Scientific Journal of Informatics
Vol. 12, No. 1, Feb 2025

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Rice Price Forecasting for All Provinces in Indonesia Using The Time Series Clustering Approach and Ensemble Empirical Mode Decomposition

Erdanisa Aghnia Ilmani[1*], I Made Sumertajaya[2], Anwar Fitrianto[3]
[1, 2, 3]Department of Statistics and Data Science, IPB University, Indonesia

Abstract.
Purpose: Accurate forecasting of rice prices is essential to ensure food security and a healthy economy for a country like Indonesia. Problems regarding time-series phenomena, such as trends or seasonality, are problematic for traditional approaches like ARIMA (Autoregressive Integrated Moving Average). This study analyzes the effect of EEMD (Ensemble Empirical Mode Decomposition) combined with time-series data clustering on forecasting accuracy.
Methods: From 2009 until 2023, the thirty-two Indonesian provincial rice prices were grouped monthly into time-series clusters using hierarchical clustering, average linkage, and DTW (Dynamic Time Warping). After clusterization, the time series were decomposed using the ensemble EEMD method to extract their IMFs (Intrinsic Mode Functions) and residual components. Each IMF was assigned an ARIMA model. The model forecast was generated by adding all individual estimates. MAPE (Mean Absolute Percentage Error) was used to measure the model's performance.
Result: The prices were divided into three clusters with an optimized region. Price changes are well captured through EEMD, where the residual components contributed predominantly to the long-term trends. The validation of the prediction showed MAPE values under 10% for the majority of the provinces, which indicates a relatively accurate prediction. On the other hand, some regions had inaccuracies that were higher than others due to uncontrollable fluctuations.
Novelty: This study integrates clustering with EEMD decomposition for monthly rice price forecasting using data from 32 Indonesian provinces from 2009 - 2023, offering a novel approach that improves traditional techniques. The model can capture distinct regional price patterns and provide essential information to policymakers to manage rice supply and price stabilization. Further studies can develop external hybrid models with economic variables.

Keywords: Time series clustering, Rice price forecasting, EEMD, DTW, ARIMA
Received April 2025 / **Revised** May 2025 / **Accepted** May 2025

INTRODUCTION

The analysis of time series assists us in making sense of elaborate phenomena like trends, seasonality, and random variations. With the growing need for forecasting in multiple domains, applying sophisticated statistical models becomes imperative. One of the most tried and tested models designed for such needs is the Autoregressive Integrated Moving Average (ARIMA) model. As its name suggests, ARIMA is a non-stationary time series model. Unlike its predecessor, the Autoregressive Moving Average (ARMA) model, which solely caters to stationary data, ARIMA takes care of this shortcoming by differencing the time series to render it stationary, then applying the ARMA model for analysis. Through this, ARIMA describes the underlying attributes of non-stationary time series data while enabling real-world conditions and patterns, thus allowing for reliable predictions of trends and values [1]. However, when dealing with large and complex data, time series analysis becomes more challenging. Data often needs to be clustered first to improve forecasting efficiency and accuracy. Cluster analysis can explore helpful information in the data, such as anomaly detection, hidden features of the data, and similarities between observations, all of which are crucial for improving predictive models [2], [3]. In time series clustering, data can be grouped based on similarities in patterns, and several approaches.

Several studies have explored clustering methods for different types of data. In the context of rice price data, a data characteristic approach was more efficient for ARIMA modeling and forecasting. In another

[*]Corresponding author.
Email addresses: erdanisaaghnia@apps.ipb.ac.id (Ilmani)[*], imsjaya@apps.ipb.ac.id, (Sumertajaya), anwarstat@gmail.com (Fitrianto)
DOI: 10.15294/sji.v12i1.23536

study, clustering rainfall time series data using the Piccolo method proved more accurate than the Maharaj distance method [4] Clustering of cooking oil prices in Indonesia based on raw data was conducted using the CID distance measurements [5]. Additionally, the accuracy of cluster-level ARIMA modeling for rice price data in the western part of Indonesia was found to be higher compared to the provincial level. The text also discusses two main approaches in cluster analysis: hierarchical and non-hierarchical. The hierarchical method involves grouping data based on similarity, resulting in a tree or dendrogram. Different linkage methods, such as single linkage, complete linkage, average linkage, ward method, and centroid linkage, are used in the hierarchical cluster analysis. Studies like those by [6] and [7] have demonstrated the success of average linkage to produce the best results for clustering COVID-19 cases and rice prices in Indonesia. Finally, clustering provinces in Indonesia based on non-oil and gas exports using Dynamic Time Warping (DTW) and Euclidean distance showed high cophenetic correlation values. Regardless of these developments, a persisting gap exists about how intricately structured time series data with its trends, seasonality, random noise, and other intricate details is handled due to the misconception that it often impairs prediction accuracy. Standalone ARIMA may not optimally differentiate these elements; therefore, decomposition methods must be added to enhance accuracy in forecasting.

Enhancement of EMD, known as Ensemble Empirical Mode Decomposition (EEMD), provides practical methods for partitioning time series data into Intrinsic Mode Functions (IMFs) for subsequent modeling and forecasting. While EMD improves some aspects of traditional methods, it struggles with mode mixing, leading to overlapping scale IMF components [8]. Adding white noise to the data, as outlined in the development of EEMD, helps to solve some of these concerns more efficiently by having a more reliable decomposition process [9]. EEMD-based hybrid ensemble forecasting has been shown to achieve high accuracy in predicting the price of rice, especially in Jakarta, where weekly price forecasts tend to be more accurate [10], [11].

The majority of Indonesia's population considers rice their primary food source, making it an integral part of the country's economic and social structure. Ensuring a sufficient rice supply is a primary concern, especially considering food security and the country's large population. Although rice is often regarded as a relatively stable commodity, its price movements are influenced by various interrelated factors. Differences in harvest times between regions, limited storage and distribution infrastructure, changes in import-export policies, and unpredictable climatic conditions all contribute to the instability of rice supply, which ultimately affects price fluctuations in the market. The Indonesian economy is significantly impacted when rice supply decreases, particularly during the off-season, which significantly affects low-income groups highly dependent on rice as a staple food. While these factors may not always cause direct price surges, their cumulative impact can disrupt market stability, especially for communities that heavily rely on rice for their daily food needs. A decline in rice supply can lead to price surges, which may seriously affect economic stability [12]. The past few years have shown a frequent fluctuation in rice prices, with some estimates showing that 179 regions in Indonesia experience some of the sharpest price increases available in the market [13]. The demand reaction to rice prices is somewhat inelastic, implying that varying prices do not significantly impact consumer demand; nonetheless, supply disruptions can lead to significant increases in price, which can have negative consequences on overall economic stability and public welfare [14]. This price uncertainty highlights the importance of having a reliable forecasting tool. By predicting potential price surges or declines, stakeholders can design more accurate strategies to manage rice's availability, distribution, and affordability. For instance, policymakers need accurate predictive information to determine the timing for interventions, such as adjusting import quotas or releasing government rice reserves. On the other hand, farmers and distributors will significantly benefit from price projections when planning production and logistics Therefore, accurate rice price forecasting is crucial for stabilizing the market, supporting food policies, and ensuring national food security.

While various studies have integrated clustering with ARIMA for forecasting time series data, the effectiveness of EEMD as a decomposition method for rice price forecasting remains underexplored. Therefore, this study aims to evaluate the effectiveness of the EEMD method when combined with ARIMA for forecasting rice prices in Indonesia. This research aims to determine the method that provides the highest prediction accuracy, ultimately contributing to more informed policy-making and better management of rice supply in Indonesia. This study is expected to demonstrate the performance of EEMD and assess its suitability for improving rice price forecasting, which will ultimately support achieving food security and national economic stability.

METHODS

Data

This research uses the time series data of rice prices every month from January 2009 until December 2023 for 32 provinces in Indonesia. The data were obtained from the publication 'Rural Consumer Price Statistics: Food Group' by the Central Statistics Agency (BPS) [15]. The price of rice was obtained from three to four traders per market at the level of sub-district, and after averaging at the district level, the figures were aggregated at the province level. However, data for North Kalimantan and Jakarta's provinces were unavailable for this period. All analyses were done using R studio. Since the dataset was already clean and well-structured, no additional preprocessing steps were necessary. The data was directly used for clustering analysis without modification, ensuring that the data was ready for modeling and minimizing any potential bias introduced by preprocessing steps.

Flowchart

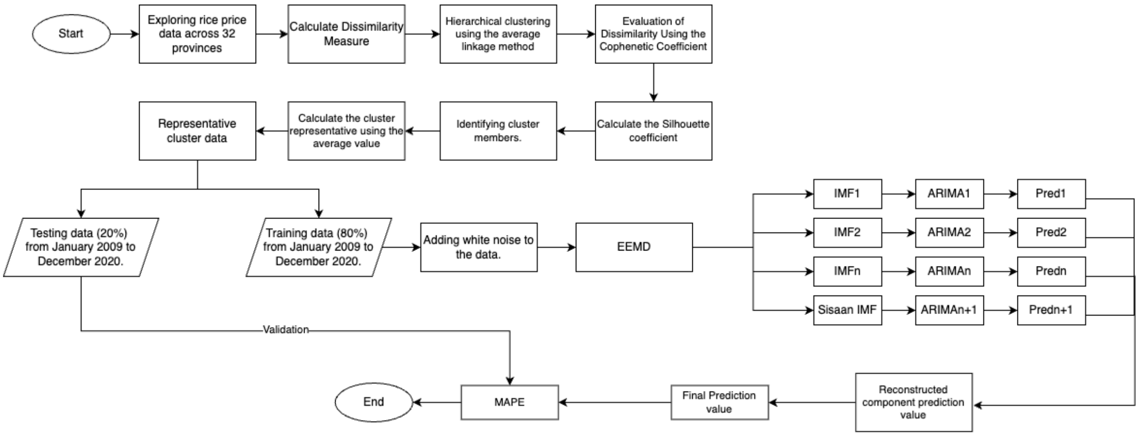
Start
Exploring rice price data across 32 provinces
Calculate Dissimilarity Measure
Hierarchical clustering using the average linkage method
Evaluation of Dissimilarity Using the Cophenetic Coefficient
Calculate the Silhouette coefficient
Identifying cluster members.
Calculate the cluster representative using the average value
Representative cluster data
Testing data (20%) from January 2009 to December 2020.
Training data (80%) from January 2009 to December 2020.
Adding white noise to the data.
EEMD
IMF1
IMF2
IMFn
Sisaan IMF
ARIMA1
ARIMA2
ARIMAn
ARIMAn+1
Pred1
Pred2
Predn
Predn+1
Validation
Reconstructed component prediction value
Final Prediction value
MAPE
End

Figure 1. Flowchart of hierarchical time series clustering and forecasting using EEMD-ARIMA

The data is divided into two parts: training data and testing data. Training data spans from January 2020 to December 2020, while testing data spans from January 2019 to December 2020. The training set is used to build the model, and the testing set is used for validation.

Hierarchical clustering

As Kaufman and Rousseeuw (1990) proposed, agglomerative hierarchical clustering begins with each object as its cluster. It progressively merges similar objects into the same cluster until all objects are grouped into one cluster [16]. This method uses linkage techniques to measure dissimilarity between clusters during the merging process. In this study, the **average linkage** method is employed, where the dissimilarity between two clusters is based on the average distance between all objects in one cluster and all objects in the other. The formula for average linkage is:

$$d_{(xy)z} = average\{d_{xz}, d_{yz}\} = \frac{d_{xz} + d_{yz}}{n_{(xy)}n_z}$$

where $n_{(xy)}$ is the number of objects in cluster (xy), and n_z is the number of objects in cluster z [17].

Similarity measure

The similarity measure quantifies the resemblance between two-time series, and Dynamic Time Warping (DTW) is a model-free approach used to measure this to a great extent. Often utilized for aligning two-time series by applying them in speech recognition, it minimizes the distance between them and is commonly used with DTW. [18]. The DTW distance is defined as:

$$\text{DTW}\ (S, T) = \min_{W}\left[\sum_{k=1}^{p} \delta(w_k)\right]$$

where S and T are time series, and $\delta(i,j) = |s_i - t_j|$ represents the distance between points on the alignment path [19].

Similarity measure evaluation

The effectiveness of a similarity measure in clustering can be evaluated using the cophenetic correlation coefficient. After clustering, a similarity matrix D and a dendrogram are generated, with the cophenetic distances *v(i,j)* representing the height of the lines in the dendrogram. The cophenetic correlation coefficient c is calculated as:

$$c = \frac{\sum_{i<j}\ (d(i,j) - \bar{d})(v(i,j) - \bar{v})}{\sqrt{\left[\sum_{i<j}\ (d(i,j) - \bar{d})^2\right]\left[\sum_{i<j}\ (v(i,j) - \bar{v})^2\right]}}$$

where $d(i,j)$ is the similarity and $v(i,j)$ is the cophenetic distance. A higher coefficient indicates a better similarity measure [20].

Cluster validity measure

The number of clusters k can range from 2 to $n-1$ (excluding $k = 1\ and\ k = n$) [3]. The **silhouette coefficient** si is used to measure clustering accuracy, calculated as:

$$s_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where $a(i)$ is the average dissimilarity between object I and other objects in its cluster, and $b(i)$ is the average dissimilarity between object I and the nearest cluster. The silhouette coefficient ranges from -1 to 1 for each object. For $k = 2$, the average silhouette coefficient for all objects across the clusters is computed [3]. With the ideal clusters established using the Silhouette coefficient, the next task is to compute the mean value for each time series data point within the specified cluster. This average value is the rice price serving as the prototype for that given cluster. Thus, we can obtain a value that appropriately describes the price movement of each group and offers a better understanding of the trends and prices for further analysis and forecasting.

EEMD

After determining the prototype, the representative cluster data is separated into training data from January 2009 to December 2020 and testing data from January 2021 to December 2023. Each cluster of rice price data is then decomposed using Ensemble Empirical Mode Decomposition (EEMD) to obtain Intrinsic Mode Functions (IMFs) and the residual component. The contribution of each IMF and the residual is analyzed using the mean period, Pearson correlation, variance, and variance ratio percentage to understand their impact on the overall price trend. Ensemble Empirical Mode Decomposition (EEMD) addresses the mode mixing issue in EMD, which occurs when an IMF contains data with large scale differences or similar scales across different IMFs. [21] developed EEMD by adding fixed-amplitude white noise to the data in each iteration, improving the naturalness of IMF characteristics. The EMD procedure is as follows:

1. Identify local maxima and minima of the data $x(t)$. A local maximum occurs when $x(t) > x(t-1)$ and $x(t) > x(t+1)$; a local minimum occurs when $x(t) < x(t-1)$ and $x(t) < x(t+1)$
2. Create upper and lower envelopes by connecting the local maxima and minima using cubic spline interpolation. These are denoted as $e_{max(t)}$ and $e_{min(t)}$.
3. Calculate the mean: $m(t) = \frac{e_{min(t)} + e_{max(t)}}{2}$
4. Extract the detail: $d(t) = x(t) - m(t)$
   Check the IMF condition for $d(t)$:
   - If $d(t)$ is not an IMF, repeat steps 1–4 using $d(t)$ as the new $x(t)$.
   - To verify IMF, check if the number of zero-crossings and extrema differ by one and whether the mean of the upper and lower envelopes approaches zero.
   - If $d(t)$ is an IMF, denote it as $c_i(t)$
5. Extract the residue: $r_i(t) = x(t) - c_i(t)$
6. Check if the residue is monotonic. If not, repeat steps 1–6 for iii iterations. If the residue is monotonic, stop the sifting process: $r(t) = r_{i-1}(t) - c_i(t)$ with $r_0(t) = x(t)$ and $1 < i < M$

The sifting process stops when the residue becomes a monotonic function or has only one extremum, meaning no more IMFs can be extracted. The original time series is obtained by summing the IMFs and the final residue: $x(t) = \sum_{i=1}^{M} c_i(t) + r(t)$

EEMD Procedure:

1. Add a sequence of white noise $\epsilon \sim N(0, \sigma^2)$ to the data: $y_i(t) = y(t) + n_i(t)$, where y(t) is the original data, y_i(t) is the data with added noise, and $n_i(t)$ is the white noise.
2. Decompose the noise-added data into several IMFs.
3. Repeat steps 1–2 *N* times with different white noise in each iteration.
4. Calculate the IMFs $c_{t(i)}$and residue *r* by averaging the ensemble: $c_{t(i)} = \frac{1}{N}\sum_{j=1}^{N} c_{t(ij)}$ and $r = \frac{1}{N}\sum_{j=1}^{N} r_j$, where *j* is the iteration number and *i* is the index of the IMF

The added white noise creates a uniform frequency-time space for IMF formation. The ensemble average reduces the noise, cleaning the results. The effect of white noise addition is measured by: $\epsilon_n = \frac{\epsilon}{N}$, where *N* is the ensemble size, ϵ is the amplitude of the white noise, and $\epsilon_n$is the final standard deviation of the error defined as the difference between the original data and the corresponding IMF. The ensemble size can be set to 100 trials with a standard deviation between 0.1 and 0.2 [22].

Autoregressive integrated moving average (ARIMA)

The Intrinsic Mode Functions (IMFs) and the residual component are modeled using ARIMA by first testing the stationarity of the data. The optimal ARIMA model is selected by first examining the ACF and PACF plots to identify potential model orders. Several candidate models are then estimated, and the best-fitting model is chosen based on the lowest Akaike Information Criterion (AIC). Once the most suitable model is identified, predictions are generated based on the selected ARIMA model. Finally, the predicted values of the IMFs and the residual component are combined to obtain the final forecast of rice prices, providing a more accurate and comprehensive price estimation. A time series $\{Y_t\}$ is a mixed autoregressive integrated moving average model if, after differencing d times, $W_t = \nabla^d Y_t$ becomes a stationary ARMA process. Thus, if $\{W_t\}$ follows an ARMA (p, q) model, then $\{Y_t\}$ is an ARIMA (p, d, q) process [23]. Suppose $d = 1$, meaning that $\{Y_t\}$ undergoes first-order differencing, resulting in $W_t = \nabla^d Y_t = \nabla Y_t = Y_t - Y_{t-1}$. If $\{W_t\}$ follows an ARMA (p, q) model, then its mixed model, using the notation W_t, is given by:

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

Evaluation Metrics

To measure the forecast accuracy of the EEMD-based technique from January 2021 to December 2023, evaluation metrics including Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are used. These metrics assess how well the model captures actual values and quantify the forecast error from different perspectives: MAPE expresses error as a percentage, RMSE emphasizes larger errors, and MAE gives the average absolute error. Together, they provide a comprehensive view of the model's performance. The formulas are as follows:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|;\ \text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2}\ ;\ \text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|$$

Where *At* is the actual value, *Ft* is the forecasted value, and *n* is the number of observations.

RESULTS AND DISCUSSIONS

The analysis was conducted using secondary data from the BPS report "Statistik Harga Konsumen Pedesaan: Pangan Kelompok," a report that spans 32 provinces in Indonesia from January 2009 to December 2023. It was analyzed using visual tools in the form of time series graphs depicting the changes in the price of rice, making it easier to identify changes, seasonal variations, or movements over time.

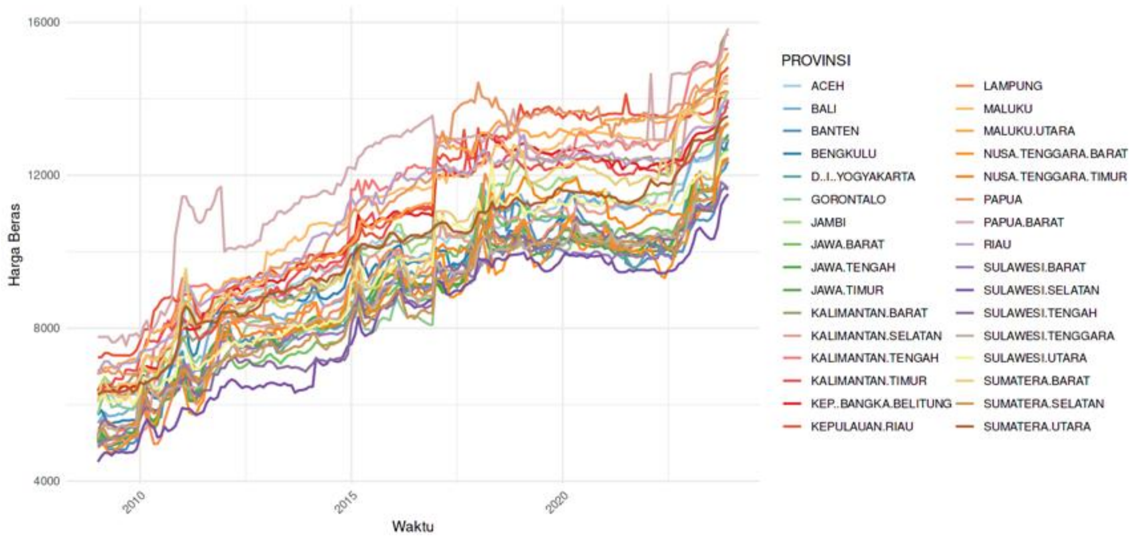

Figure 2. Graph rice price

The graph above shows the trend of rice prices across all provinces every month from 2009 to 2023. The provinces observed relatively similar trends with outliers like West Papua, where rice prices were higher than others during the set time period. Overall, there are several notable spikes in value. For example, the prices increased from below Rp7,000 per kilogram at the end of 2011 to around Rp8,000 per kilogram in early 2012. This increase was caused by the lean period before the main rice harvest, which usually starts at the end of March. [11].

The rice prices throughout Indonesia's provinces fluctuate independently, yet patterns can be identified. This research utilizes time series clustering in order to group provinces with similar movements in price, which helps in the interpretation of data and exposes the inter-provincial relations. Average linkage hierarchical clustering and dynamic time warping (DTW) distance were used and assessed by the cophenetic coefficient. The results were displayed in a dendrogram where the cophenetic correlation of the clusters was 0.64, indicating reasonable clustering quality, which is typical of grouping analysis. The optimal cluster number was determined using the silhouette coefficient for all clusters between two and ten, as illustrated in Figure 3.

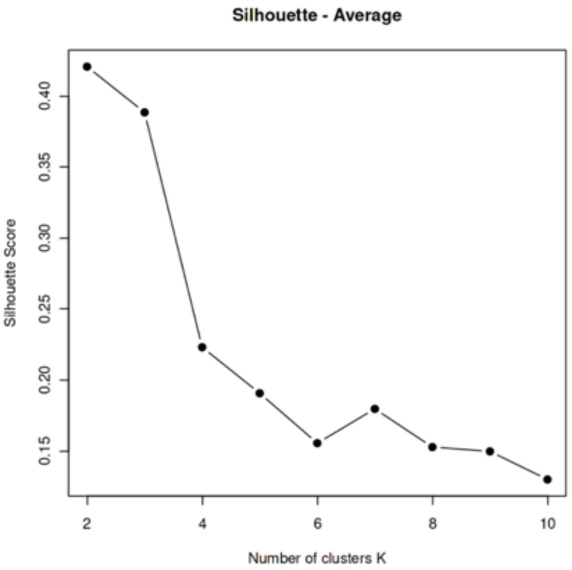

Figure 3. Silhouette coefficient plot

The silhouette score monitored cluster quality by evaluating the average distance between the data points in a single cluster and the other clusters. An increase in score corresponded with improved cluster separation [24]. From the analysis done using the rice prices in Indonesian provinces from 2009 to 2023, the optimal number of clusters was three. Even though the ideal silhouette score was achieved with two clusters, this option was less valuable because they had too sizeable inter-cluster variance, which lessened validation and forecasting accuracy. Balanced clusters had quite a good silhouette score but gave the flexibility needed for increased analysis due to their three-fold nature. The patterns observed with too few clusters did not capture data effectively, while over seven clusters resulted in excessive fragmentation. The solution with three clusters provided the optimum data distribution balance, trend clarity, and forecasting accuracy.

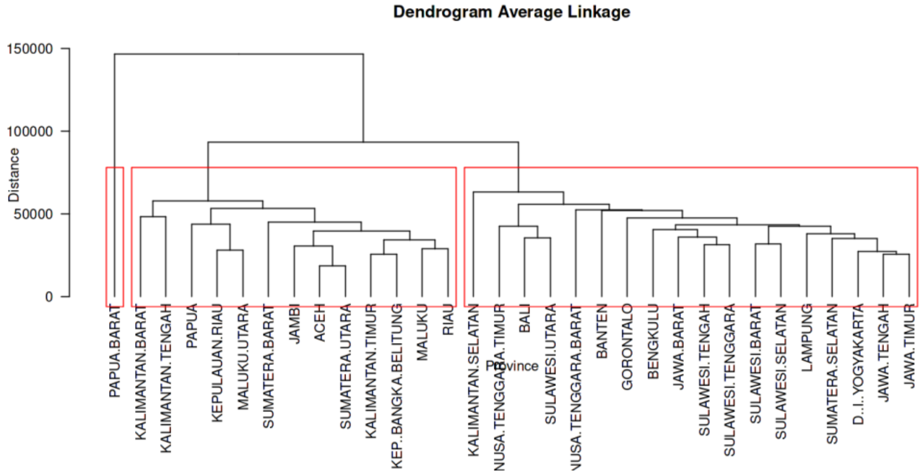

Figure 4. Dendrogram using DTW distance and average linkage for three optimal clusters

The analysis pointed out three optimal clusters. The clustering results presented in Table 1 show the division of the provinces concerning their patterns of rice price movements. It was found that the provinces within each cluster showed similar price behavior, making it possible to predict the prices within each cluster reliably. This was clear from the comparably aligned price movements of the provinces within the same cluster.

Table 1. Cluster grouping details

Cluster	Number of Provinces	Provinces
1	13	Aceh, Jambi, Kalimantan Barat, Kalimantan Tengah, Kalimantan Timur, Kep. Bangka Belitung, Kepulauan Riau, Maluku, Maluku Utara, Papua, Riau, Sumatera Barat, Sumatera Utara
2	18	Bali, Banten, Bengkulu, D.I. Yogyakarta, Gorontalo, Jawa Barat, Jawa Tengah, Jawa Timur, Kalimantan Selatan, Lampung, Nusa Tenggara Barat, Nusa Tenggara Timur, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Tenggara, Sulawesi Utara, Sumatera Selatan
3	1	Papua Barat

Hierarchical clustering with DTW distance and average linkage revealed that most provinces were aggregated into Cluster 1 (13 provinces) and Cluster 2 (18 provinces). These clusters showed relatively uniform price trends. On the contrary, Clusters 3, which contained 1 provinces, displayed greater diversity in price movements.

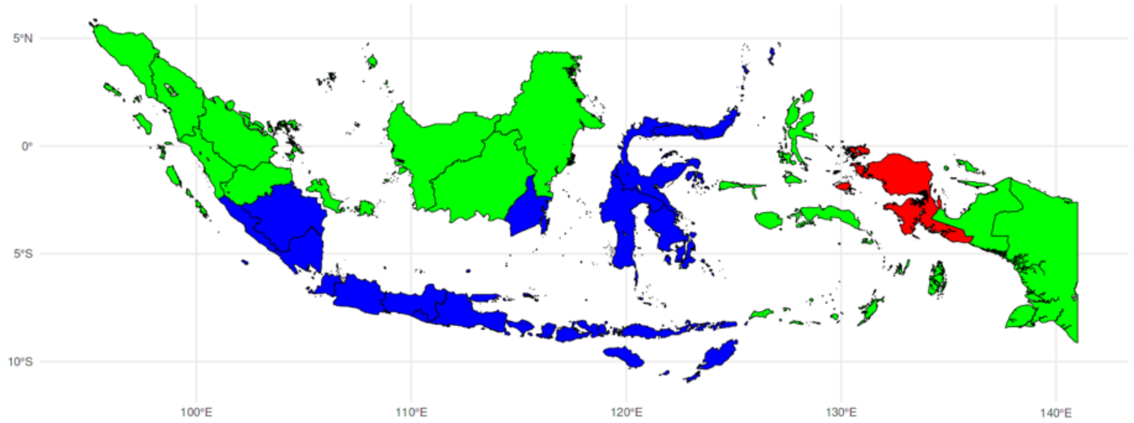

Figure 5. Rice price cluster distribution by province in Indonesia

Each cluster represented provinces with similar rice price trends, identified using DTW distance. After clustering, a prototype for each group was computed as the average rice price pattern of all provinces in the cluster, providing a representative trend for each group.

The modeling process for each cluster required representative data reflecting the unique characteristics of each group. To achieve this, the average value of all cluster members (prototype) was calculated, serving as an aggregate representation of each cluster's price behavior. Before modeling, the data underwent decomposition using the Empirical Ensemble Mode Decomposition (EEMD) method. Monthly rice prices in Indonesia from January 2009 to December 2023 exhibited a significant upward trend (Figure 6). The dataset was split into 80% training data (January 2009–December 2020) and 20% testing data (January 2021–December 2023). Despite a consistent price increase, fluctuations occurred due to factors such as seasonality, government policies, or market conditions, indicating non-stationarity in both mean and variance. Given these characteristics, an advanced analytical approach was necessary to capture complex patterns in the data. EEMD was employed to decompose rice price data into simpler components, facilitating a detailed analysis of trends, seasonality, and fluctuations. The decomposition was applied to each cluster prototype, breaking the data into Intrinsic Mode Functions (IMFs) and a residual component.

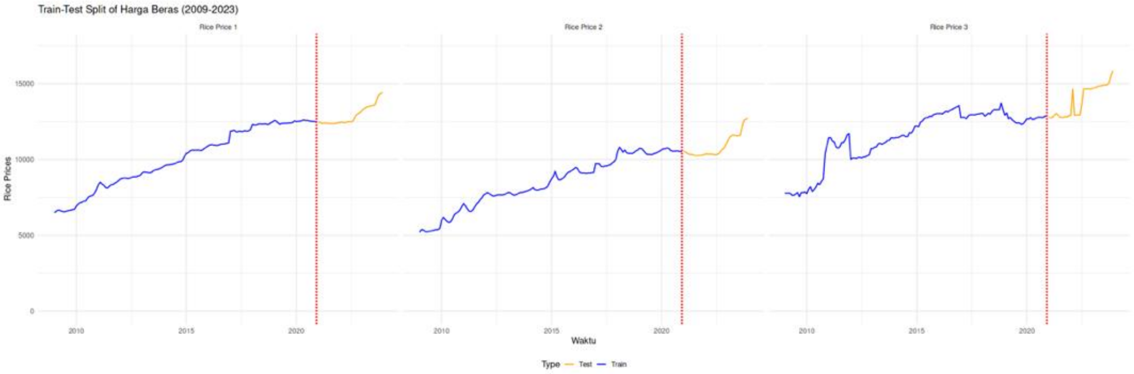

Figure 6 Monthly rice price trends in Indonesia based on prototypes (2009–2023)

The decomposition of Indonesia's monthly rice price data using Ensemble Empirical Mode Decomposition (EEMD) was applied to each prototype from the clusters obtained through clustering analysis. In this process, the data for each cluster were decomposed into Intrinsic Mode Functions (IMFs) with an ensemble size of 100 and a noise strength (standard deviation) of 0.2. The IMFs were extracted sequentially, starting from the highest to the lowest frequency components. The residual component in EEMD, unlike residuals in regression or additive decomposition, did not stem from prediction errors or deviations between observed and predicted values. Instead, it represented the final component that could no longer be decomposed into additional IMFs, reflecting the long-term trend beyond the oscillatory patterns captured by the IMFs.

In regression analysis, residuals referred to the difference between actual observed values and those predicted by the regression model. Meanwhile, in additive decomposition, residuals were typically considered as the part that could not be explained by trend and seasonality. The key difference lay in how these methods interpreted and separated data components. In EEMD, the residual was the final component containing more stable information, usually representing a long-term trend without periodic fluctuations explainable by IMFs [25]. The decomposition results, consisting of IMFs and the residual component from the EEMD application on each cluster, are presented in the following analysis.

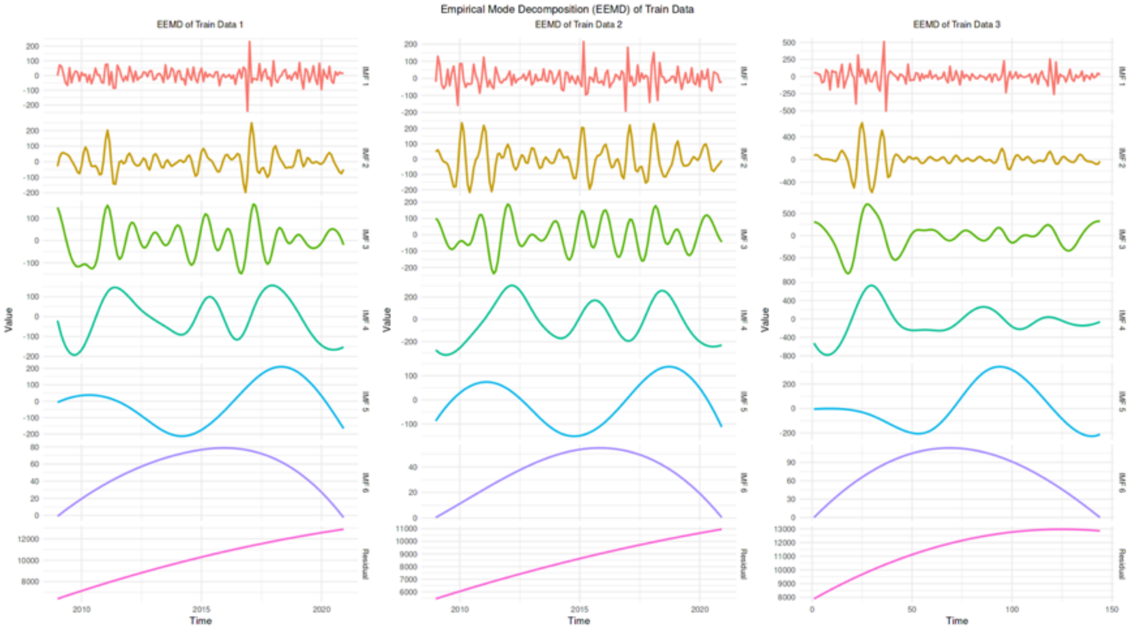

Figure 7. Decomposition results of Indonesia’s monthly rice prices based on EEMD for each cluster

The average period of each IMF was set based on the count of peaks and troughs for each component. The number of peaks, troughs, average periods, Pearson correlation, variance, and variance ratio for each IMF, as well as the residual from rice price decomposition, are shown in Table 2. The overall trend for the average period ranged higher and gradually increased due to the coarsening of the average number of peaks and troughs in the IMFs. For example, for IMFs 1 through 6, the average period in each cluster increased, meaning the components shifted from high-frequency to low-frequency.

The correlation, variance, and variance ratio analysis of the relationships among IMFs was done for each cluster's monthly rice price data. The Pearson correlation established the relevance of the individual IMFs to the overall data. The decomposition results showed that as the average period of an IMF increased, its correlation with the primary rice price trend also became more vigorous. On the other hand, short-average-period IMFs like IMFs 1 and 2 had little correlation and were more representative of intra-annual fluctuations.

Table 2 Descriptive analysis of IMF components and residuals

Cluster	IMF	Peaks	Troughs	Average Periods	Pearson Correlation	Variance	Variance Ratio (%)
Gerombol 1	IMF 1	48	48	-0.7	0.02	2433.55	0.06
	IMF 2	22	22	0.29	0.01	3797.33	0.1
	IMF 3	11	11	-8.4	0.17	5075.94	0.13
	IMF 4	3	4	-15.5	0.11	10861.01	0.28
	IMF 5	2	1	-96	0.33	16207.19	0.42
	IMF 6	1	0	NA	0.31	586.69	0.02
	Residual	0	0	NA	0.99	3603487.61	93.67
Gerombol 2	IMF 1	43	43	-1.38	0.03	3109.09	0.11
	IMF 2	19	19	5	0.12	8584.18	0.3
	IMF 3	11	11	-2.3	0.18	8314.16	0.29
	IMF 4	3	4	21	0.23	32470.04	1.14
	IMF 5	2	1	-92	0.21	8016.69	0.28
	IMF 6	1	0	NA	0.36	287.47	0.01
	Residual	0	0	NA	0.99	2604552.7	91.81
Gerombol 3	IMF 1	44	43	0.33	0.04	11986.74	0.38
	IMF 2	21	21	4.45	0.06	25432.37	0.8
	IMF 3	8	8	11.29	0.32	86924.95	2.75
	IMF 4	4	5	9	0.45	104144.13	3.29
	IMF 5	2	2	-82	0.35	31200.07	0.99
	IMF 6	1	0	NA	0.35	1204.68	0.04
	Residual	1	0	NA	0.94	2355893.38	74.5

The contribution of each IMF to the patterns of rice prices is explained by the variance and the variance ratio (%). IMF 1 and IMF 2 captured short-term fluctuations with high frequencies and, therefore, contributed the least. IMF 4 and IMF 5, on the other hand, had a more significant impact because of their extended average periods. The component that is left over has the largest share of variance, indicating that the most important feature is the trends relative to the season or short-term variation. Short-term IMFs reflect local seasonal impacts, while long-term IMFs and the residual, without having an explicit model, contain more excellent information about the trend. The residual in most clusters had the strongest correlation with the rice price data, proving the significance of the trend. However, in cluster 3, the residual variance ratio was lower, indicating a more complex pattern with both long- and short-term influences, making separating signals difficult. This means that on certain occasions, high-frequency IMFs may adequately po, portray the price variation, suggesting more complex relations between seasonal and long-term trending-term components of the IMF along with the residuals from each cluster (1st January 2009 – 31st December 2021) served as the foundation of the modeling procedure.

Every component was assigned a model, performed parameter estimation, and underwent diagnostics within the Box-Jenkins framework. A grid search was conducted to choose the best model, which involved testing many combinations of values in the model to achieve the best results. The predicted IMF components were summed along with the residuals to produce forecasts for 1st January 2021 to 31st December 2023. Actual data on the price of rice was used to estimate how close the predictions were to the received MAPE. The best ARIMA models for individual components were obtained in residual diagnostics of independence and normality and are all presented in Table 3.

Table 3. The Best model for each cluster of each IMF and residual (January 1, 2009 – December 31, 2020)

Cluster	IMF	Best ARIMA	AIC	p-value	
				Ljung-Box p-value	Jarque-Bera p-value
Cluster 1	IMF 1	ARIMA(1,0,1)	1488.01	0.14	0.00
	IMF 2	ARIMA(6,0,0)	1098.18	0.55	0.00
	IMF 3	ARIMA(0,0,0)	NA	NA	NA
	IMF 4	ARIMA(0,0,0)	NA	NA	NA
	IMF 5	ARIMA(0,0,0)	NA	NA	NA
	IMF 6	ARIMA(1,4,0)	-1541.17	0.94	0.00
	Residual	ARIMA(1,3,0)	-6720.94	0.11	0.00
Cluster 2	IMF 1	ARIMA(5,0,2)	1497.93	0.66	0.01
	IMF 2	ARIMA(0,0,0)	NA	NA	NA
	IMF 3	ARIMA(0,0,0)	NA	NA	NA
	IMF 4	ARIMA(1,5,2)	-308.93	0.14	0.00
	IMF 5	ARIMA(0,0,0)	NA	NA	NA
	IMF 6	ARIMA(1,5,0)	-1752.74	0.96	0.00
	Residual	ARIMA(1,3,0)	-6681.68	0.11	0.00
Cluster 3	IMF 1	ARIMA(3,0,1)	1674.72	0.58	0.00
	IMF 2	ARIMA(6,0,2)	1263.22	0.79	0.00
	IMF 3	ARIMA(3,0,5)	600.91	0.13	0.00
	IMF 4	ARIMA(0,5,1)	-160.19	0.28	0.00
	IMF 5	ARIMA(1,5,0)	-1205.11	0.56	0.00
	IMF 6	ARIMA(0,4,0)	-1415.7	0.99	0.00
	Residual	ARIMA(0,3,2)	-6604.12	0.11	0.00

According to Table 3, the Ljung-Box test residuals for each model were independent (p > 0.05). Meanwhile, the residuals show no normal distribution based on the Jarque-Bera test (p < 0.05). After determining the best model, forecasting was done utilizing each cluster member's decomposed IMF and residual components. The final forecast values, which reflected the data patterns for each cluster member, were obtained by summing the predictions from each IMF with the residual component. Forecast validation was performed by checking the predicted values against the actual data from January 1, 2021, to December 31, 2023. Validation was done using two approaches: cluster-level validation and validation of individual members within each cluster. Error metrics such as Mean Absolute Percentage Error (MAPE) were used to assess the model's accuracy to assess the model's accuracy. The results for the evaluation of prediction and forecasting with the application of clustering on the prototypes of clusters 1 to 3 based on the training and testing data are presented in Table 4.

Table 4 MAPE evaluation for each cluster using EEMD decomposition

Cluster	MAPE (%)		RMSE		MAE	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
1	1.75	3.99	207.64	575.46	172.76	511.08
2	1.91	5.65	190.84	689.27	154.88	605.84
3	0.54	3.05	79.85	595.86	59.04	442.41

The smaller the MAPE value, the better the forecast accuracy. Based on Table 4, the MAPE values for the training data in all clusters are below 2%, and for the testing data, they remain well under the 10% threshold, ranging from 3.05% to 5.65%. This suggests the EEMD-based approach performs reliably across clusters in the training and testing phases. The relatively low RMSE and MAE values further support this, indicating that the prediction errors, both in magnitude and percentage, are consistently minimal across all clusters. Cluster 3 shows the best performance overall, with the lowest MAPE, RMSE, and MAE values, reflecting the most accurate forecast among the clusters.

Table 5. MAPE testing for each province in the cluster using EEMD

Cluster	MAPE (%)				
1	Aceh	Jambi	Kalbar	Kalteng	Kaltim
	12.32	10.94	5.71	4.84	6.69
	Kep. Babel	Kepri	Maluku	Malut	Papua
	5.50	3.33	2.00	3.87	3.53
	Riau	Sumbar	Sumut		
	4.12	6.48	9.70		
2	Bali	Banten	Bengkulu	DIY	Gorontalo
	4.57	9.95	7.33	8.05	5.59
	Jabar	Jateng	Jatim	Kalsel	Lampung
	4.95	7.23	8.87	7.98	6.34
	NTB	NTT	Sulbar	Sulsel	Sulteng
	10.04	2.71	9.14	14.37	6.39
	Sultra	Sulut	Sumsel		
	6.80	2.65	6.95		
3	Papbar				
	3.05				

When the rice price time series data was clustered using the EEMD evaluation, the Mean Absolute Percentage Error (MAPE) showed that different provinces had various prediction accuracy levels. In Cluster 1, the MAPE values ranged from 2.00% to 12.32%. Provinces like Maluku (2.00%), Kepulauan Riau (3.33%), and Papua (3.53%) displayed extremely low error rates, suggesting that EEMD successfully captured price trends in those areas. This suggests the potential for reliable forecasting in relatively stable markets, which can assist local governments in planning regional food distribution or stabilizing prices through timely interventions. However, some provinces like Aceh (12.32%) and Jambi (10.94%) exhibited higher errors, reflecting more volatile conditions that may require adaptive strategies or further model refinement.

In **Cluster 2**, which included a broader set of provinces, MAPE values varied from **2.65% to 14.37%.** While Sulawesi Selatan (14.37%) and NTB (10.04%) displayed higher levels of forecast error, Sulawesi Utara (2.65%), Nusa Tenggara Timur (2.71%), and Gorontalo (5.59%) had relatively low MAPE values, indicating consistent price patterns. These findings suggest that dynamic factors like seasonal supply shifts, infrastructure constraints, and local demand shocks influence price behaviors in specific areas. By acknowledging these regional differences, stakeholders, including suppliers and market regulators, can adjust their monitoring and policy strategies to the unique features of each cluster.

The potential of EEMD as a support tool for early warning systems in remote areas where logistics are crucial was highlighted by Cluster 3, which only included Papua Barat, and had a low MAPE of 3.05%. The model did well in the majority of provinces, primarily those where MAPE values were less than 10%, but the other differences highlight how crucial it is to incorporate additional data or techniques for better prediction. Policymakers and agricultural stakeholders looking to improve food security will find these insights especially pertinent. Targeted interventions like price subsidies, import adjustments, or infrastructure improvements can be better informed by identifying areas with stable versus volatile price trends. This will increase market stability and facilitate more effective resource allocation.

After we had analyzed the MAPE for every cluster, the next step was to estimate the rice prices for the upcoming 24 time periods. Regarding clustering, each group was treated individually for trend and seasonal components to improve the accuracy of the forecast. Using EEMD on the extracted components allowed us to capture the significant patterns later used in the prediction. The following graph at Figure 8 shows the expected price of rice for different clusters for the following 24 time periods, along with their probable movements over the designated periods.

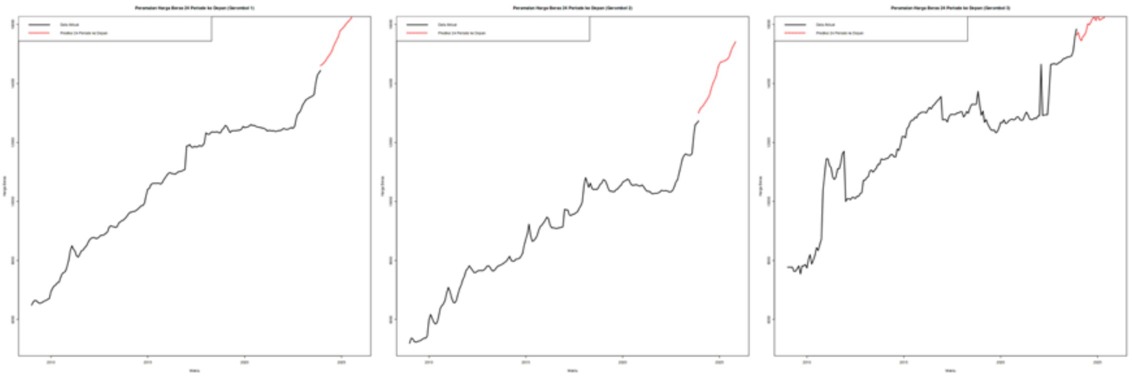

Figure 8. Forecasting rice prices for the next 24 periods for each cluster

CONCLUSION

This study uses a *time series clustering* approach and Ensemble Empirical Mode Decomposition (EEMD) to forecast rice prices in Indonesia, demonstrating promising results. Applying Dynamic Time Warping (DTW) and *hierarchical clustering* with the *average linkage* method, three optimal clusters were identified based on price movement patterns across various provinces. The EEMD decomposition process separated price components based on frequency, with the IMF reflecting short-term fluctuations and the residual capturing the overall trend. The ARIMA model built on each component from the decomposition showed an MAPE range from 2.65% to 12.32%. Some provinces exhibited low prediction errors, suggesting the model performs well under relatively stable market conditions. However, in areas with higher volatility, the results indicate that this approach has limitations and could be further developed. The province grouping also opens up opportunities to utilize the model's results to inform decision-making in the food sector. For example, mapping regions based on price pattern similarities could assist in devising more efficient strategies for rice distribution and supply control. This approach has the potential for further development by incorporating external factors such as weather, distribution costs, or government policies to improve model accuracy and relevance to the ever-changing market dynamics.

REFERENCES

[1] J. Liu, “Navigating the Financial Landscape: The Power and Limitations of the ARIMA Model,” Highlights in Science Engineering and Technology, vol. 88, pp. 747–752, 2024, doi: https://doi.org/10.54097/9zf6kd91.

[2] N. Khamidah, R. A. Astari, A. Fitrianto, Erfiani, and A. N. Pradana, “PENERAPAN MULTI-CLUSTERING DALAM PENGELOMPOKAN KABUPATEN/KOTA DI PROVINSI JAWA BARAT BERDASARKAN INDEKS DESA MEMBANGUN,” Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika, vol. 4, no. 1, pp. 1651–1665, 2023.

[3] D. A. N. Sirodj, I. M. Sumertajaya, and A. Kurnia, “Analisis Clustering Time Series untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indeks Pembangunan Manusia Jenis Kelamin Perempuan,” STATISTIKA Journal of Theoretical Statistics and Its Applications, vol. 23, no. 1, pp. 29–37, 2023, doi: 10.29313/statistika.v23i1.2181.

[4] S. Fadhlia, I. M. Sumertajaya, and A. Djuraidah, “Penggerombolan Deret Waktu Dengan Pendekatan Ukuran Kemiripan Piccolo Untuk Peramalan Curah Hujan Provinsi Banten,” Indonesian Journal of Statistics and Its Applications, vol. 4, no. 2, pp. 382–391, 2020, doi: 10.29244/ijsa.v4i2.607.

[5] W. Adinugroho, “Pendekatan Clustering Time Series pada Peramalan Harga Minyak Goreng,” Jurnal Ilmiah Populer Median, vol. 4, pp. 47–55, 2021.

[6] R. Novidianto, A. Tri, R. Dani, and P. Hubei, “Analisis Klaster Kasus Aktif Covid-19 Menurut Provinsi Di Indonesia Berdasarkan Data Deret Waktu,” Jurnal Aplikasi Statistika & Komputasi Statistik, pp. 15–24, 2020.

[7] Y. Rahkmawati and S. Annisa, “Clustering Time Series Using Dynamic Time Warping Distance in Provinces in Indonesia Based on Rice Prices,” TIERS Information Technology Journal, vol. 4, no. 2, pp. 115–121, 2023, doi: 10.38043/tiers.v4i2.5081.

[8] Z. Ma, G. Wen, and C. Jiang, “EEMD independent extraction for mixing features of rotating machinery reconstructed in phase space,” Sensors (Switzerland), vol. 15, no. 4, pp. 8550–8569, Apr. 2015, doi: 10.3390/s150408550.

[9] Y. Wan, P. Song, J. Liu, X. Xu, and X. Lei, "A hybrid model for hand-foot-mouth disease prediction based on ARIMA-EEMD-LSTM," BMC Infect Dis, vol. 23, no. 879, Dec. 2023, doi: 10.1186/s12879-023-08864-y.

[10] H. Fransiska, H. Wijayanto, and B. Sartono, "Ensemble Decomposition Method for Predicting the Price of Rice in Jakarta," IOSR Journal of Mathematics, vol. 10, no. 4, pp. 92–97, 2014, doi: 10.9790/5728-10419297.

[11] D. Fajriani, "Pemerintah diminta antisipasi kenaikan harga beras," www.antaranews.com. Accessed: Feb. 16, 2025. [Online]. Available: https://www.antaranews.com/berita/295507/pemerintah-diminta-antisipasi-kenaikan-harga-beras

[12] D. H. Darwanto and E. S. Rahayu, "Analisis Faktor-Faktor Yang Mempengaruhi Impor Beras Indonesia," Caraka Tani: Journal of Sustainable Agriculture, vol. 23, no. 1, p. 1, 2017, doi: 10.20961/carakatani.v23i1.13732.

[13] A. Novelino, "Data BPS: Harga Beras di 179 Daerah Naik Tinggi," Jakarta, 2024. [Online]. Available: https://www.cnnindonesia.com/ekonomi/20240219155728-92-1064615/data-bps-harga-beras-di-179-daerah-naik-tinggi-pekan-lalu

[14] E. Siswanto, B. Marulitua Sinaga, and Harianto, "The Impact of Rice Policy on Rice Market and The Welfare of Rice Producers and Consumers in Indonesia," Jurnal Ilmu Pertanian Indonesia, vol. 23, no. 2, pp. 93–100, 2018, doi: 10.18343/jipi.23.2.93.

[15] Badan Pusat Statistik, Statistik Harga Konsumen Perdesaan Kelompok Makanan 2010. Jakarta: Badan Pusat Statistik, 2010.

[16] M. Yohansa, K. A. Notodiputro, and E. Erfiani, "Dynamic Time Warping Techniques for Time Series Clustering of Covid-19 Cases in DKI Jakarta," ComTech: Computer, Mathematics and Engineering Applications, vol. 13, no. 2, pp. 63–73, 2022, doi: 10.21512/comtech.v13i2.7413.

[17] M. A. Zen, S. Wahyuningsih, and A. T. R. Dani, "Aplikasi Pendekatan Agglomerative Hierarchical Time Series Clustering untuk Peramalan Data Harga Minyak Goreng di Indonesia," Seminar Nasional Official Statistics, vol. 2022, no. 1, pp. 293–302, 2022, doi: 10.34123/semnasoffstat.v2022i1.1394.

[18] W. Wang, G. Lyu, Y. Shi, and X. Liang, "Time Series Clustering Based on Dynamic Time Warping," Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, vol. 2018-Novem, pp. 487–490, 2018, doi: 10.1109/ICSESS.2018.8663857.

[19] L. P. W. Adnyani and P. R. Sihombing, "Analisis Cluster Time Series Dalam Pengelompokan Provinsi Di Indonesia Berdasarkan Nilai Pdrb," Jurnal Bayesian : Jurnal Ilmiah Statistika dan Ekonometrika, vol. 1, no. 1, pp. 47–54, 2021, doi: 10.46306/bay.v1i1.5.

[20] A. D. Munthe, "Penerapan Clustering Time Series Untuk Menggerombolkan Provinsi Di Indonesia Berdasarkan Nilai Produksi Padi," Jurnal Litbang Sukowati : Media Penelitian dan Pengembangan, vol. 2, no. 2, p. 11, 2019, doi: 10.32630/sukowati.v2i2.61.

[21] S. Wang, N. Zhang, L. Wu, and Y. Wang, "Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method," Renew Energy, vol. 94, pp. 629–636, Aug. 2016, doi: 10.1016/j.renene.2016.03.103.

[22] C. Nursyifa, H. Wijayanto, and B. Sartono, "Indentifikasi Pola Pergerakan Harga Beras melalui Dekomposisi Deret Waktu secara Ensemble," Prosiding Seminar Nasional Statistika Universitas Diponegoro, pp. 89–100, 2013.

[23] X. Liu, Y. Zhang, and Q. Zhang, "Comparison of EEMD-ARIMA, EEMD-BP and EEMD-SVM algorithms for predicting the hourly urban water consumption," Journal of Hydroinformatics, vol. 24, no. 3, pp. 535–558, May 2022, doi: 10.2166/HYDRO.2022.146.

[24] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index Pada Hasil Cluster K-Means dan DBSCAN," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5001.

[25] Z. Chen, B. Liu, X. Yan, and H. Yang, "An improved signal processing approach based on analysis mode decomposition and empirical mode decomposition," Energies (Basel), vol. 12, no. 16, Aug. 2019, doi: 10.3390/en12163077.