Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/AP06.20250192
Published: August 5, 2025

Regional Clustering of Chili Prices in Indonesia Using Six Clustering Methods and VARD Modeling

Ani Andriyati and Embay Rohaeti
Mathematics Study Program
Pakuan University
Bogor, West Java, Indonesia
ani.andriyati@unpak.ac.id
embay.rohaeti@unpak.ac.id

Muhammad Edy Rizal
Data Science Study Program
Tadulako University
Palu, Central Sulawesi, Indonesia
muhedyrizal@untad.ac.id

Abstract

This study aims to cluster Indonesian regencies and cities based on chili price movement patterns using a model-based Multivariate Time Series Clustering (MTSClust) approach. Daily price data for four chili types: big red chili, curly red chili, red bird's eye chili, and green bird's eye chili, were collected from 72 regions between January 2022 and December 2024. Missing values were imputed using the Last Observation Carried Forward (LOCF) method. Price dynamics for each region were modeled using Vector Autoregressive with Differencing (VARD), producing coefficient matrices that captured temporal and cross-variable relationships. These matrices served as input for clustering, which was performed using six scenarios combining K-means and K-medoids algorithms with three distance measures: Euclidean, Squared Euclidean, and Canberra. Evaluation using Root Mean Square Standard Deviation (RMSSTD) and R-Squared (RS) identified the K-means algorithm with Canberra distance as the best-performing method, constantly has lower RMSSTD and higher R-Squared, with an average RMSSTD of 51.02559 and an average R-Squared of 0. 998623.

Keywords
Chili Prices, MTSClust, VARD, K-Means Canberra, Clustering

1. Introduction
Chili is a crucial horticultural commodity in Indonesia, widely consumed and traded across regions (Lukas et al., 2023; Sundari et al., 2023; Surya and Tedjakusuma, 2022). However, chili prices are highly volatile, often fluctuating due to regional disparities in production, distribution, and market conditions (Muflikh et al., 2021; Van et al., 2017; Webb and Kosasih, 2011). These fluctuations have significant implications, particularly for consumer purchasing power and farmer incomes. Sudden price increases can reduce affordability, while drastic drops may harm farmer livelihoods. Effective price control strategies are therefore essential for food security and economic stability.

Indonesia's geographical diversity, with 514 regencies and cities, contributes to a complex price dynamic, with each region exhibiting unique price behavior. The complexity is further magnified when using multivariate time series

(MTS) data, especially when the data consist of multiple commodities observed across long time periods. Traditional forecasting techniques become inefficient when applied to such large, scale multivariate datasets.

To address this, clustering regions with similar price movement patterns becomes a practical solution. Clustering allows grouping of areas based on similarity in chili price behaviors, thereby enabling region-level forecasting strategies instead of managing each region independently (Angga Juliartha et al., 2024; Fariz Fadillah Mardianto et al., 2024). This study employs Multivariate Time Series Clustering (MTSClust) to identify these groupings using Vector Autoregressive with Differencing (VARD) models as the underlying data representations (Embay Rohaeti et al., 2023; Srinivasan, 2022).

In addition, the study investigates multiple clustering configurations by combining different algorithms (K-means and K-medoids) with various distance measures (Euclidean, Squared Euclidean, and Canberra) (Jollyta et al., 2023; Raeisi and Sesay, 2022). This approach is essential to identify the most effective clustering method for data, which is critical for downstream tasks like cluster-level forecasting and policy planning.

Unlike previous approaches that directly clustered raw data, this study proposes a model-based clustering framework using VARD-transformed features (Ajeng et al., 2023; Anjani and Bahtiar, 2024; Meiriza et al., 2023). By converting each regional time series into a set of coefficient matrices, the approach captures the dynamic relationships among chili price variables while reducing noise and dimensionality.

Despite previous studies exploring time series clustering, few have addressed chili price behavior across multiple regions using model-based MTSClust. By doing so, this research contributes not only to methodological advancement but also to practical applications in agriculture and market stability.

1.1 Objectives

This study aims to analyze regional chili price patterns in Indonesia and group regions based on the similarity of their price movements using model-based multivariate time series clustering. The specific objectives of the study are as follows:

1. To model chili price data from 72 regencies and cities across four chili types using the Vector Autoregressive with Differencing (VARD) model.
2. To perform clustering using the MTSClust framework with six different clustering scenarios, combining K-means and K-medoids algorithms with three types of distance measures (Euclidean, Squared Euclidean, and Canberra).
3. To evaluate the clustering performance based on Root Mean Square Standard Deviation (RMSSTD) and R-Squared (RS) metrics to determine the best clustering method.

2. Literature Review

2.1 Chili in Indonesia

Chili peppers are a horticultural crop from the Capsicum genus that are widely used as a cooking ingredient in various countries, including Indonesia. The main ingredient in chili peppers is capsaicin , the active compound that gives them their spicy sensation. Chili peppers also contain various essential nutrients, such as vitamin C and provitamin A, which are beneficial for health. Common types of chili peppers consumed in Indonesia include red cayenne peppers, green cayenne peppers, large red chili peppers, and curly red chili peppers. Chili peppers are a strategic commodity because their price fluctuations often affect inflation rates and economic stability. Sharp fluctuations in chili prices can also impact people's purchasing power, especially in the food sector (Yuditya et al., 2023).

Chili prices are heavily influenced by market and distribution conditions. Significant price fluctuations often occur due to an imbalance between supply and demand, as well as logistical constraints affecting interregional distribution, such as transportation issues or inadequate storage (Yuditya et al., 2023) . Differences in chili prices between regions are common in Indonesia, which has 514 regencies and cities with diverse characteristics. Data shows that 72 regencies and cities serve as chili price monitoring centers. East Java, for example, tends to experience price declines during harvest time, while West Papua often experiences price spikes due to limited distribution and the long distance from production centers (Bank Indonesia, 2018).

2.2 Last Observation Carried Forward (LOCF)

Last Observation Carried Forward (LOCF) is an imputation method that uses the last available value before the missing data is filled in. This approach is simple and frequently used because it is easy to implement, especially in time series data(Risnayah and Sagala, 2023).

If the missing data is at the beginning of the series, this method cannot retrieve previous values. Therefore, filling is done using the first value that is completely available in the data (Hasibuan and Novialdi, 2022).

2.3 Vector Autoregressive (VAR)

According to Sumertajaya et al. (2023), the Vector Autoregressive (VAR) model is a regression system consisting of several equations. Each variable is regressed against other variables, including itself, at a previous time. According to Hamilton (2020), the VAR model is formulated as follows:

$$x_t = \Phi_0 + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \cdots \Phi_p x_{t-p} + w_t$$

Where $x_t$is data vector in the th period t, Φ_0 the intercept vector, $\Phi_1, \Phi_2, \ldots, \Phi_p$ is coefficient matrix, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$is the data at previous time, w_t is white noise vector, and $t$is observation period.

Vector Autoregressive with Differencing (VARD) is a form of VAR model used on time series that have been transformed into first-difference form. This model is used when the data is non-stationary and does not exhibit a cointegration relationship. Transformation to first-difference form is performed to eliminate long-term trends, thus rendering the data stationery. With stationary data, a VAR model can be constructed accurately (Embay Rohaeti et al., 2023).

The VARD modeling equation is as follows:

$$\Delta x_t = \Phi_0 + \Phi_1 \Delta x_{t-1} + \Phi_2 \Delta x_{t-2} + \cdots \Phi_p \Delta x_{t-p} + w_t$$

Where Δx_t is the data vector of differencing results period to-t.

According to Bashir & Wei (2017), the optimal lag is determined based on criteria such as the Akakike Information Criterion (AIC). The smallest AIC value is selected as the optimal lag . The AIC formula is written as follows:

$$AIC(m) = ln\left|\tilde{\Sigma}_u(m)\right| + \frac{2\mathrm{m}K^2}{T}$$

Where ln is natural logarithm, $\tilde{\Sigma}_u(m)$ is error covariance estimator at *lag* nm, T is number of observation, $K$is number of variables, and m is order estimation($\hat{p}$).

According to Brockwell and Davis (2016), parameter estimation for the VAR(1) model with two variables is carried out using the Maximum method. Likelihood Estimation (MLE). Likelihood function can be written as follows:

$$L(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

The parameter values satisfy $-\infty < \mu < \infty$and $\sigma > 0$.

Maximum Process Likelihood Estimation (MLE) is obtained by maximizing the likelihood function $L(\mu, \sigma^2)$. This maximization is equivalent to minimize the negative two times the log-likelihood , namely:

$$-2\ln L(\mu, \sigma^2) = n\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Maximum likelihood estimation for μ and σ^2 is obtained by solving the estimated equation resulting from the differentiation process for the parameters concerned.

2.4 ADF Test

According to Febrianti et al. (2021), stationarity testing is performed using the Augmented Dickey-Fuller (ADF) method. This method is used to test for the presence of a unit root in time series data. The test is conducted through hypothesis testing as follows:

i. Null Hypothesis (H_0) : Data contains *a unit root* or not stationary.
ii. Alternative hypothesis (H_1) : The data does not contain *a unit root* or stationary.

The ADF test statistic is obtained using the following formula:

$$ADF_{hitung} = \frac{\hat{\Phi}}{SE(\hat{\Phi})}$$

The calculation components include the standard error and residual error variance used in the ADF test statistic. Standard error:

$$SE(\hat{\Phi}) = \left[\hat{\sigma}_e^2 \sum_{t=1}^{n} (Y_{t-1}^2)^3\right]^{1/2}$$

Residual error variety:

$$\hat{\sigma}_e^2 = \sum_{t=1}^{n} \frac{(Y_{t-\hat{\Phi}} Y_{t-1})^2}{(n-1)}$$

Where n is Total number of observations, $t = 1, \ldots, n$ is Time index, $Y_0 = 0$ is Initial data value, $\hat{\Phi}$ is Lag coefficient, $SE(\hat{\Phi})$ is Standard error of$\hat{\Phi}$, and $\hat{\sigma}_e^2$ is Variety of residual errors.

The null hypothesis (H_0) is rejected if the value ADF_{value} falls below the critical value in the ADF Critical Value table. The significance level used is, for example, 5%. If H_0rejected, the data is deemed to meet the stationarity assumption.

2.5 **K-Means Clustering**

According to Rohaeti et al. (2023) the K- means algorithm consists of the following stages:

1. Determining the value of k, namely the optimal number of clusters.
2. Random selection of k initial cluster centers.
3. Calculation of the distance of each object to each cluster center using a certain distance measure.
4. Placement of each object into a cluster with the closest distance to the cluster center.
5. The new cluster center is determined by calculating the average of all objects in the same cluster. According to Azrahwati et al. (2022), the centroid of the cluster is obtained using the following formula:

$$\boldsymbol{C_k} = \frac{\mathbf{1}}{\boldsymbol{n_k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n_k}} \boldsymbol{d_i}$$

   Where $\boldsymbol{C_k}$ is centroid of the th cluster$\boldsymbol{k}$, $\boldsymbol{n_k}$ is number of members in cluster k, $\boldsymbol{d_i}$ is each object included in the cluster$\boldsymbol{k}$,
6. Repeat steps 3 to 5 until the cluster center is stable and does not experience significant changes.

2.6 **K-Medoid Clustering**

According to Nahdliyah et al. (2019) the K- medoids algorithm consists of the following stages:

1. Determination of the value $k$that indicates the number of clusters to be formed.
2. Determination of k random medoid .
3. Calculate the distance between non-medoid and medoid objects in each cluster and then locate the nearest non-medoid object .
4. Random selection of non-medoid objects in each cluster as new medoid candidates .
5. Calculate the distance between the non-medoid object and the new medoid candidate and then place the non-medoid object to the nearest medoid candidate .
6. Calculation of the difference in total distance (S total distance) with S total distance = total distance on the new medoid candidate minus the total distance on the old medoid .
7. If S total distance < 0, the new medoid candidate will become the new medoid. If S total distance > 0, the iteration is stopped.
8. Repeat steps 4 to 7 until there is no change in medoid or S total distance > 0.

2.7 **Distance Measurements**

According to Rohaeti et al (2023), Euclidean distance is calculated as the square root of the sum of the squared differences between the objects being compared. The equation for calculating Euclidean distance is as follows:

$$d_{i,j} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Where $d_{i,j}$ is distance between object $i$and object$j$, x_{ik} is value of the object $i$in the variable$k$, x_{jk} is value of the object $j$in the variable$k$, and p is number of variables observed

Meanwhile, according to Fathia et al. (2016), the Squared Euclidean distance is used to measure the closeness between two variables being compared. This method is included in the category of distance measures considering quadratic differences. The Squared Euclidean distance between the nth object $k$and the nth object $j$based on $p$the variables is:

$$d_{i,j} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

Where $p$is the number of observed variables, $d_{i,j}$ is distance between object $i$and object$j$, x_{ik} is data value of the object to- $i$in the variable to-k, and x_{jk} is data value of the object $j$in the variable$k$.

According to Hasanah and Sofro (2022), the Canberra distance is a measure of proximity used in clustering methods. This measure compares two objects by calculating the ratio between the absolute value of the difference between the two variables and the sum of the two values for each object. This measure is only suitable for positive data. The Canberra distance formula is stated as follows:

$$d_{i,j} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$$

Where $p$is number of observed variables, $d_{i,j}$ is distance between the th object $i$and the th object$j$, x_{ik} is value of the object to the variable to the variable $i$to the variable to the variable to the object ...k, and x_{jk} is value of the object $j$in the variable$k$.

3. Methods

This study adopts a quantitative approach using multivariate time series analysis to cluster regencies and cities in Indonesia based on chili price data. The overall methodology involves four key stages: data preparation, time series modeling, clustering, and cluster evaluation.

The dataset consists of daily chili price observations from January 2022 to December 2024 across 72 regencies and cities, covering four types of chili: big red chili, curly red chili, red bird's eye chili, and green bird's eye chili. Missing values are handled using the Last Observation Carried Forward (LOCF) method, where each missing entry is imputed with the last known observation in the same series.

Following imputation, the time series are modeled using the Vector Autoregressive with Differencing (VARD) approach to address non-stationarity. Each region's data yields a VARD model capturing multivariate price dynamics among the four chili types. In total, 288 VARD models are built (four per region), and each modeling process is repeated 100 times to ensure stability in selection and evaluation.

The resulting model structures, specifically the coefficient matrices, are then used as inputs for clustering using the Multivariate Time Series Clustering (MTSClust) framework. Clustering is carried out using two non-hierarchical algorithms: K-means and K-medoids. Each algorithm is evaluated under three distance measures: Euclidean, Squared Euclidean, and Canberra, resulting in six clustering scenarios. Every clustering scenario is also repeated 100 times, yielding 600 total clustering results from which the best-performing method is selected.

To assess clustering performance, two internal validation metrics are applied: Root Mean Square Standard Deviation (RMSSTD), which measures within-cluster homogeneity, and R-Squared (RS), which quantifies the variance explained between clusters. The clustering method with the best combination of low RMSSTD and high RS is selected for final interpretation and forecasting.

All data processing, modeling, and clustering procedures were implemented in R programming language, utilizing relevant statistical and time series packages to handle data imputation, model estimation, and clustering evaluation.

4. Data Collection

This study uses secondary data obtained from the *Pusat Informasi Harga Pangan Strategis* (PIHPS), available through the official website of Bank Indonesia at https://www.bi.go.id/hargapangan. The dataset comprises daily chili price observations from January 2022 to December 2024 across 72 regencies and cities in Indonesia.

The data includes prices for four chili types: big red chili, curly red chili, red bird's eye chili, and green bird's eye chil. Each commodity contains 78,768 daily observations, resulting in a total of 315,072 data points across all variables and regions. The use of daily data allows for detailed analysis of both short-term and long-term price fluctuations, while the wide geographical coverage ensures that the regional price diversity in Indonesia is adequately captured.

5. Results and Discussion

5.1 Numerical Results

The analysis began with the construction of multivariate time series models for each region to capture the dynamic interaction between the four chili price variables. To ensure the validity of the models, it was necessary to first assess the stationarity of the series and apply differencing when required. Therefore, we performed ADF test before modeling process. As an illustration, Table 1 shows the results for ADF test on all the commodities in Ambon City.

Table 1. ADF Test Results for Imputed Data from Ambon City

Variable	*p-values*	Stationarity
Big Red Chili	0.7829	Not Stationary
Curly Red Chili	0.0609	Not Stationary
Green Bird's Eye Chili	0.1303	Not Stationary
Red Bird's Eye Chili	0.2829	Not Stationary

The data being tested in Table 1 is the data after imputation using Last Observation Carried Forward (LOCF). Table 1 is an illustration of the conditions across the 72 regencies and cities, where all the commodities are not stationary most of the time. Therefore, the use of VAR is not possible, and we had to use VAR in difference (VARD) instead.

Table 2. ADF Test Results for Imputed Data from Ambon City After Differencing with the order $\boldsymbol{d = 1}$

Variable	*p-values*	Stationarity
Big Red Chili	0.01	Stationary
Curly Red Chili	0.01	Stationary
Green Bird's Eye Chili	0.01	Stationary
Red Bird's Eye Chili	0.01	Stationary

Table 2 shows the ADF test results for all the commodities from Ambon City after data differencing with order $d = 1$. As can be seen, all the commodities are now stationary, and therefore suitable to be fitted using VAR model.

Table 3. AIC Scores for Imputed Differenced Data from Ambon City

Lag	AIC Score	Lag	AIC Score
1	6.70130	6	6.70808
2	6.70298	7	6.70921
3	6.70476	8	6.70639
4	6.70570	9	6.70853
5	6.70734	10	6.70135

To get the best model for the data, we tried modeling the data on multiple lags. Table 3 shows the AIC scores for all tested lags. Based on the table, the optimal lag is $p = 1$ with an AIC score of 6.70130. Therefore, in the case of Ambon City, we use VAR(1,1) or VARD(1) for fitting the data. Portmanteau test also shows a p-value of 0.5234, indicating that there is no residual autocorrelation, supporting the argument that VARD(1) is indeed suitable for this case.

The modelling process was done to all the 72 regencies and cities, resulting in 72 coefficient matrices. These matrices are used as input data for the clustering phase, which is explained in the section 5.3. The modeling and clustering steps

were performed 100 times, which then evaluated using RMSSTD and RS. Table 4 shows the evaluation for all the clustering methods.

Table 4. Optimal Number of Clusters, Average RMSSTD, and Average RS of All the Clustering Methods

Methods	Distance	Avg. RMSSTD	Avg. RS
K-Means	Canberra	51.02559	0.998623
K-Means	Euclidean	238.2891	0.970106
K-Means	Squared Euclidean	556.4499	0.878225
K-Medoid	Canberra	1281.183	0.349681
K-Medoid	Euclidean	544.1965	0.874486
K-Medoid	Squared Euclidean	527.1069	0.877088

5.2 Graphical Results

Figure 1 illustrates that the missing data in one of the cities. that is. in Ambon.

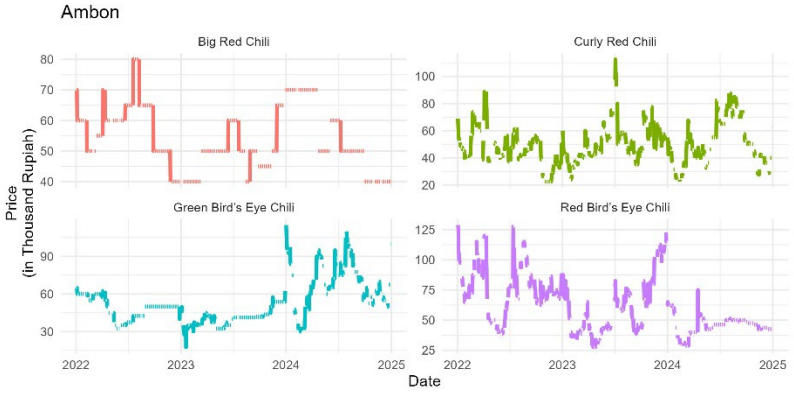


Figure 1. Illustration of Missing Data in The Raw Dataset

Among the 72 regencies and cities. Blitar Regency has the lowest number of missing values with 1.304 values across the four commodities. equivalent to around 30% of the whole dataset for Blitar Regency. Bontang is the worst regency in terms of the number of missing values. which amounts to 1.536 values. equivalent to around 35% of the whole data for the Bontang Regency.

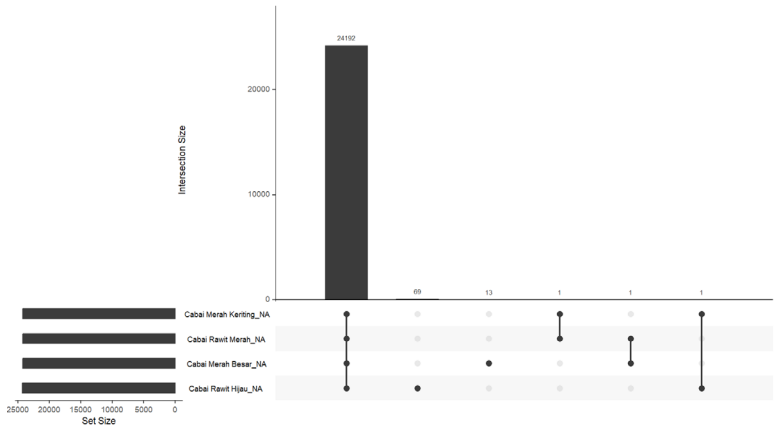


Figure 2. Number of Missing Value Cases based on Commodity

In terms of commodities. all the four commodities have relatively similar number of missing values: 24.206 (30.7%) missing values for big red chili; 24.194 (30.7%) missing values for curly red chili (Figure 2); 24.262 (30.8%) missing values for red bird's eye chili. and 24.194 (30.7%) missing values for green bird's eye chili. It is also found that. as shown in 2. that most of those missing values occur on the same day for across all of the commodities.

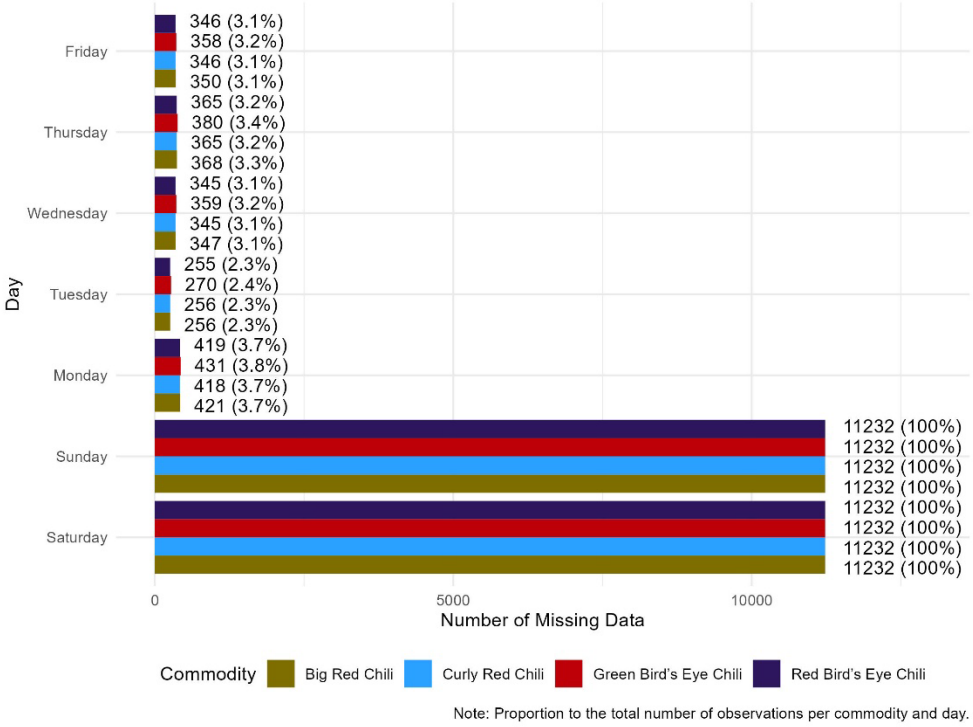
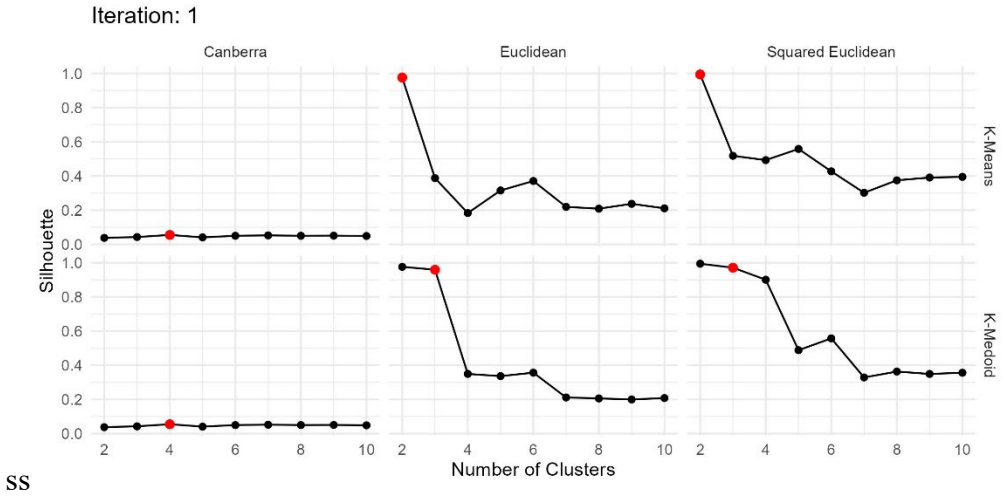


Figure 3. Number of Missing Values based on Days

Missing values is an important problem since time series modeling depends on the order of the data (Figure 3). Therefore. dealing with missing values is a top priority in data preprocessing. Most of the missing values happened on either Saturday or Sunday. as clearly shown in 3. Since prices on Friday are most likely the same as prices on Monday. it can be safely assumed that simple data imputation techniques such as Last Observation Carried Forward (LOCF) should suffice.



SS

Figure 4. Optimal Number of Clusters for All the Clustering Methods for the First Repetition

After VAR modeling process in section 5.2, we clustered the resulted 72 coefficient matrices. To get the optimal clusters, we calculated the silhouette scores for $k = 2$ to $k = 10$. Figure 4 shows the optimal number of clusters for each clustering methods. As shown in Figure 5, the optimal number of clusters varies, from $k = 2$ to $k = 4$. These optimal number of clusters were then evaluated using RMSSTD and RS. After 100 repetitions, the average RMSSTD and RS were calculated and were shown in Table 4. The full distribution of the RMSSTD and RS scores across 100 repetitions is shown in Figure 5.

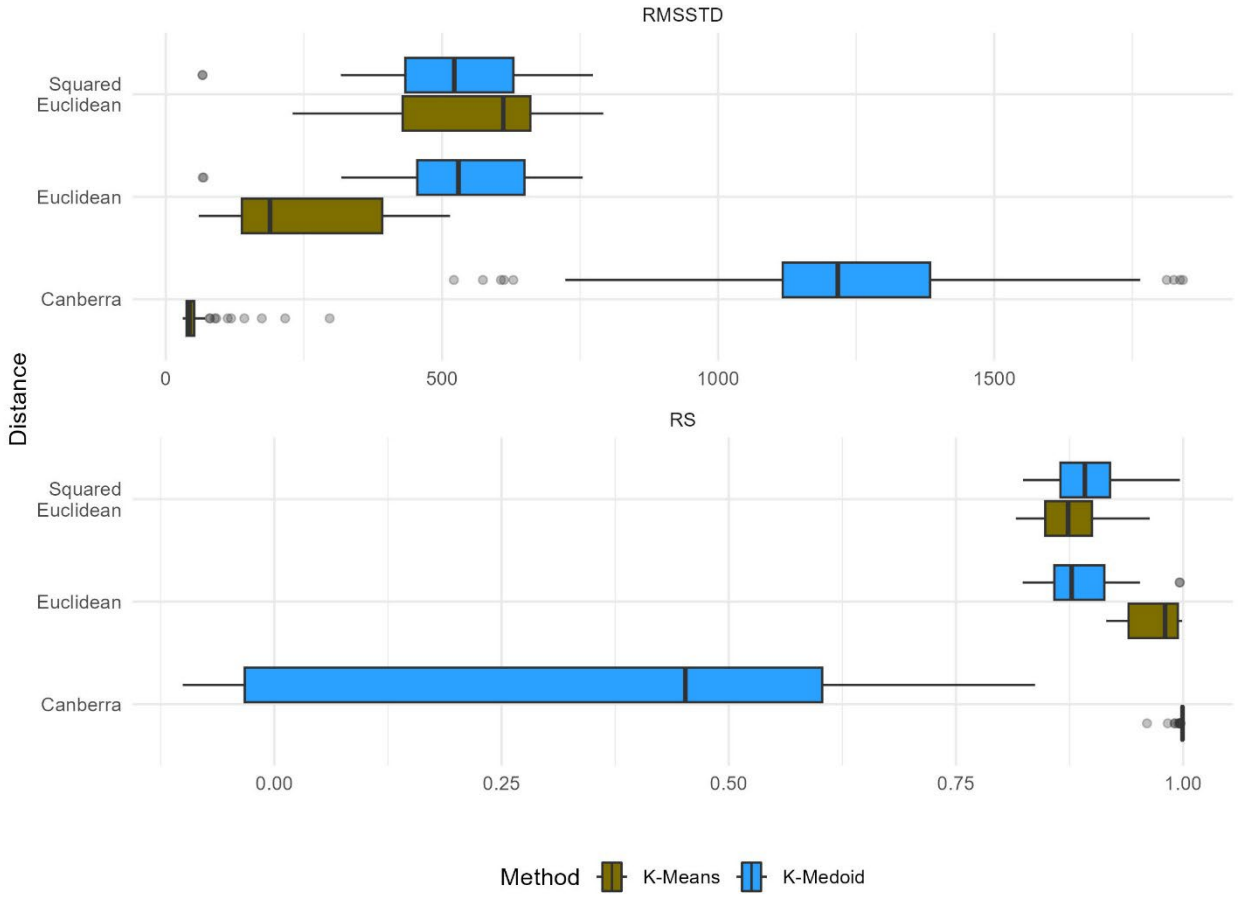


Figure 5. RMSSTD and RS Distribution Across 100 Repetitions

Figure 5 shows that, among the six clustering methods, across 100 repetitions, K-means Canberra consistently better than the other clustering methods.

Figure 6 shows the same distributions as shown in Figure 6. Figure 6 shows that K-means Canberra is indeed perform better across all 100 repetitions, consistently score lower RMSSTD and higher RS.

Smaller RMSSTD and high RS indicate that the formed clusters are homogeneous within clusters and well separated between clusters. K-medoid Canberra, on the other hand, has high RMSSTD and low RS overall, indicating that this method may not as well fit as the K-means Canberra for the current study. The other clustering methods are moderately well clustered, especially K-means Euclidean which is just slightly worse than K-means Canberra in some repetitions.

5.3 Proposed Improvements

Instead of using raw chili price time series, this study transforms each regional time series into model-based coefficient matrices using the VARD approach. This reduces noise and dimensionality while capturing essential dynamics

between chili types. Furthermore, six clustering scenarios were evaluated extensively over 100 repetitions, ensuring robustness in determining the optimal clustering method. Such model-based representation, repeated experiments, and extensive validation offer significant improvement over traditional clustering approaches that use raw price data directly. The RMSSTD and RS results confirm this improvement, particularly with the use of K-means and Canberra distance.

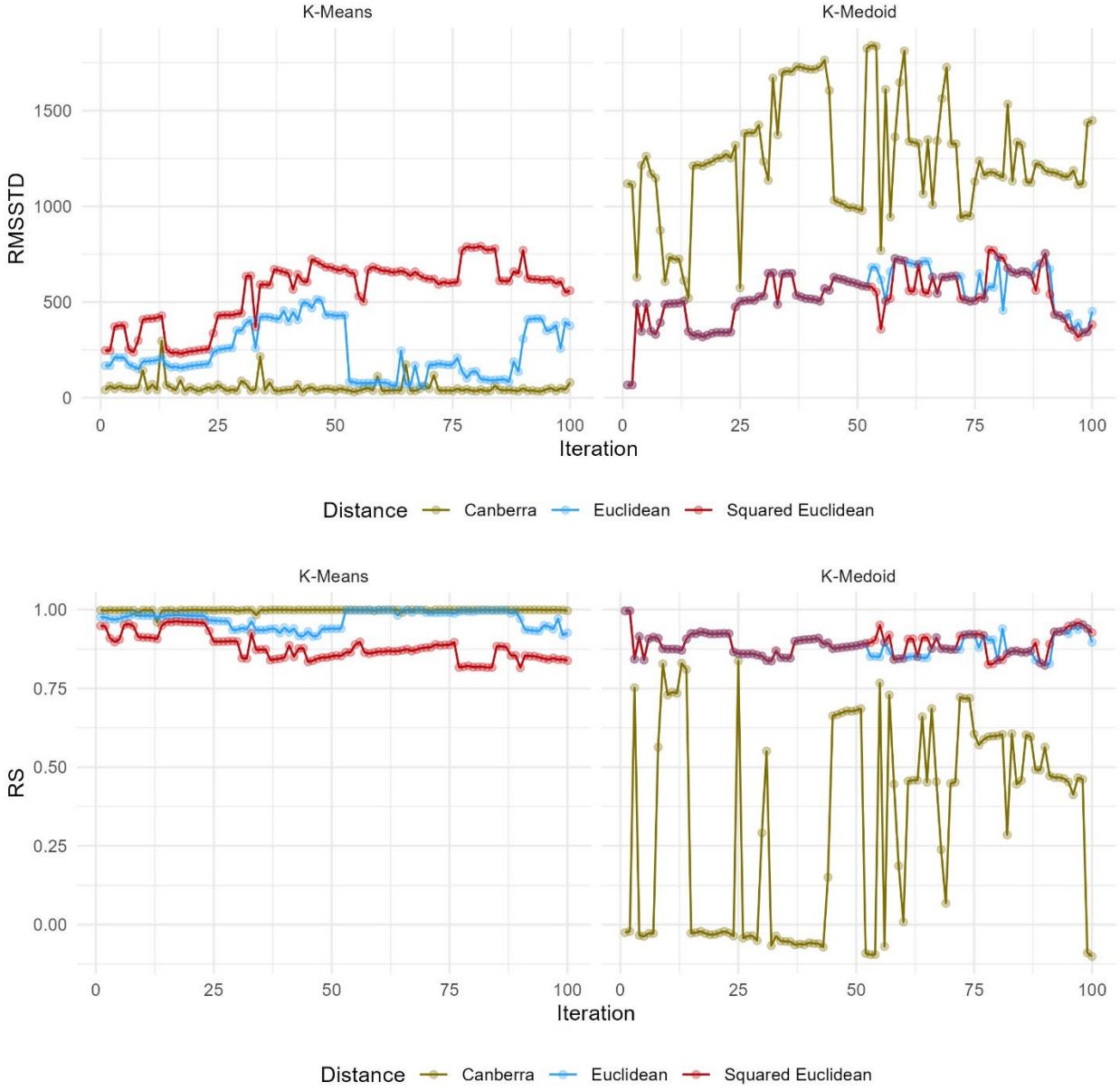


Figure 6. RMSSTD and RS of the Six Clustering Methods Across 100 repetitions

5.4 Validation

The validation process was conducted at two levels: model fitting and clustering performance. In the modeling phase, the dataset was split into training and validation sets. The validation sets consists of 5% of the whole data, which is equivalent to 55 observations per regency/city per commodity.

In the clustering phase, all six clustering methods were repeated 100 times. This allowed consistency checks of RMSSTD and RS values across runs. Figures 5 and 6 demonstrate that K-means with Canberra distance consistently outperformed other methods, confirming the reliability and robustness of the chosen clustering configuration.

6. Conclusion

This study aimed to cluster regencies and cities in Indonesia based on their chili price movement patterns using a model-based multivariate time series approach. The research successfully addressed this objective by first modeling the price dynamics of four chili commodities: big red chili, curly red chili, red bird's eye chili, and green bird's eye chili, using Vector Autoregressive with Differencing (VARD) models. The resulting coefficient matrices served as compact representations of the temporal and cross-variable relationships, reducing noise and dimensionality.

Clustering was performed across six scenarios, combining K-means and K-medoids algorithms with three distance measures: Euclidean, Squared Euclidean, and Canberra. Internal evaluation metrics, RMSSTD and R-Squared, were used to assess cluster compactness and separation. The results consistently identified the K-means algorithm with Canberra distance as the best-performing method across 100 repeated trials.

A key contribution of this study lies in the integration of model-based transformation with robust clustering evaluation, offering an alternative to raw-data clustering approaches. The use of VARD modeling as a preprocessing step improved the reliability and interpretability of the clusters, which may support more targeted price monitoring or policy interventions in the future.

Further research may explore the use of external validation metrics, inclusion of additional explanatory variables, or the extension of this framework into predictive clustering for early warning systems.

References

Ajeng, I., Afifah, N. and Nurdiyanto, H., Data mining clustering dalam pengelompokan buku perpustakaan menggunakan algoritma K-Means, *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, vol. 8, no. 3, pp. 802–814, 2023.

Anjani, I. D. and Bahtiar, A., Penerapan algoritma K-Means clustering untuk mengelompokkan penerima bantuan sosial tunai (BST) di Jawa Barat, *JATI (Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 3, pp. 2743–2747, 2024.

Bank Indonesia, Pusat Informasi Harga Pangan Strategis (PIHPS), *Bank Indonesia*. Available: https://www.bi.go.id/hargapangan/, Accessed on July 29, 2025.

Brockwell, P. J. and Davis, R. A., *Introduction to Time Series and Forecasting*, 3rd edition, Springer, New York, 2016.

Fadillah Mardianto, M. F., Siregar, N. R. A. A., Soewignjo, S., Putri, F. R., Prayogi, H., Imama, C., Amelia, D., Sediono, D. and Dewi, D. A., Time series clustering analysis for increases food commodity prices in Indonesia based on K-Means method, *Journal of Human, Earth, and Future*, vol. 5, no. 3, pp. 319–329, 2024.

Fathia, A. N., Rahmawati, R. and Tarno, Analisis klaster kecamatan di Kabupaten Semarang berdasarkan potensi desa menggunakan metode Ward dan Single Linkage, *Jurnal Gaussian*, vol. 5, no. 4, pp. 801–810, 2016.

Febrianti, D. R., Tiro, M. A. and Sudarmin, Metode vector autoregressive (VAR) dalam menganalisis pengaruh kurs mata uang terhadap ekspor dan impor di Indonesia, *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, vol. 3, no. 1, pp. 23–30, 2021.

Hasanah, I. N. and Sofro, A., Analisis cluster berdasarkan dampak ekonomi di Indonesia akibat pandemi COVID-19, *MATHunesa Jurnal Ilmiah Matematika*, vol. 10, no. 2, pp. 239–249, 2022.

Hasibuan, L. S. and Novialdi, Y., Prediksi harga minyak goreng curah dan kemasan menggunakan algoritme Long Short-Term Memory (LSTM), *Jurnal Ilmu Komputer Agri-Informatika*, vol. 9, no. 2, pp. 149–157, 2022.

Jollyta, D., Prihandoko, P., Priyanto, D., Hajjah, A. and Marlim, Y. N., Comparison of distance measurements based on k-numbers and its influence to clustering, *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 1, pp. 93–102, 2023.

Juliartha, M. A., Purnamasari, I. and Goejantoro, R., Penerapan automatic clustering pada fuzzy time series pada data wisatawan mancanegara Kalimantan Timur, *Eksponensial*, vol. 15, no. 2, pp. 110–118, 2024.

Lukas, A., Kairupan, A. N., Hendriadi, A., Arianto, A., Manalu, L. P., Sumarno, L., Munarso, J., Hadipernata, M., Elmatsani, H. M., Benyamin, B. O., Junaidi, A., Djafar, M. J., Elizabeth, R., Sahlan, Nasruddin, Astuti, P., Subandrio, Yohanes, H., Koeslulat, E. E., Susetyo, E. B., Ngudiwaluyo, S., Halik, A., Purwanto, W., Haryanto, G., Ramlah, S., Sjafrina, N., Budiyanto, A., Kailaku, S. I., Rienoviar, Yani, A., Yustiningsih, N., Adinegoro, H.,

Henanto, H., Layuk, P., Lintang, M., Joseph, G. H., Polakitan, D., Wahyudianto, H., Saptohadi, J., Budiarti, K. D., Hidayat, T., Ginting, J., Rasyid, A. and Polakitan, A., Fresh chili agribusiness: Opportunities and problems in Indonesia, Available: https://doi.org/10.5772/intechopen.112786, 2023.

Meiriza, A. and Ali, E., Perbandingan algoritma K-Means dan K-Medoids untuk pengelompokan program BPJS Ketenagakerjaan, *The Indonesian Journal of Computer Science*, vol. 12, no. 2, 2023. https://doi.org/10.33022/ijcs.v12i2.3184

Muflikh, Y. N., Smith, C., Brown, C. and Aziz, A. A., Analysing price volatility in agricultural value chains using systems thinking: A case study of the Indonesian chilli value chain, *Agricultural Systems*, vol. 192, 103179, 2021.

Nahdliyah, M. A., Widiharih, T. and Prahutama, A., Metode K-Medoids clustering dengan validasi silhouette index dan c-index, *Jurnal Gaussian*, vol. 8, no. 1, pp. 161–170, 2019.

Raeisi, M. and Sesay, A. B., A distance metric for uneven clusters of unsupervised K-Means clustering algorithm, *IEEE Access*, vol. 10, 2022.

Risnayah, S. and Sagala, L. O., Penerapan imputasi LOCF dan cross mean dalam pengisian data kosong pada curah hujan harian ARG, *Megasains*, vol. 14, no. 2, pp. 23–31, 2023. https://doi.org/10.46824/megasains.v14i2.138

Rohaeti, E., Sumertajaya, I. M., Wigena, A. H. and Sadik, K., MTSClust with handling missing data using VAR-Moving Average imputation, *Mathematics and Statistics*, vol. 11, no. 2, 2023. https://doi.org/10.13189/ms.2023.110201

Rohaeti, E., Sumertajaya, I. M., Wigena, A. H. and Sadik, K., Vector autoregressive-moving average imputation algorithm for handling missing data in multivariate time series, *IAENG International Journal of Computer Science*, vol. 5, no. 4, pp. 727–735, 2023.

Srinivasan, S., Modeling marketing dynamics using vector autoregressive (VAR) models, *Handbook of Market Research*, pp. 515–547, 2022. https://doi.org/10.1007/978-3-319-57413-4_10

Sundari, M. T., Darsono, Sutrisno, J. and Antriyandarti, E., Analysis of trade potential and factors influencing chili export in Indonesia, *Open Agriculture*, vol. 8, 2023. https://doi.org/10.1515/opag-2022-0205

Surya, R. and Tedjakusuma, F., Diversity of sambals, traditional Indonesian chili pastes, *Journal of Ethnic Foods*, 2022. https://doi.org/10.1186/s42779-022-00142-7

Van, J. C., Huang, W.-C., Anindita, R., Chang, W.-I., Yang, S.-H. and Shonhiwa, C., Price volatility of cayenne pepper and red chili pepper in Papua and Maluku provinces, Indonesia, *Scholars Journal of Economics, Business and Management*, vol. 4, no. 9, 2017. https://doi.org/10.36347/sjebm.2017.v04i09.002

Webb, A. J. and Kosasih, I. A., Analysis of price volatility in the Indonesia fresh chili market, 2011.

Yuditya, A., Hardjanto, A. and Sehabudin, U., Fluktuasi harga dan integrasi pasar cabai merah besar (Studi kasus: Pasar Induk Kramat Jati dan pasar eceran di DKI Jakarta), *Indonesian Journal of Agricultural Resource and Environmental Economics*, vol. 2, no. 1, pp. 1–13, 2023. https://doi.org/10.29244/ijaree.v2i1.50669

Biographies

Embay Rohaeti is a lecturer in the Mathematics Study Program at Universitas Pakuan, with research expertise in mathematical modeling. Her scholarly work spans various applied mathematics topics, including time series analysis, statistical modeling, and numerical methods. She has co-authored numerous publications in national and international journals, contributing to fields such as tuberculosis modeling, financial forecasting, and environmental statistics. Her research has been cited multiple times, reflecting her academic influence and collaboration with fellow researchers.

Ani Andriyati is a lecturer in the Mathematics Study Program at the Faculty of Mathematics and Natural Sciences, Universitas Pakuan, specializing in statistics. Her research focuses on statistical methods and their applications in various scientific and social contexts. She has published several scholarly articles in national and international journals, contributing to the advancement of statistical science. Her academic work has been cited by other researchers, reflecting her impact in the field. Ani is actively involved in teaching and mentoring students.

Muhammad Edy Rizal is a lecturer in the Data Science Study Program at the Faculty of Mathematics and Natural Sciences, Universitas Tadulako. He earned his master's degree in 2023 from the Department of Statistics and Data Science at IPB University, where he developed a strong foundation in statistical modeling and computational techniques. His research interests lie primarily in time series analysis and machine learning, with a particular focus on their applications in disaster prevention and mitigation. He has authored several studies exploring predictive models and data-driven approaches to enhance early warning systems and improve resilience against natural hazards. Through his academic work, he aims to contribute to the development of intelligent solutions for disaster risk management in Indonesia and beyond.