. https://www.bi.go.id/hargapangan/ [18 September 2025]
10. Rohaeti, E., Sumertajaya, I. M., Wigena, A. H., & Sadik, K., 2023. MTSClust with Handling Missing Data Using VAR-Moving Average Imputation. *Mathematics and Statistics*, Vol. 11 No. 2, 229–244.
11. Sucipto, C. S., Sulandari, W., & Susanti, Y., 2025. Implementation of Cross-Validation of Hang Seng Index Forecasting Using Holt's Exponential Smoothing and Auto-Arima Method. *Barekeng: Journal Of Mathematics and Its Applications*, Vol. 19, No. 1, 13–24.
12. Sudarwanto, S., Puteri, S. M., Santi, V. M., & Alwansyah, M. A., 2026. Forecasting Indonesia's Export Revenue through a Vector Autoregressive Exogenous Approach. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, Vol. 10, No. 2, 396-411.
13. Sumertajaya, I. M., Rohaeti, E., Fitrianto, A., & Adi, W. P. P., 2024. Development of Food Commodity Price Forecasting Model as an Early Warning System with a Multivariate Time Series Clustering. *Sains Malaysiana*, Vol. 53, No. 11, 3779–3789.
14. Sumertajaya, I. M., Rohaeti, E., Wigena, A. H., & Sadik, K., 2023. Vector Autoregressive - Moving Average imputation Algorithm for Handling Missing Data in Multivariate Time Series. *IAENG International Journal of Computer Science*, Vol. 50, No. 2, 727–735.
15. Suraya, G. R., & Wijayanto, A. W., 2022. Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia According to the Special Index for Handling Stunting. *ISJA: Indonesian Journal of Statistics and Its Applications*, Vol. 6, No. 2, 180–201.
16. Tardini, G. A., & Suharjito., 2024. Selection of Modelling for Forecasting Crude Palm Oil Prices Using Deep Learning (GRU & LSTM). *Emerging Science Journal*, Vol. 8, No. 3, 875–898.
17. Theresia, A., Ikhsan, M., Kacaribu, F. N., & Sumarto, S., 2025. Spillover Effect of Food Producer Price Volatility in Indonesia. *Economies*, Vol. 13, No. 9, 256.

18. Tsabihah, D. U., Anggraini, Y., & Sumertajaya, I. M., 2025. Implementation of Time Series Clustering with DTW to Clustering and Forecasting Rice Prices Each Provinces in Indonesia. *INFERENSI*, Vol. 8, No. 1, 2721–3862.
19. Ulama, B. S. S., Hibatullah, F., Habibi, M. R., & Makkulau., 2024. Comparative Analysis of ANFIS and State-ANFIS for Forecasting Cooking Oil Prices Based on Processed Palm Oil Yield (Crude Palm Oil). *Applied Computational Intelligence and Soft Computing*, Vol. 2024, 1-16.
20. Wei, W. W. S., 2019. *Multivariate Time Series Analysis and Applications*, First Edition. John Wiley & Sons Inc. Hoboken, USA.
21. Wongoutong, C., 2025. Performance Comparison of Various Imputation Methods for Missing Data Mechanisms (MAR, MCAR, and MNAR) in a Nonstationary Time Series. *International Journal of Mathematics and Mathematical Sciences*, Vol. 2025, 1–16.
22. Zamri, N., Bakar, N. A. A., Aziz, A. Z. A., Madi, E. N., Ramli, R. A., Sukono, Koon, C. S., & Marhadi., 2026. A Conceptual of Merging of Intuitionistic Fuzzy C-Means with Chebyshev for Genomic Clustering Solutions Addressing Cancer Issues. *Malaysian Journal of Mathematical Sciences*, Vol. 20, No. 1, 395–421.