Sains Malaysiana 53(11)(2024): 3779-3789
http://doi.org/10.17576/jsm-2024-5311-20

Development of Food Commodity Price Forecasting Model as an Early Warning System with a Multivariate Time Series Clustering

(Pembangunan Model Peramalan Harga Komoditi Makanan sebagai Sistem Amaran Awal dengan Pengelompokan Siri Masa Multivariat)

I MADE SUMERTAJAYA[1,*], EMBAY ROHAETI[2], ANWAR FITRIANTO[1] & WINDHIARSO PONCO ADI P[3]

[1]Department of Statistics, Faculty of Mathematics and Science, Bogor Agricultural University, 16680 Bogor, West Java, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Science, Pakuan University, 16129 Bogor, West Java, Indonesia
[3]Badan Pusat Statistik, 10440 Jakarta, West Java, Indonesia

Received: 17 June 2024/Accepted: 30 September 2024

ABSTRACT

Fluctuations in food commodity prices have a significant impact on a country's food security, purchasing power, and economic growth. Therefore, good governance is needed to maintain price stability, one of which is by developing a forecasting model as an early warning system. This study aims to develop a food commodity price forecasting model using Multivariate Time Series Clustering (MTSClust) and Vector Autoregressive Imputation Method with Moving Average (VAR-IMMA) approaches for food commodities in the Indonesian region. The data used in this study consisted of daily prices of 13 commodities from 103 districts/cities in Indonesia. Data analysis was conducted in several stages, namely VAR modeling, K-means Euclidean clustering, profiling, and forecasting. The results show that 103 sample districts/cities across Indonesia can be grouped into four types of regions based on food price movement patterns. There are homogeneous islands such as Maluku where the sample district/city are in the same cluster, but there are also heterogeneous islands such as Kalimantan and Papua with their four clusters. The forecasting evaluation results show good accuracy with Root Mean Square Error (RMSE) scores below IDR 1000.00 in most cases, which is equivalent to Mean Absolute Percentage Error (MAPE) scores below 10%. However, two commodities, namely cayenne pepper and red chili, need more attention due to relatively high RMSE and MAPE scores, although not exceeding 30% MAPE in most cases. These results show that the MTSClust and VAR-IMMA approaches are accurate in forecasting food commodity prices, although further research is needed for the two chili commodities.

Keywords: Early warning system; food security; MTSClust; VAR; VAR-IMMA

ABSTRAK

Turun naik dalam harga komoditi makanan mempunyai kesan yang besar terhadap keselamatan makanan, kuasa beli dan pertumbuhan ekonomi sesebuah negara. Oleh itu, tadbir urus yang baik diperlukan untuk mengekalkan kestabilan harga, salah satunya dengan membangunkan model peramalan sebagai sistem amaran awal. Kajian ini bertujuan untuk membangunkan model peramalan harga komoditi makanan menggunakan pendekatan Siri Masa Multivariat Berkelompok (MTSClust) dan Kaedah Pengimputan Vektor Autoregresif dengan Purata Bergerak (VAR-IMMA) bagi komoditi makanan di wilayah Indonesia. Data yang digunakan dalam kajian ini terdiri daripada harga harian 13 komoditi dari 103 daerah/bandar di Indonesia. Analisis data dijalankan dalam beberapa peringkat iaitu pemodelan VAR, K-means Euclidean berkelompok, pemprofilan dan peramalan. Hasil kajian menunjukkan bahawa 103 sampel daerah/bandar di seluruh Indonesia boleh dikumpulkan kepada empat jenis wilayah berdasarkan corak pergerakan harga makanan. Terdapat pulau homogen seperti Maluku di mana daerah/bandar sampel berada dalam kelompok yang sama, tetapi terdapat juga pulau heterogen seperti Kalimantan dan Papua dengan empat kelompoknya. Keputusan penilaian peramalan menunjukkan ketepatan yang baik dengan skor Punca Min Ralat Kuasa Dua (RMSE) di bawah IDR 1000.00 dalam kebanyakan kes, yang bersamaan dengan skor Min Ralat Peratusan Mutlak (MAPE) di bawah 10%. Walau bagaimanapun, dua komoditi iaitu lada cayenne dan cili merah memerlukan lebih perhatian kerana markah RMSE dan MAPE yang agak tinggi, walaupun tidak melebihi 30% MAPE dalam kebanyakan kes. Keputusan ini menunjukkan bahawa pendekatan MTSClust dan VAR-IMMA adalah tepat dalam meramalkan harga komoditi makanan, walaupun kajian lanjut diperlukan untuk kedua-dua komoditi cili ini.

Kata kunci: Keselamatan makanan; MTSClust; sistem amaran awal; VAR; VAR-IMMA

INTRODUCTION

Food commodity prices reflect various conditions, ranging from the availability of supply, smooth distribution, international trade conditions, successful implementation of government policies, and conditions of people's purchasing power to the welfare of the population (FAO Food Price Index 2014; Kusnadi 2022). Fluctuations in food commodity prices can also affect regional and national food security (Salasa 2021). Based on the Global Food Security Index (GFSI), Indonesia ranks 69th out of 113 countries, or 13th in the Asia Pacific region, with a score of 59.2 (Global Food Security Index 2022; Kementerian Pertanian 2022). Domestically, data from the Badan Pangan Nasional shows that food security in Indonesia is uneven, with some regions categorized as highly vulnerable (Badan Pangan Nasional 2022).

Long-term increases in food commodity prices can lead to inflation, affecting people's purchasing power (Roziah et al. 2023). This phenomenon causes economic growth problems, making it an essential focus of regional (TPID) and national (TPIN) inflation control teams. In Indonesia, TPID is located in every city and is tasked for keeping the inflation rate stable (Nurhasanatun, Bustami & Afrijal 2023). Therefore, it is necessary to manage commodity prices in each region to maintain the stability of food commodity prices.

One key strategy to maintain the stability of food commodity prices is forecasting future prices. The results of such forecasting can serve as a crucial tool in supporting policy and strategy making, particularly in efforts to curb the level of food commodity price increases. The multivariate time series model, a type of forecasting model that is particularly suitable for the case of food commodity prices due to their interdependence, can be instrumental in this regard. The Vector Autoregression (VAR) model, for instance, has been successfully used in several previous studies (Embay et al. 2023, 2022; I Made et al. 2023).

A few studies have examined food commodity price forecasting models in Indonesia. Firmansyah, Maruli and Harahap (2023) compared the VAR and Vector Error Correction Model (VECM) in modeling beef prices in several provinces on the Sumatra, Java, Bali, and Nusa Tenggara Islands. Meanwhile, Effendy et al. (2021) used a univariate Autoregressive Integrated Moving Average (ARIMA) model to predict rice and corn prices in Central Sulawesi. For different agricultural commodities, Kunandar et al. (2023) predicted the price of red chili in the Banyumas Regency using an ARIMA model. Meanwhile, Primageza et al. (2021) predicted rice prices in six provinces on the island of Java using a hybrid Neural Network ARIMA with Explanatory Variable (NN-ARIMAX) and Neural Networks Generalized Space Time ARIMAX (NNs-GSTARIMAX) model, while Christine et al. (2023) used a random forest model to predict the national monthly average rice price.

While previous studies have made valuable contributions to the field, they often focus on a single commodity, whereas food security depends on a variety of food commodities. Moreover, the challenges of data collection in vast areas like Indonesia, with its 514 districts/cities, have limited the scope of previous studies to a few areas. These two aspects, which are often overlooked, are the primary gaps that this study aims to address, thereby making a unique contribution to the existing literature.

This research aims to develop a food commodity price forecasting model using multivariate time series clustering (MTSClust) and VAR-IMMA methods to support Indonesia's price stability and food security. The development of this model is expected to provide an overview of the future price conditions of 13 food commodities and support policy-making and strategies to reduce price increases of food commodities in all regions of Indonesia.

MATERIALS AND METHODS

DATA COLLECTION

The data used in this study are food commodity price data from 1 September 2022 to 1 February 2024 at the district/city level in Indonesia. The data is in the form of daily data from 20 food commodities, namely rice, shallots, garlic, red chilies, cayenne pepper, purebred chicken meat, beef, sugar, cooking oil, eggs, raw tofu, tempeh, wheat flour, milk powder, toddler milk, bananas, fish, oranges, instant dry noodles, and wet shrimp. The data source was obtained from the Indonesian Statistics Bureau.

DATA PREPROCESSING

Data preprocessing is the first stage of analysis carried out in this study. The aim is to ensure that the data processed in the next stage is clean and of good quality so that the forecasting model obtained can also be accurate. One of the data quality problems that often arise is the problem of missing data. This problem is critical in time series forecasting because each data point is sequential and related to others. In this study, missing data is handled using a previous research method, the VAR-IMMA imputation method (I Made et al. 2023).

VAR-IMMA is a multivariate time series data imputation method based on the VAR model. However, before using the VAR model, initial imputation is performed using an Exponential Moving Average, which is mathematically defined as:

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

where α is the weight; and k is half the MA window size (Moritz & Bartz-Beielstein 2017).

The initial imputation results are then used for the VAR model-based imputation iteration, which is mathematically defined as:

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

where y_t, y_{t-i} are $n \times 1$ vectors; n is the number of variables; t and $t-i$ are time indexes where $i = 1,2,\dots,p$. p is the order of the VAR model. A_0 is an intercept vector and A_i is a coefficient matrix sized $n \times n$. u_t is a white noise vector at time t (Akkaya 2021).

In addition to the missing data problem, there are also data quality issues related to the assumptions of the VAR model, especially the stationarity assumption—techniques such as differencing need to be used. First-order differencing is defined as the difference between the observed value at time t and the observed value at the previous time:

$$y'_t = y_t - y_{t-1}$$

The second-order differencing is the difference of the first-order differencing:

$$y''_t = y'_t - y'_{t-1}$$

The differencing order can be increased with the same logic until stationary data is obtained (Kamalov 2021).

DATA ANALYSIS

The data analysis stage is divided into the following stages: *Data Preprocessing and Exploration* Data exploration aims to find patterns in the data and serves as a second filter if data problems are missed during preprocessing. Data exploration also aims to observe data characteristics, especially characteristics in time series data, such as indications of seasonality and trends that can affect stationarity and modeling.

Stationarity Test Stationarity is one of the assumptions of VAR modeling, so a stationarity test is mandatory before performing VAR modeling. The stationarity test uses the Augmented Dickey-Fuller (ADF) test (Batarseh 2021). ADF test is a test in a category of tests called 'unit root test'. Unit root is a characteristic of a time series indicating that the time series is non-stationary. Assume a simple Autoregressive (AR) model

$$y_t = \rho y_{t-1} + u_t$$

where y_t is the target variable; t is a time index; ρ is a coefficient; and u_t is an error term. The regression model can be written as

$$\Delta y_t = (\rho - 1) y_{t-1} + u_t = \delta y_{t-1} + u_t$$

where Δ is the first difference operator. The unit root is present if $\rho = 1$, since $\delta = \rho - 1 = 0$, implying the model to be non-stationary. This is called Dickey-Fuller Test. Augmented Dickey-Fuller Test has similar procedure, but applied to this model

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

where α is a constant; β, γ, and δ are coefficients; while p is the lag order of the autoregressive process. The unit root test is tested under the null hypothesis of $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. The value of test statistic

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

is compared to the critical value for Dickey-Fuller test. If the test statistic is less than the critical value, then the null hypothesis is rejected and no unit root is present, implying the model is stationary (Ajewole, Adejuwon & Jemilohun VG 2020; Chang & Park 2002; Dickey & Fuller 1979).

VAR Modeling Data availability is one of the issues that often arise in some regions in Indonesia. Several districts/ cities considered representative of each province were selected as survey data to address this. Modeling using the VAR model was conducted in each of these survey areas. The result of VAR modeling is a model with estimated coefficients. The estimated coefficients are in the form of a matrix whose rows represent regions and whose columns represent coefficients.

Euclidean K-means Clustering The coefficient matrix from the previous stage is clustered using the K-means Euclidean method. The choice of method is based on the results of previous research comparing six clustering methods with similar cases (Ikotun et al. 2023; Embay et al. 2023). The clustering results were then evaluated with two criteria: the

silhouette score and the presence or absence of singleton clusters. Silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to +1, where a high value indicates that the object is more similar to its own cluster than other clusters (Januzaj, Beqiri & Luma 2023). Meanwhile, a singleton cluster is defined as a cluster formed with only one object, often because the object is too different than the other objects.

Profiling Profiling is the process of interpreting the price movement patterns of all regions. The profile of each region is generalized by observing the general pattern of each cluster formed in the previous stage.

Forecasting The profiling results from the previous stage can be used to forecast the future movement of food commodity prices. Before forecasting, the data is divided into training and test data to evaluate the accuracy of the forecasting model. The test data used is the last month of profile data. If the forecasting model has a small RMSE and MAPE value, then the model is considered feasible and can be used to predict commodity prices for the next month. RMSE and MAPE are calculated using these formulas (Hodson 2022; Vivas, Allende-Cid & Salas 2020).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAPE = 100\left(\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}\right)$$

RESULTS AND DISCUSSION

DATA PREPROCESSING AND DATA EXPLORATION

This research successfully collected daily price data for 20 food commodities from 514 districts/cities in Indonesia from September 1, 2022, to February 1, 2024. However, some data quality issues were identified during data preprocessing and exploration. Below are the data quality issues encountered and how they were addressed: 1). The data collected is generally only available on weekdays. To supplement data availability, prices on weekends and public holidays were estimated using the VAR-IMMA method, 2). Data noise exists for all commodities where the commodity price is below IDR 1,000.00 or even IDR 0.00. These observations are considered missing data and were then imputed, 3). Fourteen districts/cities only had data for 2022, 2023, or 2024. Since there were too many missing data points in these 14 districts/cities, the following analysis was conducted on the remaining 500 districts/cities, 4). The percentage of missing data in the remaining 500 districts/cities varied from 35% to 50%. To address this, sampling was conducted using two criteria. First, sampling was done by selecting three district/city with the smallest percentage of missing data in each province, filtering 114 districts/cities (3 districts/cities × 38 provinces). Second, the filtered district/city were filtered again so that none had a more significant missing data percentage than 51%. These two sampling criteria aimed to equalize the number of district/city in each province, considering that provinces in the eastern part of Indonesia tend to have a more significant percentage of missing data. This sampling process selected 103 districts/cities for the following analysis, and 5). Of the 103 sample districts/cities, seven commodities had a percentage of missing data of more than 40%. These commodities are milk powder, toddler milk, bananas, fish, oranges, instant dry noodles, and wet shrimp. Therefore, these seven commodities were excluded from the study, and further analysis used only the remaining 13 commodities.

DATA STATIONARITY

The preprocessed data consists of price data for 13 food commodities from 103 districts/cities spread across all provinces in Indonesia. Data from each commodity in each district/city was tested for stationarity before being modeled using a VAR model. The stationarity test used the Augmented Dickey-Fuller (ADF) test with $\alpha = 5\%$. The ADF test results showed that only some commodities in some districts/cities were stationary in the original data ($d = 0$). In the original data ($d = 0$), only Kota Jayapura and Kabupaten Polewali Mandar were stationary for a maximum of 7 commodities, while other districts/cities were stationary for only 2 or 3 commodities. This non-stationarity is common in price data, especially since the data used is daily, but it can cause inaccuracies in the model and produce inaccurate predictions. Therefore, it is necessary to handle this so that the original data becomes stationary before modeling.

One approach to overcoming non-stationarity in time series data is by differencing. This study performed differencing from the first order ($d = 1$) to the third order ($d = 3$). The ADF test results for the first-order differencing ($d = 1$) showed a more significant stationary series than in the original data ($d = 0$), where all commodities were stationary in most districts/cities. There were only a few districts/cities where one or two commodities were not stationary. Specifically, in the first-order differencing data ($d = 1$), 23 districts/cities were non-stationary for one commodity, four districts/cities were non-stationary for two commodities, and one city was non-stationary for three commodities. Considering the sample has 13 commodities and 103 districts/cities, this non-stationarity number is relatively tiny and deemed acceptable

in VAR modeling. Moreover, the ADF test results at higher orders, i.e., $d = 2$ and $d = 3$, did not improve stationarity conditions significantly and even tended to be worse than at order $d = 1$. Therefore, the data used in VAR modeling was differenced at order $d = 1$.

VAR MODELING

VAR models can be modeled with various lags p, so the optimal value of p must be determined before modeling the entire data set. The optimal lag p can be determined based on domain knowledge. Intuitively, one might tend to use $p = 365$ (i.e., 365 days) to accommodate the 'seasonal' prices of some commodities, such as meat and chili. However, the data exploration results did not show any seasonal pattern in the data used. Hence, using lag $p = 365$ is inappropriate. Alternatively, the optimal lag can be determined by comparing the information criterion of various values of p. In this study, the optimal lag was determined by comparing the information criterion of VAR models with various lag p values, ranging from $p = 1$ to $p = 7$. Each district/city was modeled based on their optimal lags in the range of $p = 1$ to $p = 7$.

The output of the VAR modeling is the coefficient matrix of all districts/cities. To facilitate clustering, the coefficient matrices of all districts/cities were combined into one giant matrix of size 103 x 104, where 103 rows represent each district/city, and 104 columns represent the coefficients. The 104 columns were obtained from 104 = 8 x 13, where 8 is the maximum number of lags ($p = 7$) plus one constant value, and 13 is the number of commodities.

However, the optimal lag for each district is different, so districts with optimal lag $p < 7$ will have fewer coefficients and, hence, fewer columns than districts with optimal lag $p = p_{max} = 7$. For example, districts with optimal lag $p = 4$ will have coefficients of 0 in columns corresponding to lags 5,6 and 7. This approach is in line with the interpretation of the VAR model; in the case of optimal lag $p = 4$, the model coefficients for lags 5, 7 and 7 are irrelevant and do not affect the model.

CLUSTERING (ENTIRE SAMPLE AREA)

Clustering begins with determining the optimal number of clusters. Based on the silhouette score, the optimal number of clusters is $k = 4$. The four clusters formed using the Euclidean K-means method do not produce singleton clusters, meaning no district/city are 'outliers' forming their own clusters, as the cluster with the fewest members has six districts/cities in it. This is good, considering the clusters formed group district/city based on price movement patterns over 18 months. If there are singleton clusters, it could indicate one of two cases: either the Euclidean K-means model is unsuitable for the case being analyzed, or the 'outlier' district/city have very different price movement patterns from other district/city in Indonesia. Of these possibilities, a mismatch between the model and the data is more likely. Therefore, the absence of singleton clusters in the four clusters formed indicates that the K-means Euclidean clusters are suitable for clustering the 103 districts/cities sampled.

In addition to the absence of singleton clusters, a measure of the goodness of a clustering result can also be seen from the silhouette score. The average silhouette score for the formed clusters is around 0.18, indicating that the district/city have been grouped into clusters with similar characteristics. Moreover, when the silhouette scores between the selected clusters and neighboring clusters are compared, positive medians are obtained in all cases except in comparing Cluster 2 and neighboring Cluster 4, where the median value is close to 0. This indicates that some districts/cities in Cluster 2 are in the gray area and could fit into Cluster 4.

Figure 1 shows the map of clusters formed across the survey area.

The clusters formed show some unique patterns when viewed by island. The Maluku Islands tend to have the same pattern of food price movements in all sample districts/cities, as all (six) sample districts/cities in the Maluku Islands are in the same cluster, namely cluster 2. The island of Java is slightly more varied, with district/city belonging to clusters 3 or 4, while districts/cities in Sulawesi, Bali, and Nusa Tenggara belong to clusters 2, 3, or 4. Districts/cities on the island of Sumatra mainly belong to cluster 3, although some districts/cities belong to clusters 1 or 4. The districts/cities on the islands of Kalimantan and Papua have at least one district in each cluster, indicating different food price patterns despite being on the same islands.

PROFILING (ENTIRE SAMPLE AREA)

Cluster profiling aims to obtain the typical characteristics of food commodity prices from the clusters formed. Profiling is done by calculating the average of each food commodity at each time t. The profiling results of the four clusters formed are presented in Figure 2.

In general, Figure 2 shows a consistent pattern in forming the four clusters, where clusters 1 and 2 have similar average prices, as do clusters 3 and 4. However, there are significant differences in beef prices, where Cluster 1 is similar to Cluster 3, while Cluster 2 is similar to Cluster 4. Some commodities show clear differences between the clusters, particularly the price of tempeh, which varies significantly across clusters. Cluster 1 has expensive tempeh, tofu, and beef commodities but relatively cheap broiler chicken prices. In contrast, cluster 2 has the highest broiler chicken prices but the lowest beef prices. Clusters 3 and 4 have similar price patterns but differ in beef and broiler meat prices; Cluster 3 has higher beef prices, but Cluster 4 has higher broiler meat prices.

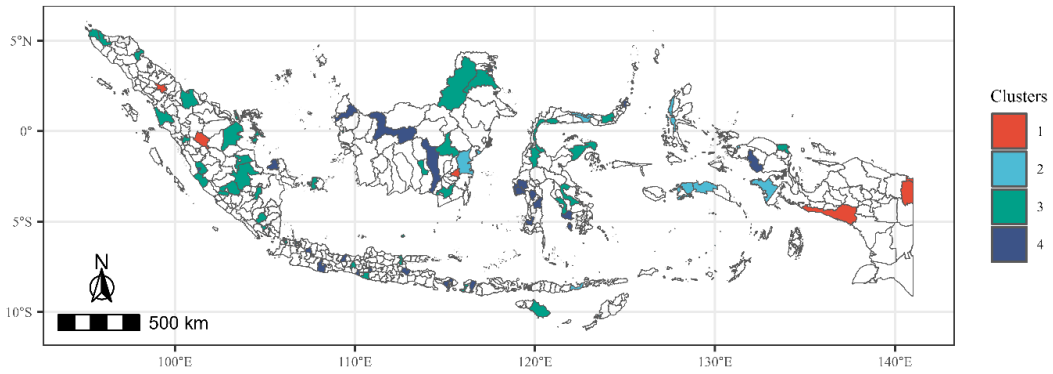

FIGURE 1. Cluster map of the entire survey area

FIGURE 2. Profiling (region-wide) by average food commodity prices

FORECASTING (ENTIRE SAMPLE AREA)

The profiles of the four clusters can be used for forecasting. However, before it is declared feasible as a forecasting model, it is necessary to evaluate it. The evaluation uses holdout validation where the last 32 days (January 1, 2024, to February 1, 2024) are used as test data while the previous data is used as training data.

Table 1 shows the results of a relatively accurate forecasting evaluation with a percentage difference below 1% in each cluster. Some commodities, such as red chili and cayenne pepper, have large RMSE values, indicating that the forecasting results and actual values differ in the range of thousands to tens of thousands. This rather significant difference is less than 1% of the actual value based on the MAPE value. This relatively more significant difference, but with a small percentage difference, is due to the price of red chili and cayenne pepper being more expensive than other commodities. However, there are also expensive commodities such as beef, whose accuracy is much better, with the highest difference around IDR 2,000.00.

Table 1 shows relatively accurate forecasting results, where the RMSE values of most commodities are below IDR 1,000.00. This means that the forecasting errors for these commodities are no more than IDR 1,000.00 from the actual prices. In addition, when evaluated based on the MAPE value, the errors obtained are generally at most 10%, indicating that the forecasting error is not more than 10% of the actual price. However, two commodities that need attention are cayenne pepper and red chili, which have relatively large errors compared to other commodities. However, the MAPE value is still below 30% in most clusters.

Based on Table 1, the forecasting model is relatively accurate and can be used to forecast commodity prices for the next 30 days. Figure 3 shows the forecasting results of each cluster for each food commodity for the next 30 days. Prices are expected to remain relatively stable or even decrease, except in certain cases, such as the price of cayenne pepper in cluster 3, purebred chicken meat in clusters 1, 3, and 4, red chili pepper in clusters 3 and 4, red onion in cluster 3, and tempeh in cluster 3. Overall, the prices in cluster 3 should be monitored closely.

JAVA ISLAND

One of the important results to be studied further is the modeling results on the island of Java. Java is the most populous island in Indonesia, with a population of more than 157 million people, which is equivalent to 56% of Indonesia's population by the end of 2023 (Fadhlurrahman 2024). Such a dense population makes the issue of food price volatility an important issue that must be closely monitored. In addition, district/city in Java also have relatively more complete data availability than other district/city. Complete data availability allows for more accurate modeling that can produce interesting results for further study.

The previous clustering of all survey areas shows that Java is divided into two clusters, as shown in Figure 1. Not many unique patterns are seen in the formed clusters, as districts/cities in Java are clustered with districts/cities in the rest of Indonesia from different islands. However, if the districts/cities in Java are clustered with the districts/cities in Java only, some unique patterns can be seen, as shown in Figure 4.

Based on Figure 4, districts/cities in Java are still optimally divided into two clusters. Some districts/cities previously shown in Figure 1, were in different clusters, but after being clustered specifically on the island of Java alone, they became one cluster. One of the patterns formed is that district/city in Banten, DKI Jakarta, and East Java are in one cluster, namely cluster 2. Meanwhile, six out of nine districts/cities in West Java, Central Java, and DI Yogyakarta are in cluster 1, the same cluster as the city of Yogyakarta, known as one of the cities with the lowest cost of living in Indonesia. Meanwhile, three districts/cities in these three provinces, namely Kulon Progo District (DI Yogyakarta), Cirebon City (West Java), and Wonosobo District (Central Java), are classified in cluster 2. This indicates that the pattern of food price movements in districts/cities in Banten, DKI Jakarta, and East Java are relatively similar. In contrast, there are still differences in food price movement patterns between districts/cities in West Java, Central Java, and DI Yogyakarta. Furthermore, both clusters are characterized by the prices of five commodities: Shallots, garlic, beef, cooking oil, and tempeh. The profile of the two Java Island clusters can be seen in Figure 5.

Figure 5 shows the average food price profile of the two Java Island clusters. It can be seen in Figure 5 that Cluster 2 has higher average prices than Cluster 1, especially for commodities such as shallots, garlic, beef, cooking oil, and tempeh.

TABLE 1. Evaluation of forecasting results on the region-wide profile

Komoditas	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Cayenne Pepper	27.093,80	0,29	40.870,88	0,63	18.521,67	0,31	26.154,15	0,48
Cooking Oil	466,25	0,01	475,33	0,02	609,16	0,03	257,30	0,01
Cow Meat	2.143,57	0,01	1.140,80	0,01	1.037,43	0,01	863,15	0,00
Flour	628,60	0,03	121,59	0,01	137,55	0,01	386,04	0,03
Granulated Sugar	381,90	0,01	529,42	0,03	210,10	0,01	688,11	0,04
Purebred Chicken Egg	1.030,70	0,03	688,71	0,02	894,16	0,03	1.985,68	0,07
Purebred Chicken Meat	1.448,01	0,04	493,15	0,01	1.184,07	0,03	2.033,19	0,05
Raw Tofu	411,96	0,02	264,78	0,01	74,30	0,01	140,84	0,01
Red Chilli Pepper	14.541,66	0,23	13.254,70	0,25	7.003,48	0,13	4.685,74	0,07
Red Onion	3.994,85	0,07	3.880,59	0,08	6.131,09	0,14	4.166,40	0,09
Rice	478,49	0,01	460,10	0,03	417,84	0,03	544,49	0,03
Tempe	769,65	0,02	646,92	0,03	528,12	0,02	162,68	0,01
White Onion	1.724,98	0,03	2.019,98	0,04	2.101,52	0,04	2.779,60	0,06

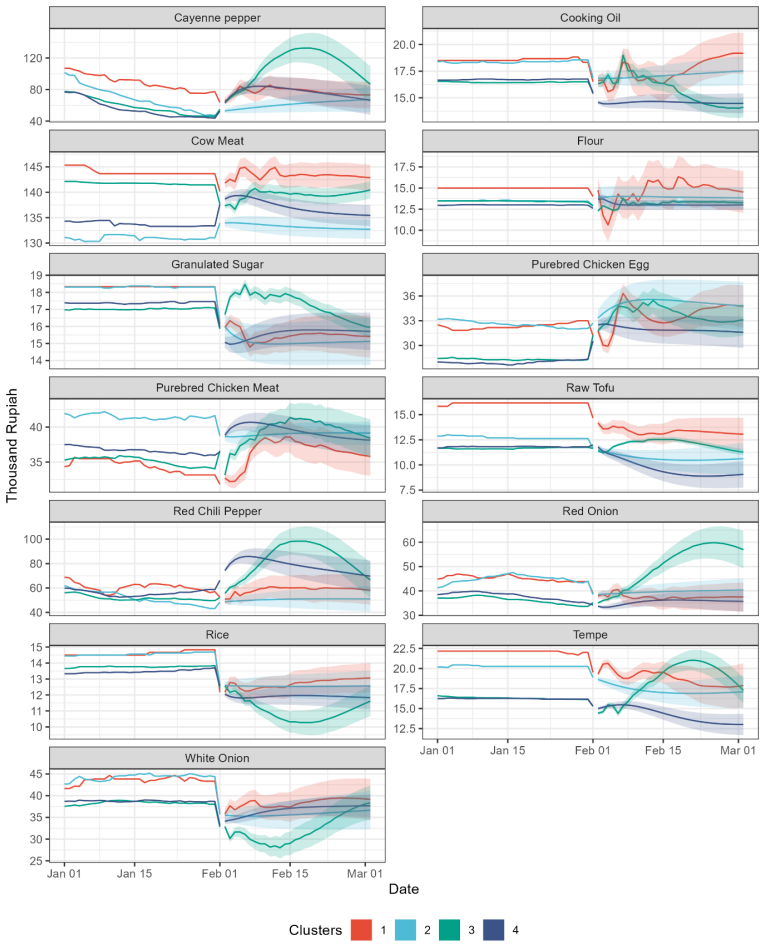

FIGURE 3. Forecasting based on average food commodity prices

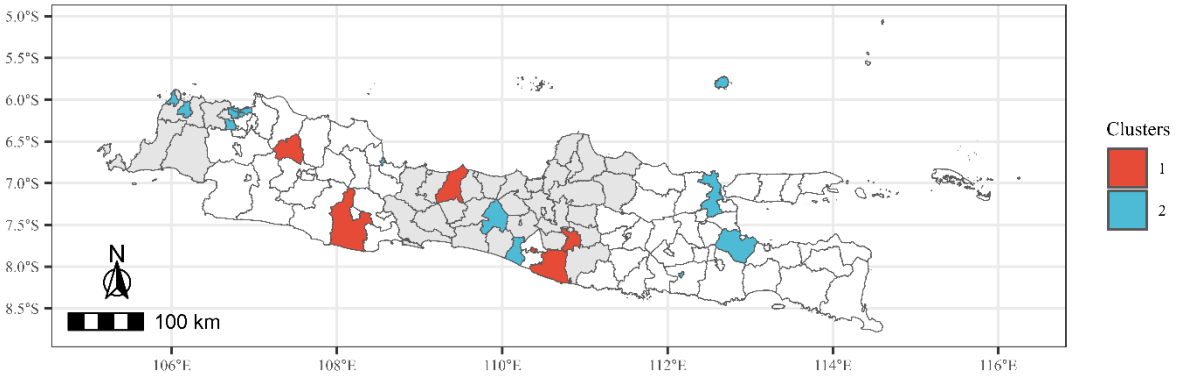

FIGURE 4. Special cluster map of Java Island

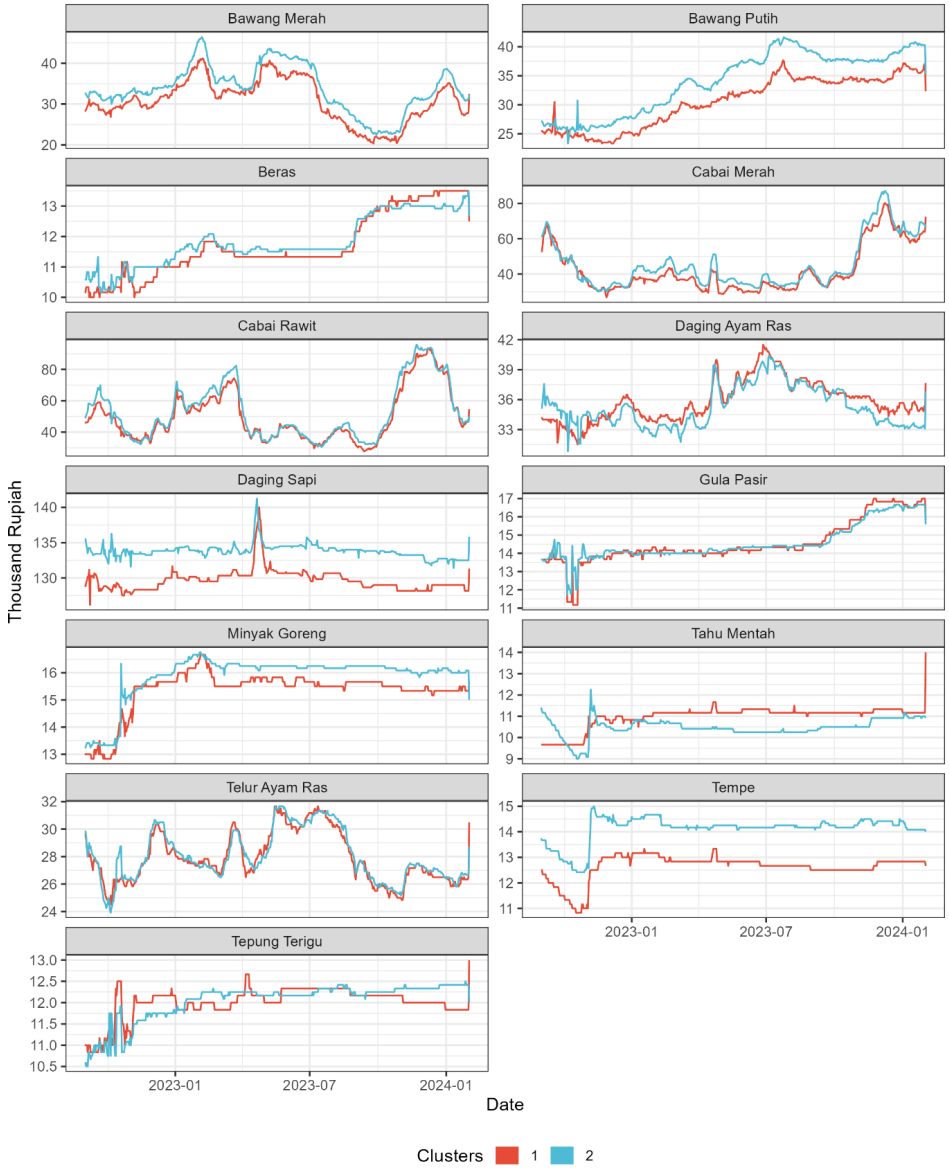

FIGURE 5. Profiling (Java Island) based on average food commodity prices

CONCLUSION

This study successfully developed a food commodity price forecasting model using the MTSClust and VAR-IMMA approaches. The results show that, nationally, Indonesia can be divided into four types (clusters) of regions based on the pattern of food commodity price movements. The Maluku region tends to be homogeneous, with all sampled districts/cities exhibiting relatively similar food price movement patterns. Other islands are more heterogeneous; for example, the sample districts/cities in Kalimantan are divided into four clusters. Further analysis shows that Java, the most populous island in Indonesia, is divided into two large clusters, where districts/cities at the western end (Banten and DKI Jakarta) and the eastern end (East Java) are in the same cluster. Regarding accuracy, the forecasting model produces good results, with RMSE values of less than IDR 1,000.00 for most food commodities. This RMSE value is equivalent to less than 10%, meaning that the error of the forecasting results is at most 10% of the actual price. However, some commodities, such as cayenne pepper and red chili, still show relatively large errors with MAPE scores of around 30%. Therefore, future research could focus on improving the forecasting accuracy for these two commodities.

REFERENCES

Ajewole, K.P., Adejuwon, S.O. & Jemilohun, V.G. 2020. Test for stationarity on inflation rates in Nigeria using augmented Dickey Fuller test and Phillips-Persons test. *IOSR Journal of Mathematics* 16(3): 11-14.

Akkaya, M. 2021. Vector autoregressive model and analysis. *In Handbook of Research on Emerging Theories, Models, and Applications of Financial Econometrics*, edited by Adigüzel Mercangöz, B. Springer, Cham. https://doi.org/10.1007/978-3-030-54108-8_8

Badan Pangan Nasional. 2022. 'Indeks Ketahanan Pangan'. https://badanpangan.go.id/storage/app/media/2023/Buku%20Digital/Buku%20Indeks%20Ketahanan%20Pangan%202022%20Signed.pdf

Batarseh, A. 2021. The nature of the relationship between the money supply and inflation in the Jordanian Economy (1980-2019). *Banks and Bank Systems* 16(2): 38-46. https://doi.org/10.21511/bbs.16(2).2021.04

Chang, Y. & Park, J.Y. 2002. On the asymptotics of ADF tests for unit roots. *Econometric Reviews* 21(4): 431-447. https://doi.org/10.1081/ETC-120015385

Dewi, C., Prasatya, G.S.K., Christanto, H.J., Widiarto, S.O.B., & Dai, G. 2023. Modified random forest regression model for predicting wholesale rice prices. *Journal of Theoretical and Applied Information Technology* 101(23): 7749-7759.

Dickey, D.A. & Fuller, W.A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74(366): 427-431. https://doi.org/10.1080/01621459.1979.10482531

Effendy, Evansyah D., Antara, M., Noli, K., & Pratama, F.M. 2021. Forecasting model of production and price of grains commodity in Central Sulawesi. *Journal of Theoretical and Applied Information Technology* 99(14): 3555-3563. https://doi.org/10.46300/9103.2021.9.8

Fadhlurrahman, I. 2024. "Jumlah Penduduk Di 38 Provinsi Indonesia Desember 2023." Databoks. February 15, 2024. https://databoks.katadata.co.id/datapublish/2024/02/15/jumlah-penduduk-di-38-provinsi indonesia-desember-2023.

FAO Food Price Index. 2014. *Africa Research Bulletin: Economic, Financial and Technical Series* 51(9): 20574A. https://doi.org/10.1111/j.1467-6346.2014.06047.x.

Firmansyah, Maruli, P. & Harahap, A. 2023. Analysis of beef market integration between consumer and producer regions in Indonesia. *Open Agriculture* 8(1): 20220221. https://doi.org/10.1515/opag-2022-0221

Global Food Security Index. 2022. *The Economist Intelligence Unit*.

Hodson, T.O. 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development* 15(14): 5481-5487. https://doi.org/10.5194/gmd-15-5481-2022

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. & Heming, J. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* 622: 178-210. https://doi.org/10.1016/j.ins.2022.11.139

Januzaj, Y., Beqiri, E. & Luma, A. 2023. Determining the optimal number of clusters using silhouette score as a data mining technique. *International Journal of Online and Biomedical Engineering* 19(4): 174-182. https://doi.org/10.3991/ijoe.v19i04.37059

Kamalov, F. 2021. A note on time series differencing. *Gulf Journal of Mathematics* 10(2): 50-56. https://doi.org/10.56947/gjom.v10i2.609

Kementerian Pertanian. 2022. Statistik Ketahanan Pangan 2022. https://satudata.pertanian.go.id/assets/docs/publikasi/Statistik_Ketahanan_Pangan_2022.pdf

Kusnadi, N.A. 2024. Pengaruh fluktuasi harga komoditas pangan terhadap inflasi di Provinsi Jawa Timur. *Thesis. FEB Universitas Brawijaya* 6 (2).

Moritz, S. & Bartz-Beielstein, T. 2017. ImputeTS: Time series missing value imputation in R. *R Journal* 9(1): 207-218. https://doi.org/10.32614/rj-2017-009

Nurhasanatun, Usman, B. & Afrijal. 2023. Analisis peran tim pengendalian inflasi Daerah Kota Banda Aceh dalam pengendalian inflasi. *Jurnal Ilmiah Mahasiswa FISIP USK* 8(2). www.jim.unsyiah.ac.id/Fisip

Prasetyo, K., Putri, D.D., Wijayanti, I.K.E. & Zulkifli, L. 2023. Forecasting of red chilli prices in Banyumas Regency: The ARIMA approach. *E3S Web of Conferences* 444: 02017. https://doi.org/10.1051/e3sconf/202344402017

Primageza, H., Vinarti, R.A., Tyasnurita, R., Riksakomara, E. & Muklason, A. 2021. Comparison of NNs-ARIMAX and NNs-GSTARIMAX on rice price forecasting in Indonesia. *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Depok, Indonesia. pp. 1-8. https://doi.org/10.1109/ICACSIS53237.2021.9631332

Rohaeti, E., Sumertajaya, I.M., Wigena, A.H. & Sadik, K. 2023. MTSClust with handling missing data using VAR-Moving average imputation. *Mathematics and Statistics* 11(2): 229-244. https://doi.org/10.13189/ms.2023.110201

Rohaeti, E., Sumertajaya, I.M., Wigena, A.H. & Sadik, K. 2022. The prominence of vector autoregressive model in multivariate time series forecasting models with stationary problems. *BAREKENG: Jurnal Ilmu Matematika dan Terapan* 16(4): 1313-1324. https://doi.org/10.30598/barekengvol16iss4pp1313-1324

Roziah, T.R., Septiani, R., Amapoli, E.V. & Muhammad, R. 2023. Inflasi di Indonesia: Perkembangan dan pengendaliannya. *Jurnal Ilmiah Multidisiplin* 1(10): 9-18.

Salasa, A.R. 2021. Paradigma dan dimensi strategi ketahanan pangan Indonesia. *Jejaring Administrasi Publik* 13(1): 35-48. https://doi.org/10.20473/jap.v13i1.29357

Sumertajaya, I.M., Rohaeti, E., Wigena, A.H. & Sadik, K. 2023. Vector autoregressive-moving average imputation algorithm for handling missing data in multivariate time series. *IAENG International Journal of Computer Science* 50(2): IJCS_50_2_42.

Vivas, E., Allende-Cid, H. & Salas, R. 2020. A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score. *Entropy* https://doi.org/10.3390/e22121412

*Corresponding author; email: imsjaya@apps.ipb.ac.id