

Robust PCA-Based Multivariate Time Series Clustering for Weekly Indonesian Food Commodity Prices with Missing Data

Nina Valentika, I Made Sumertajaya*, Aji Hamim Wigena, Farit Mochamad Afendi

Statistics and Data Science Study Program, IPB University, Indonesia

Abstract This study proposes an integrated procedure for analyzing multivariate time series (MTS) of strategic food commodity prices across Indonesian regencies/cities, encompassing missing-data imputation, outlier-robust dimensionality reduction, MTS modeling, regional clustering, and cluster-prototype-based forecasting. Weekly price data for eight strategic food commodities across 70 regencies/cities in Indonesia are analyzed through imputation and preprocessing, dimensionality reduction using lag-1 autocovariance-based RPCA (RPCA lag-1), and time series modeling of the lag-1 RPCA principal component scores (RPCA scores) on the first and second principal components (PC1 and PC2) using a first-order vector autoregressive model (VAR(1)) with an intercept term. Regional clustering is performed using non-ensemble methods (K-means Euclidean and PAM Euclidean) to identify the best non-ensemble approach, and Random Forest–General Iterative Clustering (RF-GIC) is also employed as a benchmark. For each clustering result, namely K-means Euclidean as the selected best non-ensemble method and RF-GIC, cluster prototypes are constructed and used for forecasting. The two schemes are compared using a four-fold expanding-window evaluation (root mean square error (RMSE) and mean absolute deviation (MAD)) and a final mean absolute percentage error (MAPE) assessment at 4-step and 26-step horizons, evaluated at both the cluster-prototype and regency/city levels. Based on overall MAPE at the cluster-prototype level, both K-means Euclidean and RF-GIC achieve very accurate performance at the 4-step horizon and accurate performance at the 26-step horizon, with RF-GIC yielding lower overall MAPE than K-means Euclidean at both horizons. At the regency/city level, both methods fall into the accurate category at the 4-step horizon and the moderately accurate category at the 26-step horizon, with only a small difference in overall MAPE (slightly lower for K-means Euclidean). The decrease in accuracy is primarily driven by longer forecasting horizons and commodity-specific volatility, particularly for bird’s-eye chili and red chili.

Keywords Missing Data Imputation, Robust Principal Component Analysis, Multivariate Time Series Clustering, Non-Ensemble Clustering, Random Forest Clustering, Predictive Performance

AMS 2010 subject classifications 62M10, 62H25, 62H30

DOI: 10.19139/soic-2310-5070-3618

1. Introduction

This study is designed to address several challenges in the analysis of complex multivariate time series (MTS) data, particularly when the dataset involves many observational units, multiple variables, a long time span, and the presence of missing values and outliers. Missing values can hinder the analysis process because most methods require complete datasets to ensure accurate estimation. Accurate imputation of missing values can improve the accuracy of subsequent MTS analysis [1]. Therefore, imputation is often an essential step in the analytical workflow [2]. Handling missing data should account for the underlying missingness mechanism, which is commonly classified into Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) [3, 4]. In this study, missing values in the time series are first filled using linear interpolation implemented in the `imputeTS` package [5].

*Correspondence to: I Made Sumertajaya (Email: imsjaya@apps.ipb.ac.id). Statistics and Data Science Study Program, IPB University. Jalan Raya Dramaga, Kampus IPB Dramaga Bogor, West Java, Indonesia (16680).

In addition to missing data, the high dimensionality of multivariate data and the presence of outliers can reduce the accuracy of dimensionality reduction and subsequent modeling. Accordingly, this study employs Principal Component Analysis (PCA) to reduce dimensionality, thereby decreasing the number of parameters without discarding most of the data variability [6, 7]. For MTS, PCA can be implemented either through PCA of MTS or through lag-based classical PCA constructed from sample autocovariance or autocorrelation matrices [8, 9, 10, 11]. Nevertheless, PCA that relies on the sample covariance or correlation remains sensitive to outliers, because outliers can distort the covariance structure and bias the estimated principal components (PCs) toward non-representative directions [12, 13, 14]. To address this issue, several studies have developed robust covariance or correlation estimators as the basis for Robust PCA (RPCA) [9, 10, 14, 15]. This study applies lag-1 autocovariance-based RPCA (RPCA lag-1) for dimensionality reduction. The principal component scores obtained from RPCA lag-1 (RPCA scores) are subsequently modeled using an MTS model.

After the RPCA scores are obtained and modeled, clustering is employed to improve analytical efficiency when the number of observational units is large through model-based approaches. This study considers both non-ensemble methods, including K-means Euclidean (i.e., K-means with Euclidean distance) and Partitioning Around Medoids (PAM) Euclidean (i.e., PAM with Euclidean distance) [16], and a random-forest-based clustering approach that leverages bagging [17] and uses a proximity measure as the basis for clustering [18]. In this context, clustering methods for complex temporal data need to be adaptive, robust, and scalable [19]. The final stage is forecasting, which aims to predict future values from historical information to support decision-making and enhance analytical efficiency [20]. In this context, multivariate models can exploit inter-variable relationships to improve predictive accuracy [21].

The novelty of this study lies in formulating an integrated procedure for analyzing MTS of strategic food commodity prices at the regency/city level. The proposed pipeline combines missing-data imputation, outlier-robust dimensionality reduction via autocovariance-based RPCA, modeling of RPCA scores using an MTS model, coefficient-feature extraction to represent regional dynamics, clustering (non-ensemble and random-forest-based), and cluster-prototype-based forecasting. In contrast to the PCA–Vector Autoregressive (PCA–VAR) framework, which typically uses classical PCA as a pre-modeling step [22], and to the MTS Clustering (MTS-Clust) approach that emphasizes imputation using the Vector Autoregressive Imputation Method (VAR-IM) and the Vector Autoregressive Moving Average Imputation Method (VAR-IMMA) together with non-ensemble clustering [16, 23], this study places RPCA lag-1 as the initial step to obtain RPCA scores that are robust to outliers. The imputation is then refined through an iterative reconstruction procedure based on these RPCA scores. In addition to non-ensemble methods, the clustering stage is further extended by incorporating a random-forest-based approach combined with the iterative General Iterative Clustering (GIC) scheme [18], hereafter referred to as Random Forest–General Iterative Clustering (RF-GIC), to construct a general model for each cluster and evaluate forecasting performance.

A more efficient approach is needed to analyze weekly prices of multiple strategic food commodities across regencies/cities in Indonesia under missing-data and outlier conditions. However, integrated frameworks to support analysis under such characteristics remain relatively limited. Without such a framework, the analysis tends to be conducted by modeling each regency/city separately, particularly when the data involve many variables and contain missing values and outliers. Therefore, the main contribution of this study is the development of an integrated framework that combines imputation, dimensionality reduction, RPCA score modeling, clustering, and forecasting for weekly strategic food price data across regencies/cities under missing-data and outlier conditions. Within this framework, the study applies RPCA lag-1 to obtain a low-dimensional representation of MTS data that is robust to outliers, and then employs clustering to support cluster-prototype-based forecasting. Accordingly, this study aims to formulate and evaluate an integrated procedure encompassing missing-data imputation, outlier-robust dimensionality reduction, RPCA score modeling, regional clustering, and cluster-prototype-based forecasting for weekly strategic food commodity price data across regencies/cities in Indonesia under missing-data and outlier conditions.

2. Materials and Methods

The research workflow is shown in Figure 1.

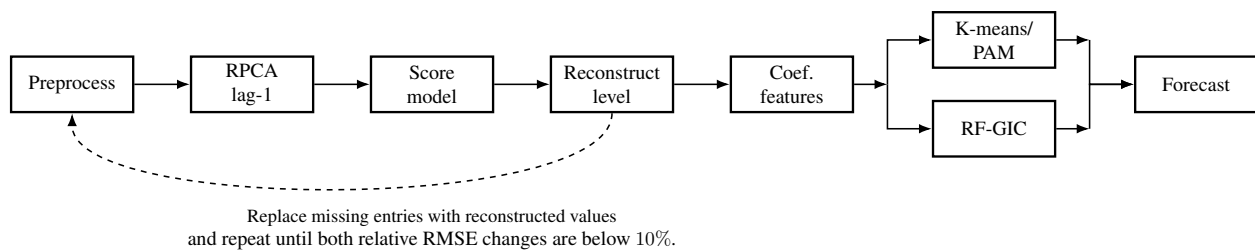


Figure 1. Research workflow.

The analysis begins with initial imputation and data preprocessing, followed by dimensionality reduction using RPCA lag-1. The RPCA scores are modeled using a Vector Autoregressive model (VAR), VAR in Differences (VARD), or a Vector Error Correction Model (VECM), and are then used to reconstruct the series on the original (level) scale. The missing entries are then updated using the reconstructed values, and the preprocessing, RPCA lag-1, score modeling, and reconstruction steps are repeated until the relative change in the root mean square error (RMSE) satisfies the convergence criterion. Next, the model coefficients are extracted as features for regional clustering using non-ensemble methods (K-means Euclidean and PAM Euclidean) as well as RF-GIC. For both clustering results (the selected best non-ensemble method and RF-GIC), cluster prototypes are constructed and forecasting performance is evaluated.

2.1. Data and preprocessing

This study used weekly price data for eight strategic food commodities, namely rice, chicken eggs, shallots, garlic, red chili, bird's-eye chili, cooking oil, and granulated sugar. The data were obtained from the National Strategic Food Price Information Center (PIHPS) of Bank Indonesia [24] for the period from 21 July 2021 to 26 February 2025, comprising 189 weekly observations. Of the 74 regencies/cities available, the analysis focused on 70 regencies/cities because four regions exhibited insufficient data completeness over the observation period. In addition, two other commodities were excluded due to incomplete and inconsistent availability across locations. Missing values were identified and initially imputed using linear interpolation. Stationarity was assessed using the Augmented Dickey–Fuller (ADF) test at the 10% significance level; differencing was applied iteratively until stationarity was achieved, followed by mean centering. Consistent with [11], who analyzed the same dataset, the level series are generally non-stationary and become stationary after first-order differencing at the 10% significance level, seasonal patterns are not dominant, and there are indications of additive outliers.

2.2. RPCA lag-1

This study employed RPCA lag-1 based on prior comparative evidence obtained from the same dataset, namely weekly price data of strategic food commodities observed across 70 regencies/cities in Indonesia [11]. In that study, Classical PCA and RPCA were evaluated at lags 1 through 7, and lag 1 was selected as the optimal lag because it was the smallest lag that consistently showed differences in the evaluation results of the two methods in terms of the minimum number of principal components and the proportion of cumulative explained variance. At that lag, Classical PCA lag-1 and RPCA lag-1 generally required one to two principal components to achieve at least 85% cumulative explained variance, whereas PCA of MTS tended to require two to three principal components. The simulation results also showed that RPCA was the most efficient and effective method, with the optimal lag consistently identified as lag 1. These findings are consistent with [10], who showed that the strongest correlations in multivariate time series generally occur at small lags. Therefore, lag 1 was considered adequate as the basis for dimensionality reduction in this study. RPCA lag-1 was obtained via eigenvalue decomposition of the total sample

autocovariance matrix,

$$\widehat{M} = \sum_{h=1}^{h_0} \widehat{\Gamma}^*(h) \widehat{\Gamma}^*(h)^\top, \quad (1)$$

with $h_0 = 1$, where $\widehat{\Gamma}^*(h)$ denotes the robust sample autocovariance matrix based on the $Q_n(\cdot)$ scale estimator [10]. The number of RPCA principal components was determined based on the majority of regencies/cities achieving a minimum cumulative explained-variance proportion of 85%.

2.3. Modeling RPCA Scores and Lag Selection

The RPCA scores were modeled using VAR, VARD, or VECM. Model specification was determined based on stationarity testing and the Johansen cointegration test (trace and maximum eigenvalue statistics). The maximum lag was set to the cube root of the number of observations used in model estimation [25], whereas the optimal lag was selected using the Schwarz Information Criterion (SC). To ensure consistent coefficient-feature dimensionality across regions, the final model class and lag order were determined by the modal (most frequently selected) specification across all regencies/cities.

2.4. Reconstruction to the Level Scale

After the RPCA scores were modeled using the selected MTS model (VAR, VARD, or VECM) with lag order ℓ , the fitted scores S_{fitted} were obtained. Data reconstruction followed the standard PCA formulation [26, 27]:

$$\widehat{X} = S_{\text{fitted}} V_{\text{reduced}}^T, \quad (2)$$

where V_{reduced} is the loading matrix for the k selected principal components. The reconstructed series were then returned to the final scale by undoing mean centering:

$$Y_\Delta = \widehat{X} + \mu_\Delta, \quad (3)$$

where μ_Δ is the mean vector on the final (differenced) scale. If differencing of order d was applied during preprocessing, Y_Δ was integrated d times using the corresponding initial conditions to obtain the level (original) scale.

2.5. Iterative Reconstruction and Convergence Criterion

The imputation did not stop at the initial interpolation stage, but was refined through an iterative reconstruction procedure based on the RPCA scores and the fitted scores obtained from the modeling results. At each iteration, the values at the missing-data positions were updated using the reconstruction results from the previous iteration. The updated data were then preprocessed again, and the entire modeling pipeline—from RPCA lag-1 to time series model estimation for the RPCA scores—was reapplied to the updated data to obtain fitted scores and new reconstruction results on the level scale for updating the missing values in the subsequent iteration. At iteration i for regency/city j , inter-iteration changes were quantified using RMSE computed over the missing entries and over all data elements.

Let $X^{(i,j)} \in \mathbb{R}^{T \times p}$ denote the reconstructed matrix at iteration i for regency/city j , and let M_j be the set of indices corresponding to cells that were missing in the original data for regency/city j , with $m_j = |M_j|$. The RMSE over the missing entries was defined as

$$\text{RMSE}_{\text{miss}}^{(i,j)} = \sqrt{\frac{1}{m_j} \sum_{(t,r) \in M_j} \left(X_{t,r}^{(i,j)} - X_{t,r}^{(i-1,j)} \right)^2}, \quad (4)$$

where $t = 1, \dots, T$ indexes time and $r = 1, \dots, p$ indexes commodities.

The RMSE over all elements was defined as

$$\text{RMSE}_{\text{all}}^{(i,j)} = \sqrt{\frac{1}{Tp} \sum_{t=1}^T \sum_{r=1}^p \left(X_{t,r}^{(i,j)} - X_{t,r}^{(i-1,j)} \right)^2}. \quad (5)$$

RMSE values were then averaged across regencies/cities at iteration i :

$$\overline{\text{RMSE}}_{\text{miss},i} = \frac{1}{n_{\text{miss}}} \sum_{j \in J_{\text{miss}}} \text{RMSE}_{\text{miss}}^{(i,j)}, \quad (6)$$

$$\overline{\text{RMSE}}_{\text{all},i} = \frac{1}{n} \sum_{j=1}^n \text{RMSE}_{\text{all}}^{(i,j)}, \quad (7)$$

where n denotes the total number of regencies/cities, J_{miss} is the set of regencies/cities with at least one missing value, and $n_{\text{miss}} = |J_{\text{miss}}|$ is the number of regencies/cities with at least one missing value. The stopping rule followed the relative iterative error criterion [28], defined as the relative change in RMSE between successive iterations (for $i \geq 2$):

$$\overline{\text{RMSE}}_{\text{miss,rel},i} = \frac{|\overline{\text{RMSE}}_{\text{miss},i} - \overline{\text{RMSE}}_{\text{miss},i-1}|}{\overline{\text{RMSE}}_{\text{miss},i}}, \quad (8)$$

$$\overline{\text{RMSE}}_{\text{all,rel},i} = \frac{|\overline{\text{RMSE}}_{\text{all},i} - \overline{\text{RMSE}}_{\text{all},i-1}|}{\overline{\text{RMSE}}_{\text{all},i}}. \quad (9)$$

Iterations were terminated at the first iteration that simultaneously satisfied

$$\overline{\text{RMSE}}_{\text{miss,rel},i} < 0.10, \quad (10)$$

and

$$\overline{\text{RMSE}}_{\text{all,rel},i} < 0.10. \quad (11)$$

The dataset obtained at the iteration satisfying these criteria was treated as the stable data for subsequent analyses. A 10% threshold was used as the stopping tolerance in the iterative reconstruction procedure. Sensitivity to alternative convergence thresholds, comparisons with other modern imputation methods, and evaluation of imputation accuracy using RMSE and mean absolute error (MAE) were beyond the scope of this study and may be examined in future research.

2.6. Coefficient features and clustering

For each regency/city, the model coefficients estimated from the RPCA-score models were extracted as a feature vector representing the region's temporal dynamics, including the intercept term when applicable. The tendency of the data to form clusters was assessed using the Hopkins statistic; values close to 1 indicate that the data are likely to be clusterable [29]. When clusterability was indicated, the number of clusters was selected based on the Silhouette Score [30], with the Dunn Index and the interpretability of cluster profiles considered as additional criteria. The selected number of clusters was then used consistently across all clustering methods, namely the non-ensemble methods (K-means Euclidean and PAM Euclidean) and RF-GIC. The best non-ensemble method was chosen based on the Silhouette Score and Dunn Index at the selected number of clusters. In RF-GIC, random forest was used to construct a proximity matrix, which was transformed into a dissimilarity matrix, and PAM was then applied iteratively until the clusters converged. The implementation was carried out using the GIC package in R via the GIC() function [18].

2.7. *Prototype and forecasting evaluation*

For the selected best non-ensemble method and RF-GIC, a cluster prototype was constructed as the mean time series of the stable data from cluster members. Prototype-model specification was determined based on the ADF test at the 10% significance level and the Johansen cointegration test (trace and maximum eigenvalue statistics) at the 5% significance level. The lag order was initially selected using the SC criterion and was then increased when necessary until the residuals showed no autocorrelation [31]. Forecasting evaluation was conducted using a four-fold expanding-window scheme by comparing the fitting and forecasting performance of VAR, VARD, and VECM using RMSE and mean absolute deviation (MAD), while the Johansen test was performed on the training data within each fold. The four folds used cumulative training sets and the subsequent 11 observations as test sets: Fold 1 (1 to 145; 146 to 156), Fold 2 (1 to 156; 157 to 167), Fold 3 (1 to 167; 168 to 178), and Fold 4 (1 to 178; 179 to 189). Subsequently, the best model was estimated using the full observation period and was assessed using the Portmanteau test, Autoregressive Conditional Heteroskedasticity (ARCH) test, and model-stability diagnostics. A model was considered stable when all roots lay inside the unit circle, such that the resulting Impulse Response Functions (IRFs) and Forecast Error Variance Decompositions (FEVDs) were regarded as valid [32]. Final evaluation used mean absolute percentage error (MAPE) at a short-term horizon of 4 steps (5 March 2025 to 26 March 2025) and a long-term horizon of 26 steps (5 March 2025 to 27 August 2025), evaluated at both the cluster-prototype and regency/city levels. MAPE accuracy categories were adapted from [33]: MAPE below 10% (very accurate), between 11% and 20% (accurate), between 21% and 50% (moderately accurate), and 51% or higher (inaccurate).

3. Results and Discussion

3.1. *Stable reconstructed data and convergence*

Reconstruction convergence was assessed using the relative change in RMSE between successive iterations computed for (i) the missing entries and (ii) all data elements, with a 10% threshold for each criterion. The sixth iteration was the first stable iteration because it satisfied both criteria simultaneously. Therefore, the reconstruction from the sixth iteration was used as the stable dataset for subsequent analyses.

3.2. *Characteristics of the stable data*

The stable data obtained from the reconstruction–imputation process served as the basis for all subsequent analyses. The characteristics of the stable data were evaluated in terms of stationarity, indications of seasonality, indications of additive outliers, and the number of RPCA principal components.

3.2.1. Stationarity of the stable data. The percentages of stationary (S) and non-stationary (NS) series based on the ADF test applied to the stable data (560 weekly series) are reported in Table 1.

Table 1. Percentages of stationary (S) and non-stationary (NS) series based on the ADF test for the stable data (560 weekly series)

Data type	S (%)	NS (%)
I(0)	44.29	55.71
I(1)	100	0
I(1) then MC	100	0

Note: I(0) denotes the level series; I(1) denotes the first-order differenced series; MC denotes mean centering. The ADF test was conducted at the 10% significance level.

Table 1 indicates that, at the 10% significance level, most level series are non-stationary, whereas after first-order differencing all series become stationary. This finding is consistent with [11], who analyzed the same dataset (using the initial imputation).

3.2.2. Seasonality indications in the stable data. Seasonality indications in the stable data were evaluated using the QS test as implemented in the `seastests` package [34]. The percentages of series indicated as seasonal (M) and not indicated as seasonal (TM) based on the QS test applied to the stable data (560 weekly series) are reported in Table 2.

Table 2. Percentages of series indicated as seasonal (M) and not indicated as seasonal (TM) based on the QS test for the stable data (560 weekly series)

Data type	M (%)	TM (%)
I(0)	1.25	98.75
I(1)	1.96	98.04
I(1) then MC	2.14	97.86

Note: I(0) denotes the level series; I(1) denotes the first-order differenced series; MC denotes mean centering. Seasonality indication was determined using the QS test at the 1% significance level.

Table 2 shows that, at the 1% significance level, the proportion of time series exhibiting seasonality indications in the stable data is relatively small. This finding is consistent with [11], who analyzed the same dataset (using the initial imputation), suggesting that seasonal effects are not dominant in the stable data.

3.2.3. Indications of additive outliers. Consistent with [10, 11], Figure 2 presents the autocovariance function (ACOVF) of the stable data at lag 1, which serves as the basis for Classical PCA and RPCA. Pematang Siantar City is shown as an illustration because it had the highest proportion of missing values in the same dataset (under the initial imputation) [11].

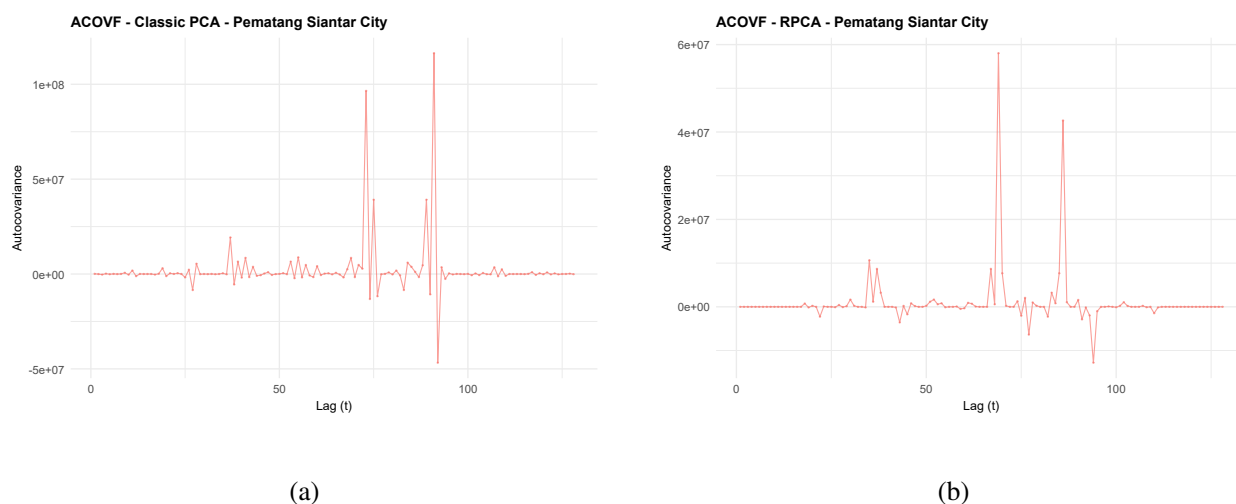


Figure 2. ACOVF of the stable data after preprocessing at lag 1 as the basis for Classical PCA (a) using the classical estimator and RPCA (b) using the robust estimator.

Figure 2(a) and Figure 2(b) display pronounced spikes in the autocovariances at several lags, indicating that the stable data exhibit additive outliers.

3.2.4. Number of RPCA principal components. Based on the stable data, RPCA lag-1 required one to two principal components across all regencies/cities to achieve at least 85% cumulative explained variance, consistent with [11]. To maintain comparability across regencies/cities in the modeling and clustering stages, this study used two RPCA scores for all regencies/cities. However, for regencies/cities in which a single principal component was already sufficient to explain at least 85% of the variance, the inclusion of the second component may have introduced less

informative variation and added noise to the feature vector. This is acknowledged as one of the limitations of the study.

3.3. *RPCA lag-1 results and modeling of RPCA scores*

The ADF test indicated that the RPCA scores were stationary in levels at the 10% significance level for all regencies/cities. Because the ADF test indicated that both RPCA scores were stationary in levels at the 10% significance level, the two scores were modeled using VAR rather than VARD or VECM. Lag-order selection based on the SC criterion chose lag 1 for 69 of 70 regencies/cities (98.57%), whereas Pontianak City selected lag 2 (1.43%). To maintain uniformity in the dimension of the coefficient feature vectors at the clustering stage, all regencies/cities were specified using a first-order vector autoregressive model (VAR(1)) with an intercept term. For Pontianak City, this specification may not fully reflect the lag-2 structure selected by the SC criterion, and may therefore affect the coefficient vector and the subsequent clustering results. This is acknowledged as one of the limitations of the study. Accordingly, the general VAR(1) specification used in this study is formulated as follows.

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{e}_t, \quad (12)$$

where $\mathbf{y}_t$ denotes the two-dimensional vector of RPCA scores at time t . The coefficients $\mathbf{A}_1$ and $\mathbf{c}$ were used as features, and all models satisfied the stability condition based on the roots of the characteristic polynomial.

3.4. *Coefficient features and clustering results (non-ensemble and RF-GIC)*

The coefficients from the VAR(1) model with an intercept term fitted to the two RPCA scores were extracted as a six-dimensional feature vector for each regency/city. A Hopkins statistic of 0.9909 indicated that the data were highly clusterable. The number of clusters was set to three across all methods based on internal validation results and the interpretability of cluster profiles.

The list of regions is as follows: (1) Banda Aceh City, (2) Lhokseumawe City, (3) Meulaboh City, (4) Denpasar City, (5) Tangerang City, (6) Bengkulu City, (7) Yogyakarta City, (8) Jakarta Pusat City, (9) Gorontalo City, (10) Bungo Regency, (11) Jambi City, (12) Bandung City, (13) Bekasi City, (14) Bogor City, (15) Cirebon City, (16) Depok City, (17) Sukabumi City, (18) Tasikmalaya City, (19) Banyumas Regency, (20) Cilacap Regency, (21) Semarang City, (22) Surakarta (Solo) City, (23) Tegal City, (24) Banyuwangi Regency, (25) Jember Regency, (26) Blitar City, (27) Kediri City, (28) Madiun City, (29) Malang City, (30) Probolinggo City, (31) Surabaya City, (32) Pontianak City, (33) Banjarmasin City, (34) Palangkaraya City, (35) Sampit City, (36) Balikpapan City, (37) Samarinda City, (38) Bulungan Regency, (39) Tarakan City, (40) Pangkal Pinang City, (41) Batam City, (42) Bandar Lampung City, (43) Ternate City, (44) Ambon City, (45) Tual City, (46) Mataram City, (47) Sumba Timur Regency, (48) Kupang City, (49) Maumere City, (50) Manokwari Regency, (51) Sorong City, (52) Jayapura City, (53) Dumai City, (54) Pekanbaru City, (55) Tembilahan City, (56) Mamuju City, (57) Makassar City, (58) Parepare City, (59) Palu City, (60) Kendari City, (61) Kotamobagu City, (62) Manado City, (63) Bukittinggi City, (64) Padang City, (65) Palembang City, (66) Gunung Sitoli City, (67) Medan City, (68) Padang Sidempuan City, (69) Pematang Siantar City, and (70) Sibolga City.

Cluster memberships (including cluster sizes) are summarized in Table 3. Table 3 shows that cluster assignments differed across methods. Under K-means Euclidean, Padang Sidempuan City (index 68) was placed in its own cluster, whereas PAM Euclidean and RF-GIC produced more balanced cluster sizes. Based on the profiles of the RPCA scores on the first principal component (PC1) and the second principal component (PC2), the three clusters produced by K-means Euclidean, PAM Euclidean, and RF-GIC were primarily differentiated by the mean score levels and the intensity of fluctuations. Under the three-cluster scheme, RF-GIC yielded the best cluster quality according to the Silhouette Score and Dunn Index (0.489 and 0.641), followed by K-means Euclidean (0.178 and 0.148) and PAM Euclidean (0.145 and 0.072). Therefore, K-means Euclidean was selected as the best non-ensemble method to be compared with RF-GIC in the prototype construction and forecasting stages. Cluster prototypes for K-means Euclidean and RF-GIC were constructed as the mean time series of the stable data from cluster members at each period. The inter-cluster movement patterns were broadly similar; however, separation was mainly observed in price levels and spike amplitudes, particularly for volatile commodities such as bird's-eye chili and red chili.

Table 3. Cluster memberships (region indices) and cluster sizes for K-means Euclidean, PAM Euclidean, and RF-GIC

Method	Cluster	Cluster size	Cluster members (region indices)
K-means Euclidean	1	1	68
	2	38	4, 5, 7, 8, 9, 10, 12, 13, 14, 16, 17, 18, 22, 25, 26, 28, 29, 31, 32, 33, 34, 35, 40, 42, 43, 46, 48, 49, 50, 55, 56, 58, 59, 60, 61, 63, 65, 70
	3	31	1, 2, 3, 6, 11, 15, 19, 20, 21, 23, 24, 27, 30, 36, 37, 38, 39, 41, 44, 45, 47, 51, 52, 53, 54, 57, 62, 64, 66, 67, 69
PAM Euclidean	1	25	1, 2, 3, 11, 15, 20, 27, 35, 36, 37, 38, 39, 44, 45, 47, 51, 52, 53, 54, 62, 65, 66, 67, 69, 70
	2	23	4, 7, 9, 10, 18, 19, 21, 26, 28, 31, 33, 40, 41, 42, 43, 48, 49, 50, 55, 57, 61, 63, 64
	3	22	5, 6, 8, 12, 13, 14, 16, 17, 22, 23, 24, 25, 29, 30, 32, 34, 46, 56, 58, 59, 60, 68
RF-GIC	1	27	1, 5, 6, 9, 11, 16, 17, 21, 24, 25, 26, 36, 37, 38, 39, 40, 44, 50, 52, 54, 56, 57, 58, 59, 62, 67, 70
	2	21	2, 3, 8, 10, 12, 14, 15, 18, 22, 23, 30, 31, 32, 34, 35, 47, 49, 61, 63, 66, 68
	3	22	4, 7, 13, 19, 20, 27, 28, 29, 33, 41, 42, 43, 45, 46, 48, 51, 53, 55, 60, 64, 65, 69

To provide a geographic view of the RF-GIC results, Figure 3 presents the geographic distribution of the 70 study regencies/cities in Indonesia across the three RF-GIC clusters.

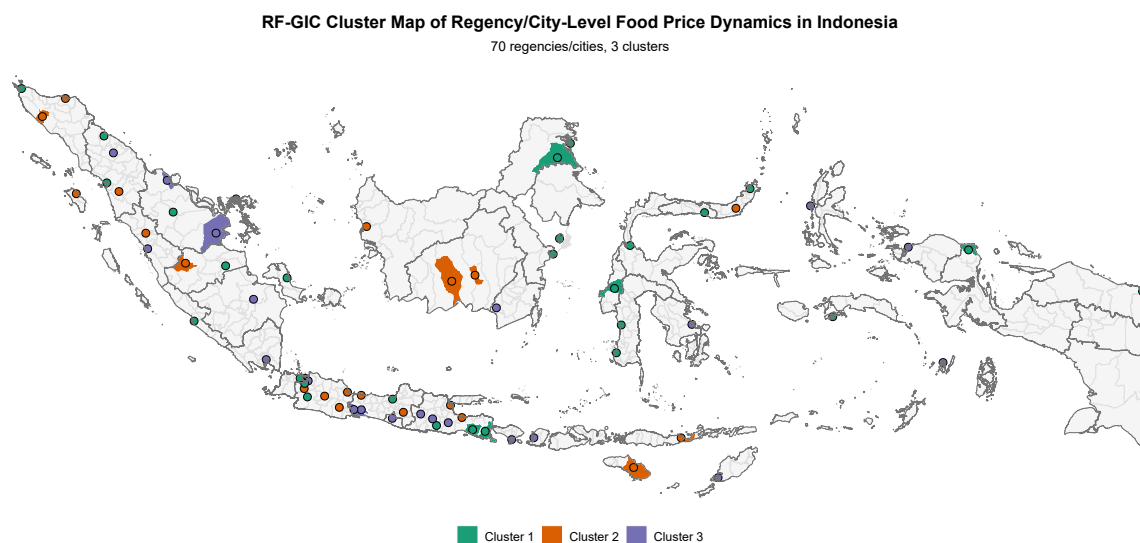


Figure 3. Geographic distribution of the 70 study regencies/cities in Indonesia across the three RF-GIC clusters. Colored circular markers were added so that study regencies/cities with small geographic areas remain easily identifiable on the national map. Regency/city administrative boundaries were obtained from Badan Informasi Geospasial [35]. Only the 70 study regencies/cities are shown with colored markers, whereas the other regencies/cities are displayed only as background boundaries.

Figure 3 shows that the three RF-GIC clusters were not formed solely on the basis of island groupings. The members of each cluster were distributed across more than one island, although in some neighboring areas

there remained a tendency for geographically adjacent regencies/cities to belong to the same cluster. This pattern indicates that the clustering results were not determined entirely by geographic proximity, but were also consistent with similarities in price dynamics across regencies/cities, including regencies/cities that are geographically far apart.

To complement the interpretation of the RF-GIC results, Table 4 presents the RF-GIC cluster profiles based on the mean and dispersion of the RPCA scores, the characteristics of the two most volatile commodities, and the geographic composition. This profile focuses on red chili and bird's-eye chili because these two commodities exhibited the highest price volatility in the empirical results.

Table 4. RF-GIC cluster profiles based on the mean and dispersion of the RPCA scores, the characteristics of the two most volatile commodities, and the geographic composition

Cluster	Mean PC1*	SD PC1**	Mean PC2*	SD PC2**	Mean red chili price	SD red chili price ***	Mean bird's-eye chili price	SD bird's-eye chili price ***	Number of provinces represented ****	Geographic summary and price characteristics
1	0.00	8916.01	0.00	6348.91	37616.26	16942.36	43379.09	18678.84	20	The most geographically dispersed cluster, with the highest mean bird's-eye chili price.
2	0.00	7167.76	0.00	5717.40	37625.25	16891.15	39616.32	16362.13	12	Relatively more concentrated in western Indonesia, with the lowest volatility for both chili commodities.
3	0.00	9941.76	0.00	6808.08	38666.22	17753.59	41823.95	21242.36	18	A cross-regional cluster, with the highest mean red chili price and the highest volatility for both chili commodities.

Note: *Because the RPCA scores were mean-centered, the mean values of PC1 and PC2 over time were close to zero for all clusters.

The standard deviations (SDs) of PC1 and PC2 represent the dispersion of the RPCA scores within each cluster. *The standard deviations of red chili and bird's-eye chili prices were used as measures of price volatility within each cluster. ****The column for the number of provinces represented indicates the number of different provinces from which the cluster members originated.

Table 4 shows that the three RF-GIC clusters differed in terms of geographic distribution, RPCA score dispersion, and the price characteristics of the two most volatile commodities. Although the mean values of PC1 and PC2 were close to zero for all clusters because the RPCA scores were mean-centered, inter-cluster differences remained evident in the dispersion of the RPCA scores, the average price levels, and the price volatility of red chili and bird's-eye chili. Cluster 1 had the widest geographic coverage, as reflected by the largest number of provinces represented, and also had the highest mean bird's-eye chili price. Cluster 2 was relatively more concentrated in western Indonesia and had the lowest RPCA score dispersion as well as the lowest price volatility for red chili and bird's-eye chili, and therefore may be regarded as the relatively most stable cluster. Cluster 3 exhibited the most fluctuating dynamics, with the highest SD of PC1 and SD of PC2, the highest mean red chili price, and the highest volatility for both chili commodities. Overall, this cluster profile strengthens the clustering results by showing that the RF-GIC clusters differed not only in geographic composition but also in average price levels, RPCA score dispersion, and commodity volatility.

3.5. Prototype forecasting evaluation

In the expanding-window evaluation, the performance of VAR, VARD, and VECM was compared using the average RMSE and MAD across folds on both the training and test sets for the K-means Euclidean and RF-GIC prototypes. After first-order differencing, all series were stationary at the 10% significance level, and the Johansen test did not indicate cointegration at the 5% significance level. Accordingly, VARD was the relevant specification for all clusters under both K-means Euclidean and RF-GIC. The maximum lag was set to 5, and the lag order was initially selected using the SC criterion and then increased when necessary until the residuals showed no autocorrelation

[31]. The final specification for the K-means Euclidean prototypes was VARD(1) for all clusters, whereas for the RF-GIC prototypes, Clusters 1 and 2 used VARD(1) and Cluster 3 used VARD(2). Final evaluation was conducted using MAPE on an out-of-sample period for 4-step and 26-step horizons, with calculations restricted to time points and variables for which actual values were available.

3.5.1. Forecasting evaluation at the cluster-prototype level. The distribution of MAPE accuracy categories at the cluster-prototype level for the 24 commodity–cluster combinations at the 4-step and 26-step horizons is reported in Table 5.

Table 5. Distribution of MAPE accuracy categories at the cluster-prototype level for 24 commodity–cluster combinations at the 4-step and 26-step horizons

Accuracy category	K-means Euclidean (4 steps)	K-means Euclidean (26 steps)	RF-GIC (4 steps)	RF-GIC (26 steps)
Very accurate	19 (79.17%)	12 (50.00%)	19 (79.17%)	12 (50.00%)
Accurate	3 (12.50%)	6 (25.00%)	4 (16.67%)	5 (20.83%)
Moderately accurate	2 (8.33%)	5 (20.83%)	1 (4.17%)	7 (29.17%)
Inaccurate	0 (0.00%)	1 (4.17%)	0 (0.00%)	0 (0.00%)
Total	24 (100.00%)	24 (100.00%)	24 (100.00%)	24 (100.00%)

Note: Overall MAPE at the cluster-prototype level was 5.69% for K-means Euclidean and 5.20% for RF-GIC at the 4-step horizon, and 13.95% for K-means Euclidean and 11.76% for RF-GIC at the 26-step horizon. Overall MAPE was computed using 96 evaluation points at the 4-step horizon and 624 evaluation points at the 26-step horizon.

Based on Table 5, at the cluster-prototype level, forecasting with both K-means Euclidean and RF-GIC fell into the very accurate category at the 4-step horizon, with RF-GIC yielding a slightly lower overall MAPE than K-means Euclidean. At the 26-step horizon, accuracy decreased to the accurate category. Nevertheless, at the cluster-prototype level, RF-GIC continued to produce a lower overall MAPE and did not yield any commodity–cluster combinations in the inaccurate category. Overall, at the cluster-prototype level, RF-GIC exhibited relatively better performance at both short and long horizons.

3.5.2. Forecasting evaluation at the regency/city level. The distribution of MAPE accuracy categories at the regency/city level (560 combinations) for the 4-step and 26-step horizons is reported in Table 6.

Table 6. Distribution of MAPE accuracy categories at the regency/city level (560 combinations) for the 4-step and 26-step horizons

Accuracy category	K-means Euclidean (4 steps)	K-means Euclidean (26 steps)	RF-GIC (4 steps)	RF-GIC (26 steps)
Very accurate	326 (58.21%)	247 (44.11%)	323 (57.68%)	242 (43.21%)
Accurate	130 (23.21%)	114 (20.36%)	130 (23.21%)	115 (20.54%)
Moderately accurate	91 (16.25%)	150 (26.79%)	94 (16.79%)	154 (27.50%)
Inaccurate	13 (2.32%)	49 (8.75%)	13 (2.32%)	49 (8.75%)
Total	560 (100.00%)	560 (100.00%)	560 (100.00%)	560 (100.00%)

Note: Overall MAPE at the regency/city level was 12.35% for K-means Euclidean and 12.44% for RF-GIC at the 4-step horizon, and 21.16% for K-means Euclidean and 21.23% for RF-GIC at the 26-step horizon. Overall MAPE was computed using 2,240 evaluation points at the 4-step horizon. At the 26-step horizon, 14,457 evaluation points were used because 103 points lacked actual values out of the 14,560 theoretical evaluation points.

Based on Table 6, the regency/city-level evaluation shows that K-means Euclidean and RF-GIC produced nearly identical distributions of MAPE accuracy categories. At the 4-step horizon, overall MAPE fell into the accurate category, whereas at the 26-step horizon accuracy decreased to the moderately accurate category. Overall, K-means Euclidean yielded a slightly lower overall MAPE than RF-GIC. Inaccuracy mainly occurred for more volatile

commodities, particularly bird's-eye chili and red chili, and at the 26-step horizon it could also occur for shallots, although only in a very limited number of cases.

To provide a more detailed view of the mean MAPE and the distribution of inaccurate cases by commodity, Table 7 presents the commodity-level profile at the regency/city level for K-means Euclidean and RF-GIC at the 4-step and 26-step horizons.

Table 7. Commodity-level profile of mean MAPE and the distribution of inaccurate cases at the regency/city level for K-means Euclidean and RF-GIC at the 4-step and 26-step horizons

Commodity	Mean MAPE (K4)	Mean MAPE (G4)	Inaccurate (K4)	Inaccurate (G4)	Mean MAPE (K26)	Mean MAPE (G26)	Inaccurate (K26)	Inaccurate (G26)
Shallots	19.37	19.30	0 (0.00%)	0 (0.00%)	25.60	25.37	1 (1.43%)	1 (1.43%)
Garlic	8.02	8.17	0 (0.00%)	0 (0.00%)	20.13	20.27	0 (0.00%)	0 (0.00%)
Rice	6.22	6.19	0 (0.00%)	0 (0.00%)	5.86	5.86	0 (0.00%)	0 (0.00%)
Red chili	21.96	21.90	3 (4.29%)	4 (5.71%)	35.71	36.48	15 (21.43%)	14 (20.00%)
Bird's-eye chili	27.80	28.24	10 (14.29%)	9 (12.86%)	61.69	61.19	33 (47.14%)	34 (48.57%)
Granulated sugar	3.45	3.40	0 (0.00%)	0 (0.00%)	5.58	5.50	0 (0.00%)	0 (0.00%)
Cooking oil	4.98	4.87	0 (0.00%)	0 (0.00%)	6.26	6.29	0 (0.00%)	0 (0.00%)
Chicken eggs	7.02	7.47	0 (0.00%)	0 (0.00%)	8.60	9.09	0 (0.00%)	0 (0.00%)

Note: K4, G4, K26, and G26 refer to K-means Euclidean at the 4-step horizon, RF-GIC at the 4-step horizon, K-means Euclidean at the 26-step horizon, and RF-GIC at the 26-step horizon, respectively. The "Inaccurate" columns report the number and percentage of regency/city-commodity combinations with MAPE greater than 50% out of the total 70 regencies/cities for each commodity.

Table 7 shows that forecasting accuracy at the regency/city level differed across commodities, and these differences became more pronounced at the 26-step horizon because the range of mean MAPE across commodities and the proportion of inaccurate cases both became larger. At the 4-step horizon, inaccurate cases under K-means Euclidean and RF-GIC occurred only for red chili and bird's-eye chili. At the 26-step horizon, this pattern became more evident: bird's-eye chili had the highest mean MAPE and the largest proportion of inaccurate cases under both K-means Euclidean and RF-GIC, followed by red chili. In contrast, rice, garlic, granulated sugar, cooking oil, and chicken eggs produced no inaccurate cases at the 26-step horizon under either method, whereas shallots showed only one inaccurate case under each method. These findings indicate that the decline in accuracy at longer horizons was primarily associated with chili commodities, particularly bird's-eye chili, which in Table 7 also had the highest mean MAPE.

To provide a more detailed view of the inaccurate cases at the 26-step horizon, Table 8 presents the distribution of these cases by commodity, island distribution, and regencies/cities with an initial missing proportion above 2%.

Table 8 shows that the inaccurate cases at the 26-step horizon were dominated primarily by specific commodities. Under both K-means Euclidean and RF-GIC, most inaccurate cases were associated with bird's-eye chili, followed by red chili, whereas shallots accounted for only a very small proportion. Table 8 also shows that, under both K-means Euclidean and RF-GIC, inaccurate cases were recorded across several islands, with the largest proportions occurring in Java and Sumatra, whereas smaller proportions were observed in Sulawesi, Nusa Tenggara, Kalimantan, and Maluku. In addition, the regencies/cities with an initial missing proportion above 2% that appeared among the inaccurate cases under both K-means Euclidean and RF-GIC were Pematang Siantar City, Cilacap Regency, Ternate City, Banyumas Regency, and Lhokseumawe City. Overall, Table 8 indicates that the inaccurate cases at the 26-step horizon were primarily associated with chili commodities, particularly bird's-eye chili and red chili, and occurred in several specific regencies/cities.

Accordingly, the proposed method appears to be more suitable for relatively stable commodities, whereas highly volatile commodities, particularly bird's-eye chili and red chili, may require additional modeling specifically designed to capture price-volatility dynamics. The cluster-prototype-based approach in this study was used to represent general dynamics at the cluster level, rather than to capture all local specificities of each regency/city. Therefore, the application of cluster-prototype-based results at the regency/city level needs to be complemented by evaluation at the regency/city level, and the development of a hybrid approach with regency/city-level adjustments may be considered in future research.

Table 8. Distribution of inaccurate forecasting cases at the 26-step horizon

Method	Bird's-eye chili	Red chili	Shallots	Island distribution (%)	Regencies/cities with an initial missing proportion > 2% (missing proportion; number of cases)
K-means Euclidean	33 (67.35%)	15 (30.61%)	1 (2.04%)	Java 46.94; Sumatra 38.78; Sulawesi 6.12; Nusa Tenggara 4.08; Kalimantan 2.04; Maluku 2.04	Pematang Siantar City (6.55%; 2 cases); Cilacap Regency (5.69%; 2 cases); Ternate City (5.69%; 1 case); Banyumas Regency (4.96%; 1 case); Lhokseumawe City (2.05%; 2 cases)
RF-GIC	34 (69.39%)	14 (28.57%)	1 (2.04%)	Java 44.90; Sumatra 42.86; Nusa Tenggara 4.08; Sulawesi 4.08; Kalimantan 2.04; Maluku 2.04	Pematang Siantar City (6.55%; 2 cases); Cilacap Regency (5.69%; 2 cases); Ternate City (5.69%; 1 case); Banyumas Regency (4.96%; 2 cases); Lhokseumawe City (2.05%; 2 cases)

Note: The percentages in the Bird's-eye chili, Red chili, and Shallots columns were calculated from the total of 49 inaccurate forecasting cases for each method at the 26-step horizon. The Island distribution (%) column shows the percentage distribution of inaccurate cases by island, which was also calculated from the total of 49 inaccurate cases for each method. Each case represents one regency/city–commodity combination. The last column lists regencies/cities with an initial missing proportion above 2% that appeared among the inaccurate cases; the values in parentheses indicate the initial missing proportion and the number of inaccurate cases for the corresponding regency/city.

3.6. Illustration of model fit and forecasting results

Following [11], who analyzed the same dataset, Pematang Siantar City had the highest proportion of missing values at 6.55%. To illustrate model fit and 26-step-ahead forecasting performance, the bird's-eye chili series is shown for Clusters 1, 2, and 3 obtained from RF-GIC in Figures 4–6, respectively. For Clusters 1 and 2, the cluster prototypes were modeled using VARD(1). The weakest performance is observed in Cluster 3, as shown in Figure 6: bird's-eye chili in Pematang Siantar City, a member of Cluster 3 under RF-GIC, yielded the highest MAPE of 184.91% when modeled using VARD(2).

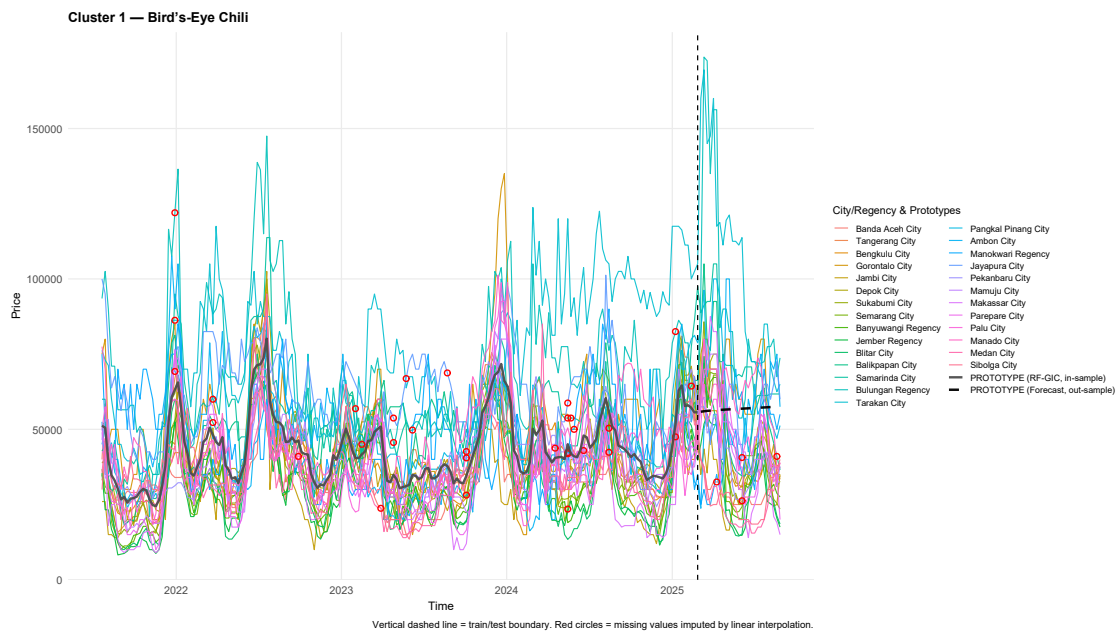


Figure 4. Comparison of actual values, fitted values, 26-step-ahead forecasts, and the cluster prototype for bird's-eye chili in RF-GIC Cluster 1 modeled using VARD(1).

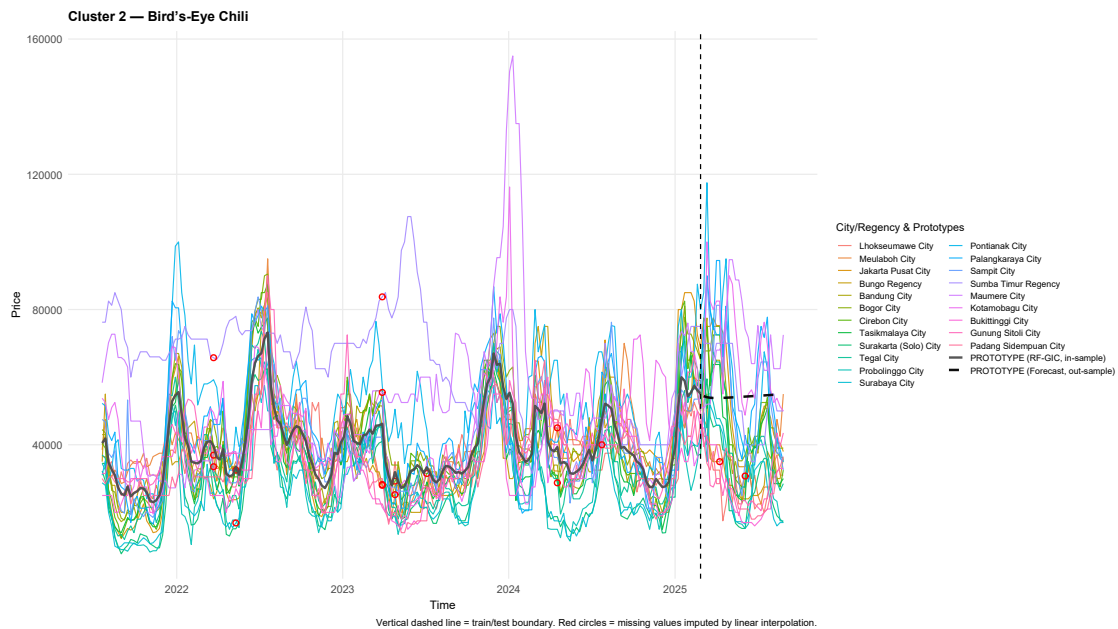


Figure 5. Comparison of actual values, fitted values, 26-step-ahead forecasts, and the cluster prototype for bird's-eye chili in RF-GIC Cluster 2 modeled using VARD(1).

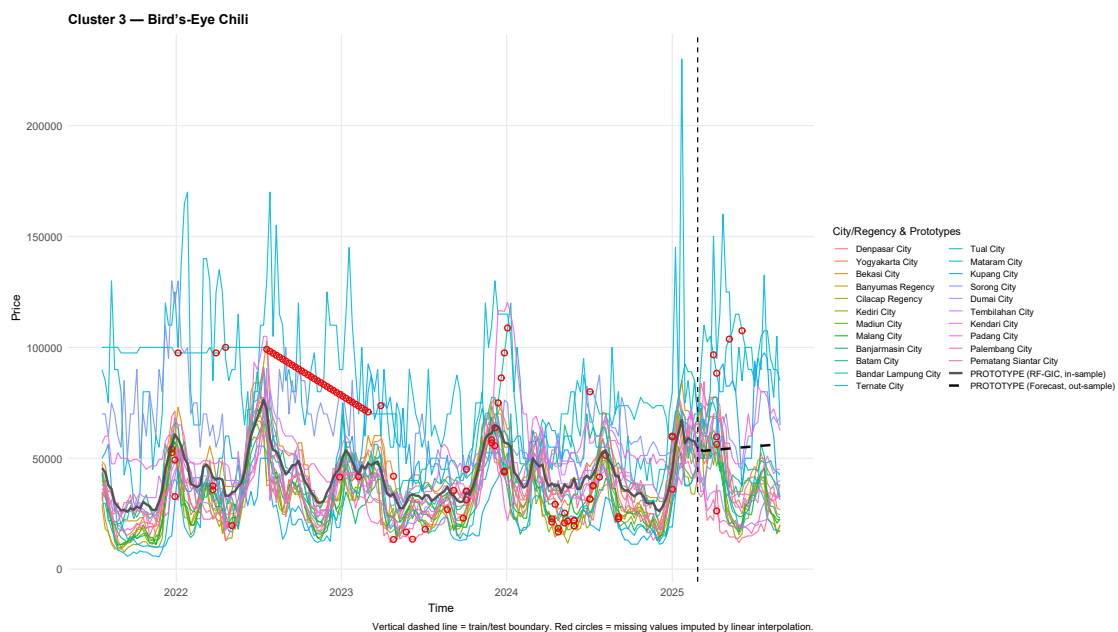


Figure 6. Comparison of actual values, fitted values, 26-step-ahead forecasts, and the cluster prototype for bird's-eye chili in RF-GIC Cluster 3 modeled using VARD(2).

Based on Figures 4–6, Cluster 1 shows wider inter-regional variability, whereas Cluster 2 appears more regular and more consistently follows the cluster prototype. Cluster 3 exhibits the sharpest spikes, so the discrepancies between actual and forecast values become larger more frequently in certain periods. Based on Figure 6, the high

MAPE is primarily attributable to the extreme volatility of bird's-eye chili, which amplifies percentage errors between actual and forecast values in several periods. Consistent with the highest proportion of missing data in Pematang Siantar City [11], this finding provides descriptive evidence suggesting that data incompleteness may be associated with reduced forecasting accuracy.

4. Conclusion

This study developed an integrated procedure for analyzing multivariate time series of strategic food commodity prices across Indonesian regencies/cities, incorporating missing-data imputation, outlier-robust dimensionality reduction, multivariate time series modeling, regional clustering, and cluster-prototype-based forecasting. Forecasting performance was evaluated at two levels, namely the cluster-prototype and regency/city levels, for 4-step and 26-step horizons. At the cluster-prototype level, both K-means Euclidean and RF-GIC achieved very accurate performance at the 4-step horizon and accurate performance at the 26-step horizon, with RF-GIC yielding lower overall MAPE at both horizons. At the regency/city level, both K-means Euclidean and RF-GIC fell into the accurate category at the 4-step horizon and the moderately accurate category at the 26-step horizon, with only a small difference in overall MAPE and slightly lower values for K-means Euclidean. The decrease in accuracy was primarily driven by longer forecasting horizons and commodity-specific volatility, particularly for bird's-eye chili and red chili. Based on the results of this study, RF-GIC is more appropriate when the primary objective is cluster-level representation, qualitative cluster profiling, and cluster-prototype-based forecasting, because RF-GIC produced better cluster quality and lower prototype-level MAPE. By contrast, for forecasting at the regency/city level, K-means Euclidean appears to be sufficient because the difference in MAPE at that level was very small. From a policy perspective, the RF-GIC clustering results support differentiated monitoring and stabilization strategies across clusters. Cluster 2, which was relatively the most stable, may be monitored with lower priority. Cluster 1 requires attention to bird's-eye chili prices. Cluster 3 requires the most intensive attention because Cluster 3 exhibited the highest volatility in red chili and bird's-eye chili as well as the highest mean red chili price. These policy implications are indicative and are based on the clustering results and cluster profiles obtained in this study.

Acknowledgement

The authors gratefully acknowledge the helpful technical discussions and constructive feedback received during the preparation of this manuscript. This research received financial support from the Indonesian Education Scholarship (BPI), the Center for Higher Education Funding and Assessment (PPAPT), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, and the Indonesian Endowment Fund for Education (LPDP), Ministry of Finance of the Republic of Indonesia.

REFERENCES

1. Y. Wang, H. Ding, and H. Li, *Multivariate Time-Series Missing Data Imputation with Convolutional Transformer Model*, Symmetry, vol. 17, no. 5, Art. 686, 2025. <https://doi.org/10.3390/sym17050686>.
2. D. J. Stekhoven, and P. Bühlmann, *MissForest—non-parametric missing value imputation for mixed-type data*, Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012. <https://doi.org/10.1093/bioinformatics/btr597>.
3. R. J. A. Little, and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., John Wiley & Sons, 2002. <https://doi.org/10.1002/9781119013563>.
4. A. Lotfipoor, S. Patidar, and D. P. Jenkins, *Transformer network for data imputation in electricity demand data*, Energy and Buildings, vol. 300, Art. 113675, 2023. <https://doi.org/10.1016/j.enbuild.2023.113675>.
5. S. Moritz, and T. Bartz-Beielstein, *imputeTS: Time Series Missing Value Imputation in R*, The R Journal, vol. 9, no. 1, pp. 207–218, 2017. <https://doi.org/10.32614/RJ-2017-009>.
6. J. Chang, B. Guo, and Q. Yao, *Principal component analysis for second-order stationary vector time series*, The Annals of Statistics, vol. 46, no. 5, pp. 2094–2124, 2018. <https://doi.org/10.1214/17-AOS1613>.
7. B. Zamprogno, V. A. Reisen, P. Bondon, H. H. A. Cotta, and N. C. Reis Jr, *Principal component analysis with autocorrelated data*, Journal of Statistical Computation and Simulation, vol. 90, no. 12, pp. 2117–2135, 2020. <https://doi.org/10.1080/00949655.2020.1764556>.

8. W. W. S. Wei, *Principal Component Analysis of Multivariate Time Series*, in *Multivariate Time Series Analysis and Applications*, 1st ed., pp. 139–161, John Wiley & Sons, 2019. <https://doi.org/10.1002/9781119502951.ch4>.
9. H. H. A. Cotta, *Análise de componentes principais robusta em dados de poluição do ar: aplicação à otimização de uma rede de monitoramento*, Master's thesis, Universidade Federal do Espírito Santo, 2014. <https://dspace5.ufes.br/items/72c34454-d695-48f9-9eec-3297f7c6a0d5>.
10. V. A. Reisen, C. Lévy-Leduc, E. Z. Monte, and P. Bondon, *A dimension reduction factor approach for multivariate time series with long-memory: A robust alternative method*, *Statistical Papers*, vol. 65, no. 5, pp. 2865–2886, 2024. <https://doi.org/10.1007/s00362-023-01504-2>.
11. N. Valentika, I. M. Sumertajaya, A. H. Wigena, and F. M. Afendi, *Dimensionality Reduction Evaluation of Multivariate Time Series of Consumer Price Index in Indonesia*, *JTAM (Jurnal Teori dan Aplikasi Matematika)*, vol. 10, no. 1, pp. 28–44, 2026. <https://journal.ummat.ac.id/index.php/jtam/article/view/34151>.
12. H. H. A. Cotta, *Robust Methods in Multivariate Time Series*, Doctoral thesis, Universidade Federal do Espírito Santo, 2019. https://sappg.ufes.br/tese_drupal/tese_14040_UFESFINALTheseHigor0809%20%282%29.pdf.
13. V. A. Reisen, E. Z. Monte, G. da Conceição Franco, A. M. Sgrancio, F. A. F. Molinares, P. Bondon, F. A. Ziegelmann, and B. Abraham, *Robust estimation of fractional seasonal processes: Modeling and forecasting daily average SO₂ concentrations*, *Mathematics and Computers in Simulation*, vol. 146, pp. 27–43, 2018. <https://doi.org/10.1016/j.matcom.2017.10.004>.
14. H. H. A. Cotta, V. A. Reisen, P. Bondon, and P. R. P. Filho, *Identification of redundant air quality monitoring stations using robust principal component analysis*, *Environmental Modeling and Assessment*, vol. 25, no. 4, pp. 521–530, 2020. <https://doi.org/10.1007/s10666-020-09717-7>.
15. H. H. A. Cotta, V. A. Reisen, P. Bondon, and W. Stummer, *Robust estimation of covariance and correlation functions of a stationary multivariate process*, *International Work-Conference on Time Series*, pp. 47–58, 2017. <https://centralesupelec.hal.science/hal-01578459>.
16. E. Rohaeti, I. M. Sumertajaya, A. H. Wigena, and K. Sadik, *MTSclust with Handling Missing Data Using VAR-Moving Average Imputation*, *Mathematics and Statistics*, vol. 11, no. 2, pp. 229–244, 2023. <https://doi.org/10.13189/ms.2023.110201>.
17. L. Breiman, *Random forests*, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. <https://doi.org/10.1023/A:1010933404324>.
18. Z. Lin, E. Laska, and C. Siegel, *A general iterative clustering algorithm*, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 433–446, 2022. <https://doi.org/10.1002/sam.11573>.
19. P. Mudgal, *Clustering of temporal and visual data: Recent advancements*, *Data*, vol. 11, no. 1, Art. 7, 2026. <https://doi.org/10.3390/data11010007>.
20. R. Pérez-Chacón, G. Asencio-Cortés, A. Troncoso, and F. Martínez-Álvarez, *Pattern sequence-based algorithm for multivariate big data time series forecasting: Application to electricity consumption*, *Future Generation Computer Systems*, vol. 154, pp. 397–412, 2024. <https://doi.org/10.1016/j.future.2023.12.021>.
21. G. Cui, T. Hu, W. Zhang, and H. Bao, *MMTransformer: A multivariate time-series resource forecasting model for multi-component applications*, *Scientific Reports*, vol. 15, no. 1, Art. 29740, 2025. <https://doi.org/10.1038/s41598-025-07162-8>.
22. S. Mamipour, M. Yahoo, and S. Jalalvandi, *An empirical analysis of the relationship between the environment, economy, and society: Results of a PCA-VAR model for Iran*, *Ecological Indicators*, vol. 102, pp. 760–769, 2019. <https://doi.org/10.1016/j.ecolind.2019.03.039>.
23. I. M. Sumertajaya, E. Rohaeti, A. Fitrianto, and W. P. Adi P., *Development of food commodity price forecasting model as an early warning system with a multivariate time series clustering*, *Sains Malaysiana*, vol. 53, no. 11, pp. 3779–3789, 2024. <https://doi.org/10.17576/jsm-2024-5311-20>.
24. Bank Indonesia, *Tabel Harga Pedagang Besar Berdasarkan Daerah*, Pusat Informasi Harga Pangan Strategis (PIHPS) Nasional, 2025. <https://www.bi.go.id/hargapangan/TabelHarga/PedagangBesarDaerah>.
25. M. D. Ariefianto, *Ekonometrika Esensi dan Aplikasi dengan menggunakan EViews*, Erlangga, 2012.
26. S. Wold, K. Esbensen, and P. Geladi, *Principal component analysis*, *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9).
27. R. Bro, and A. K. Smilde, *Principal component analysis*, *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014. <https://doi.org/10.1039/c3ay41907j>.
28. R. Munir, *Metode Numerik*, Penerbit Informatika, 2015.
29. K. Wright, *hopkins: Calculate Hopkins Statistic for Clustering*, R package version 1.1, 2025. <https://cran.r-project.org/web/packages/hopkins/hopkins.pdf>.
30. Z. A. Omar, S. N. Chin, S. R. M. Hashim, and N. Hamzah, *Exploring clusters of rare events using unsupervised random forests*, *Journal of Physics: Conference Series*, vol. 2314, no. 1, Art. 012019, 2022. <https://doi.org/10.1088/1742-6596/2314/1/012019>.
31. S. Shammugam, A. Rathgeber, and T. Schlegl, *Causality between metal prices: Is joint consumption a more important determinant than joint production of main and by-product metals?*, *Resources Policy*, vol. 61, pp. 49–66, 2019. <https://doi.org/10.1016/j.resourpol.2019.01.010>.
32. M. Firdaus, *Aplikasi Ekonometrika dengan E-Views, Stata, dan R*, IPB Press, 2020.
33. R. J. C. Chen, P. Bloomfield, and J. S. Fu, *An evaluation of alternative forecasting methods to recreation visitation*, *Journal of Leisure Research*, vol. 35, no. 4, pp. 441–454, 2003. <https://doi.org/10.1080/00222216.2003.11950005>.
34. D. Ollech, *seastests: Seasonality tests*, R package version 0.15.4, 2025. <https://cran.r-project.org/web/packages/seastests/seastests.pdf>.
35. Badan Informasi Geospasial, *RBI50K_ADMINISTRASI_KABKOTA_20230907_gdb [File Geodatabase]*, 2023. <https://geoservices.big.go.id/portal/home/item.html?id=38f5b7a705b44f959bd25cb99680f78e>.