International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Time-Series Data Modeling for Inflation Forecasting based on Subcategories of Commodity using TSClust Approach as Pre-processing

Budi Utami[a*], Hari Wijayanto[b], I Made Sumertajaya[c]

[a,b,c]*Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agriculture University,Bogor, Indonesia*

[a]*Email: budiutami@bps.go.id*

[b]*Email: hari_ipb@yahoo.com*

[c]*Email: imsjaya@gmail.com*

Abstract

High and unstable inflation will lead to a decline in the quality of life of the people and the slowing down of the economy of a region. The efforts to control inflation are continuously carried out by both central and local governments. Therefore, a precise and efficient forecasting method is needed to forecast inflation according subcategories of commodity, considering the various movement patterns of each subcategory. This research developed a model to predict inflation according to subcategories of commodity. To determine the pattern and model of each subcategory of commodity, a time-series analysis approach was used to obtain a good and efficient forecasting model in terms of time, effort and cost. The cluster-level model established were ARIMA, ARIMAX, VAR and VARX models. The research shows that the best clustering 35 inflation rate based on the subcategories of commodity were obtained by using Piccolo measure of dissimilarity, resulted in 4 clusters. The best forecasting model varies for each subcategory of commodity, but ARIMAX model shows the smallest RMSE value compared to the other three models.

Keywords: arimax; inflation; tsclust; varx.

---

* Corresponding author.

1. Introduction

Stable inflation is a prerequisite for sustainable economic growth that ultimately benefits the welfare of society. The importance of inflation control is based on the consideration that high and unstable inflation has a negative impact on the socio-economic conditions of the community [1]. Understanding the importance of inflation role, in order to achieve a low and stable inflation, Government and Bank Indonesia have established Inflation Task Force (**Tim Pemantauan dan Pengendalian Inflasi/**TPI) at the national level since 2005. To strengthen the coordination, the team continued to set up at the local level called Regional Inflation Task Force (**Tim Pemantauan dan Pengendalian Inflasi** Daerah/TPID) in 2008. The number of RITFs is currently 486 RITFs at both the provincial and municipal / regency levels [2]. Special Capital Region of Jakarta as the capital city is the barometer for the Indonesian economy. Jakarta's economy contributes the most to Indonesian economy. Meanwhile, the inflation rate in Jakarta fluctuates every year with the highest inflation weight compared to other cities in Indonesia, which is recorded at 20.15 percent [3]. The measures of Jakarta inflation weight will surely greatly affect national inflation. The pattern of inflation movements in Jakarta varies for each subcategory of commodity. In Jakarta, its characteristics are much influenced by shock factors, such as production disturbances, distribution disturbances, and fuel price changes (BBM) and others. To establish an appropriate inflation control policy, the Jakarta RITF requires an accurate forecasting model in accordance with the characteristics of each subcategory of commodity. Some models of inflation forecasting that has been developed by using time series analysis techniques, among others ARIMA, ARIMAX-NN, VAR, *Threshold Vector Error Correction* (TVCEM), multi input intervention model, ARCH/GARCH and others. The aim of the study was to provide a forecasting model to 35 inflation according to subcategories of commodity. To improve the efficiency of time, effort and cost in determining the appropriate model, researchers made use the *Time Series Clustering* (TSClust) technique in *preprocessing*. TSClust clustered 35 subcategories of commodity to k-cluster. Each cluster was identified using ARIMA model which provide the movement patterns. Next, VAR modeling was applied between established clusters, and to reduce errors, researcher included exogenous variables into the model (ARIMAX and VARX). Exogenous variables included in the model were rupiah exchange rate, BI rate, and dummy variables, which are assumed to have an effect on inflation in Jakarta.

2. Methodology

2.1. Data Sources

Data analyzed in the research were secondary data from Central Bureau of Statistics Jakarta Region and Bank Indonesia. The data collected in this study is monthly data from January 2004 to August 2017. The data were divided into 2 (two) groups that is the first 156 data into training data (January 2004 - December 2016) and the last 8 data as testing data (January 2017 - August 2017). Endogenous variables include 35 subcategories of commodity according to classification by from Central Bureau of Statistics. Those variables are subcategories of grains, tubers and its derived products (Y1); meat and its derived products (Y2); fresh fish (Y3); preserved fish (Y4); eggs, milk and its derived products (Y5); vegetables (Y6); beans (Y7); fruits (Y8); spices (Y9); fat and oil (Y10); other food (Y11); processed food (Y12); non-alcoholic beverages (Y13); tobacco and alcoholic beverages (Y14); residential cost (Y15); fuel, lighting and water (Y16); household appliances (Y17); housekeeping (Y18); men's clothing (Y19); women's clothing (Y20); children's clothing (Y21); personal goods and other clothing (Y22); health services (Y23); drugs (Y24); Physical care services (Y25); physical and cosmetic care (Y26); education services (Y27); course and training services (Y28); education equipment services (Y29); recreation (Y30); sports (Y31); transport (Y32); communication and delivery (Y33); transportation means and supporting facilities (Y34); and financial services (Y35). While exogenous variables consist of change of Rupiah exchange rate (X1), change of BI rate (X2), the beginning of school academic year (D1), changes of fuel price (D2), government policy othe than fuel (D3), Eid Al-Fitr Celebration (D4), inflation based on sub-category of commodity on cluster 1 (Yc1), inflation based on sub-category of commodity on cluster 2 (Yc2), inflation based on sub-category of commodity on cluster 3 (Yc3), inflation based on sub-category of commodity on cluster 4 (Yc4).

2.2. Procedure of Analysis

Data analysis and modeling were conducted using R and SAS statistical program. The followings are the steps performed in this study:

2.2.1. Exploring inflation data by subcategories of commodity

2.2.2. Applying TSClust technique on training data

a. The procedure was initiated by checking data stationary from 35 inflation variables according to subcategories of commodity by observing the corelogram patterns of ACF and PACF, ADF test, and Levene test [4]
b. Checking the matrix of inflation dissimilarity according to subcategories of commodity using ACF, CORR and Piccolo dissimilarity measures.

Autocorrelation Based Distance (ACF)

The distance of ACF between $\mathbf{x}_t$ and $\mathbf{y}_t$ is formulated as follows [5]:

$$d_{ACF}\,(\mathbf{x_t},\mathbf{y_t}) = \sqrt{\left(\hat{\boldsymbol{\rho}}_{xt} - \hat{\boldsymbol{\rho}}_{yt}\right)'\,\Omega\,\left(\hat{\boldsymbol{\rho}}_{xt} - \hat{\boldsymbol{\rho}}_{yt}\right)} \qquad (1)$$

where $d_{ACF}\,(\mathbf{x_t},\mathbf{y_t})$ = autocorrelation distance of vectors $\mathbf{x}_t$ and $\mathbf{y}_t$, Ω = weighing matrix, $\hat{\boldsymbol{\rho}}_{xt}$ = autocorrelation vector estimator $\mathbf{x}_t$, $\hat{\boldsymbol{\rho}}_{yt}$ = autocorrelation vector estimator $\mathbf{y}_t$. If the distance of ACF omits the weight, so the weighing matrix is presented as identity matrix.

Correlation Based Distance (CORR)

Simple criteria to measure the similarity of time series data is based on Pearson's Correlation between vectors x_T and y_T based on the following formula [6]:

$$\text{CORR}\,(x_t,y_t) = \frac{\sum_{t=i}^{T} (x_t - \bar{x}_t)\ (y_t - \bar{y}_t)}{\sqrt{\sum_{t=i}^{T} (x_t - \bar{x}_t)^2}\ \sqrt{\sum_{t=i}^{T} (y_t - \bar{y}_t)^2}} \qquad (2)$$

Where $\bar{x}_t$ and $\bar{y}_t$ are the average from x_t and y_t , based on the following distance matrix:

$$\text{dCORR.1} = (\mathbf{x_t},\mathbf{y_t}) = \sqrt{2\ (1 - CORR(x_t, y_t))} \tag{3}$$

Piccolo Based Distance (PICC)

Piccolo introduced the dissimilarity measures to cluster time series data based on ARIMA model. If the formulas $\hat{\boldsymbol{\Pi}}x_T = (\ \hat{\pi}_{1,xT,}\ldots\ \hat{\pi}_{k1,xT})^T$ and $\hat{\boldsymbol{\Pi}}_{yT} = (\ \hat{\pi}_{1,yT,}\ldots\ \hat{\pi}_{k1,yT})^T$ showing that AR (k_1) and AR (k_2) vectors are parameter estimations for $\mathbf{x}_T$ and $\mathbf{y}_T$ [7].

$$\text{dpic}\ (\mathbf{x}_T, \mathbf{y}_T) = \sqrt{\textstyle\sum_{j=1}^{k}\ (\hat{\pi}'_{\ j,x_T} - \hat{\pi}'_{\ j,y_T}\ )^2} \tag{4}$$

where: k_1 = AR ordo for $X_{T,}$ k_2= AR ordo for $Y_{T,}$ k= max (k_1,k_2) , $\hat{\pi}'_{\ j,x_T} = \hat{\pi}_{\ j,x_T}$, if $j \leq k_1$ and $\hat{\pi}'_{\ j,x_T} = 0$ for the others. $\hat{\pi}'_{\ j,y_T} = \hat{\pi}_{\ j,y_T}$, if $j \leq k_2$ and $\hat{\pi}'_{\ j,x_T} = 0$ for the others

2.2.3. Conducting clustering as many as k-cluster using k-average method based on the initial information from hierarchical clustering

2.3. Modeling cluster-level ARIMA based on Box-Jenkins procedure

Before the modeling, a set of time series data representing the pattern of each cluster was determined based on median measure approach [7]. *ARIMA* is the abbreviation of *Autoregressive Integrated Moving Average* that was introduced by Box and Jenkins in1976, that is also known as Box-Jenkin Model. In general, *ARIMA* (p,d,q) is formulated as follows [8]

$$\phi(B)(1 - B)^d Y_t = \theta(B)a_t \tag{5}$$

where B is identified as *backshift* operator $BY_t = Y_{t-1}$ dan $(1 - B)^d = \nabla^d$;

ϕ is the *autoregressive* parameter; θ is the *moving average* parameter; a_t is the error measures at period-t

In general, seasonal *ARIMA*(p,d,q)(P,D,Q)$_S$ model is formulated as follows

$$\Phi(B^S)\phi(B)(1 - B)^d(1 - B^S)^D Y_t = \theta(B)\Theta(B^S)a_t \tag{6}$$

where S is seasonal period; Φ is the seasonal *autoregressive* parameter; Θ is the seasonal *moving average* parameter

2.4. VARMA Modeling between clusters

VARMA Model is a multiple variable time series model developed from ARMA model. The formula of *VARMA* is defined as follows [9]:

$$\Phi_p(B)\boldsymbol{y}_t = \Theta_q(B)\boldsymbol{a}_t \tag{7}$$

where $\Phi_{\mathrm{p}}(B) = \Phi_0 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_{\mathrm{p}} B^p$, $\Theta_q(B) = \Theta_0 - \Theta_1 B - \Theta_2 B^2 - \cdots - \Theta_q B^q$

$\Phi_{\mathrm{p}} = \Theta_q = \mathrm{I}$, $B\mathbf{y}_t = \mathbf{y}_{t-1}$

and $\boldsymbol{y}_t$ = observed vector using $\boldsymbol{y}_t = [\mathrm{Y}_{1,t}, \mathrm{Y}_{2,t}\mathrm{Y}_{3,t}, \ldots, \mathrm{Y}_{n,t}]$ at n x 1, Φ_{p} = parameter matrix of autocorrelation vector order-p at nxn, Θ_q = matrix of moving average vector parameter order-q at nxn, B = backshift operator, $\boldsymbol{a}_t$ = *white noise* random vector assuming that $\boldsymbol{a}_t \sim MN(\mathbf{0}, \Sigma)$

2.5. VARMAX Modeling between clusters

VARMA model might include exogenous variables and is known as *Vector Autoregressive Moving Average Exogenous* (VARMAX) mode. Exogenous variables are defined variables and originated from outside the system, yet applied in model because of its effects. The study employed rupiah exchange rate, BI rate, and dummy variables as exogenous factors.

The formula of VARMAX model of (p,q,s) order is defined as follows [10]:

$$\phi(B)\boldsymbol{y_t} = \Theta_i^* (B)\boldsymbol{x_t} + \Theta(B)\boldsymbol{a_t} \tag{8}$$

where $\phi(B) = I_k - \sum_{i=1}^{p} \Phi_i B^i{}_1$, $\Theta_i^* (B)x_t = \Theta_0^* + \Theta_1^* B + \cdots + \Theta_s^* B^s$ dan $\Theta(B) = I_k - \sum_{i=1}^{q} \Theta_i B^i$

and $\boldsymbol{y}_t$ = observed vectors using $Y_t = [\mathrm{Y}_{1,t}, \mathrm{Y}_{2,t}\mathrm{Y}_{3,t}, \ldots, \mathrm{Y}_{n,t}]$ at nx1, Φ_{p} = matrix of auto regression order-p at nxn, Θ_q = matrix of moving average parameter order-q at nxn, Θ_i^* = matrix of parameter for exogenous variables at nxm for each i =1,2,…s, $\boldsymbol{x}_t$ = vectors of exogenous variables considering $\boldsymbol{x}_t = [x_{1,t}, x_{2,t} \ldots x_{n,t}]$ at mx1, B = backshift operator, $\boldsymbol{a}_t$ = *white noise* random vector assuming that $\boldsymbol{a}_t \sim MN(0, \Sigma)$

2.6. ARIMAX Modeling Cluster-level

Similar to previous VARIMAX model, ARIMA model includes exogenous variables and the model is known as *Autoregressive Moving Average Exogenous* (ARIMAX). Formula of *ARIMAX* model of (p,d,q) order is defined as follows [11]:

$$Y_t = \gamma t + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_p X_{p,t} + \frac{\theta(B)\Theta(B^S)a_t}{\phi(B)\,\Phi(B^S)} a_t \tag{9}$$

where $\phi(B) = I_k - \sum_{i=1}^{p} \Phi_i B^i{}_1$,$\Theta(B) = I_k - \sum_{i=1}^{q} \Theta_i B^i$

and Y_t = observation , ϕ_{p} = autoregression parameter at order-p , Φ_{p} = seasonal autoregression parameter at order-P , Θ_q = moving average parameter at order-q , Θ_Q = seasonal moving average parameter at order-Q , β_i = parameter coefficient for exogenous variables, $\mathrm{X_{i,}} X_{it}$ = exogenous variables stage i at period t, B = backshift operator, a_t = *white noise* random vector assuming that $a_t \sim MN(0, \Sigma)$

2.7. Model Evaluation

Evaluation of model in forecasting is conducted to ascertain whether the model has been feasible to be applied in forecasting in the next period [12]. One of the measurement for model evaluation is *Root Mean Squared Error* (*RMSE*). The model is considered good if the score of *RMSE* is relatively small. The score of *RMSE* is calculated based on the following formula:

$$\text{RMSE} : \sqrt{\frac{1}{n}\sum_{t=1}^{n}[\, y_t - \hat{y_t}\,(t-1)]^2} \qquad (10)$$

where: y_t = score at period t, t= 1,2,…, n, $\hat{y}_t(t-1)$ = forecast result at period t using data training up to t-1.

3. Result and Discussion

3.1. Data Exploration

Description of inflation data as seen on Figure 1 shows that inflation according to subcategories of commodity has various data distribution and outliers were observed in small and big measures. Subcategories of meat and its processed products (Y2); egg, milk and its processed products (Y5); vegetables (Y6); fruits (Y8), and spices (Y9) shows higher inflation measures compared to other variables. Those subcategories are included in volatile food by Indonesian Central Bureau of Statistics (Badan Pusat Statistik/BPS) since the movement is very susceptible to short-term shock such as fasting month, Eid al-Fitr Celebration and seasons. The availability of the commodity is highly dependent on its availability in the supplier area. Inflation data range also varies greatly among subcategories of commodity. The highest inflation rate was observed in subcategory of spices. The commodity had a minimum measure at -0.05 percent and the maximum measure was 35.77 percent. Subcategory of spices (Y9) experienced high inflation in December and during the celebration of Eid Al-Fitr.

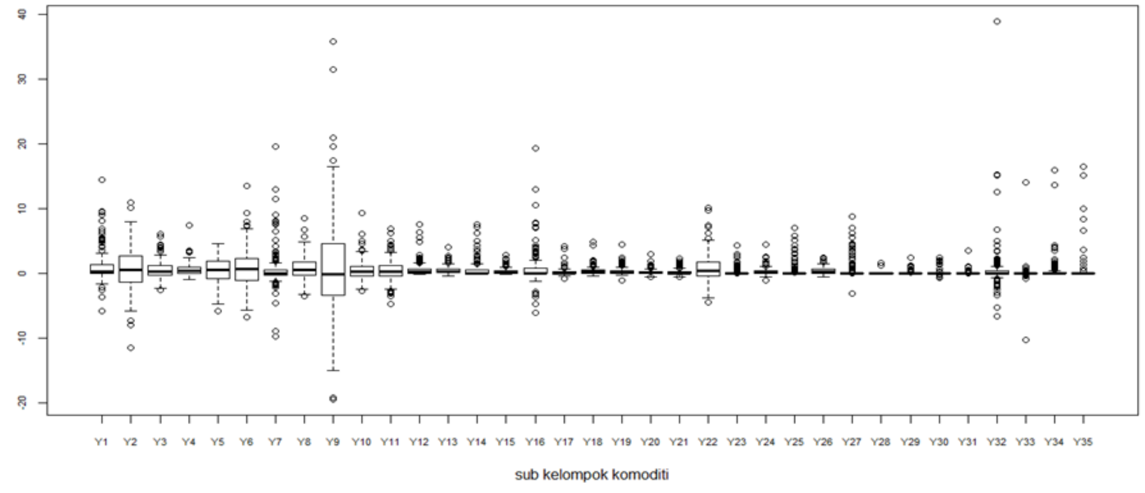

Figure 1: Distribution Pattern of Inflation Data based on Subcategories of Commodity

in Jakarta Year 2004-2016

3.2. Time Series Data Clustering

Before clustering time series data based on the distance of autocorrelation, correlation and piccolo, data must be stationary for its mean and variety. The result of plotting shows that the data of inflation based on subcategories of commodity shows no positive or negative trends. After stationary test was conducted, it was obvious that data are stationary for its mean, yet not stationary for its variety. Data were stationed by conducting data transformation using ln*(x+c)*. After that, data were hierarchically clustered based on the distance of autocorrelation, correlation and piccolo to obtain cophenetic correlation score presented in Table 1.

Table 1: Cophenetic correlation between Autocorrelation, Correlation, and Piccolo

No	Measures	Cophenetic correlation
1	Autocorrelation	0.955
2	Correlation	0.700
3	Piccolo	0.973

Table 1 shows the measures of autocorrelation, correlation, and piccolo result in the score of cophenetic correlation recorded as 0.954, 0.70 and 0.972. From the result, it is clear that the best measure to categorize inflation data according to subcategories of commodity was piccolo measure. As an illustration, the result of the clustering using these three types of hierarchical measures is presented by the dendogram as follows:

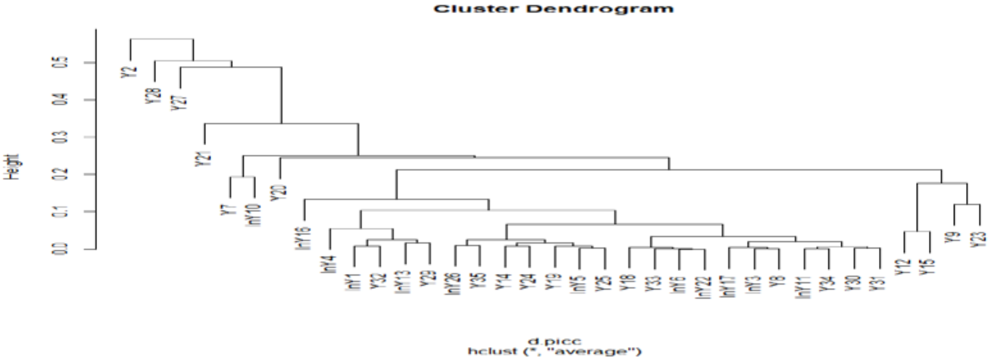

Figure 2: Dendogram of hierarchical clusters analysis using Piccolo's Measures of Dissimilarity

Figure 1 shows that the items in cluster 1 is the inflation of meat and its processed products (Y2), cluster 2 includes course services (Y28), cluster 3 includes educational services (Y27), and the others are included in cluster 4. The result of clustering shows the weakness of hierarchy method which tends to put outliers into separated cluster. Therefore, experts have started to combine hierarchical and non-hierarchical methods to establish better model. After the overview of clustering of inflation in Jakarta based on hierarchical method, the next clustering is based on k-means method by taking cluster 1 obtained from hierarchical clustering as the centroid. The combination of these methods were expected to establish better clusters.

The arrangement for the optimum number of classes was done by using elbow criteria by comparing the sum of squares within cluster (SSW) for each value of k. The elbow method will determine the actual number of clusters from a set of data. The value of SSW will continue to increase at each step and at a time the value will largely decrease. At that time, the elbow of all values of k will be established and the elbow becomes the desired value of k. The smaller value of SSW contributes to the greater number of sum squares between clusters (SSB), so the ratio of values between the two will increase significantly.

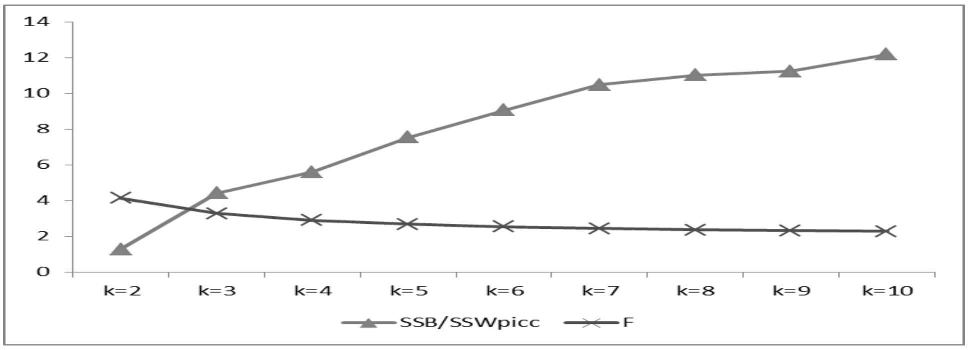

Figure 3: Comparison of the SSB by SSW at k = 2 to k = 10

Figure 3 compares SS score between cluster to SS within clusters is bigger from F table since k=3. Yet, a significant increase was observed at k=4. So, it can be concluded that the inflation according to subcategories of commodity would be clustered into 4. Next, the clustering was applied using k-means method and the result is presented in table 2.

Table 2: Inflation Clustering based on Subcategory of Commodities with Piccollo's Measures of Dissimilarity

Cluster	Members of Cluster
1	Y2, Y27, Y28
2	Y7, Y20, Y21
3	Y1*, Y3*, Y5*, Y6*, Y8, Y11*, Y13*, Y14, Y16*, Y17*, Y18, Y19, Y22*, Y24, Y25, Y26*, Y29, Y30, Y31, Y32, Y33, Y34, Y35
4	Y4*, Y9, Y10*, Y12, Y15, Y23

*) data in is presented in the form of transformation

Clustering result presented in Table 2 presents that every cluster have a unique data pattern. Cluster one is the cluster where the inflation measure tends to be stable and changes significantly in a certain period of time.

Fluctuation of inflation in Cluster 1 is affected by the high demand in July and August caused by the beginning of academic school year and Eid Al-Fitr celebration. Cluster two has bigger inflation measure and volatility compared to cluster one. Cluster three has bigger inflation measure and volatility compared to cluster one and two. Subcategories in cluster three mostly consist of strategic commodity whose volatility is prone to short term shock and government intervention. While the last cluster is the cluster with the highest measure and volatility compared to the other three. The inflation measure was noticeable at in several periods within a year. Subcategory in cluster 4 experienced inflation in January, July, November, and December.

3.3. ARIMA Inflation Modeling based on Sub-category of Commodity in Cluster-Level

After data exploration on each cluster, ARIMA modeling was conducted in each cluster. The data representing each cluster were obtained by searching for cluster's median data on each period of T, and then ARIMA modeling was conducted using Box-Jenkins procedure that is identifying ACF and PACF corelogram.

ARIMA modeling is a stationary time series model. From cluster-level inflation data plotting, the average and inflation data variables tend to be constant, so it does not require data transformation and differencing. The identification on ARIMA model resulted in the best model having the least AIC, significant parameters on significance level 0.05 and white noise residuals, yet the model had not met the assumption of normal distribution. The condition was a result from a number of outliers affecting the normality of residuals. Those outliers were allegedly caused by external factors affecting inflation like demand level, weather and intervention from the government. The best ARIMA model for each cluster is presented in Table 3.

Table 3: The Best ARIMA Model based on Cluster and Dissimilarity Measures

Cluster	Model	AIC
1	$Y_{c1,t} = 0.253 + 0.407Y_{c1,t-12} + \alpha_{c1,t}$	376.760
2	$Y_{c2,t} = 0.221 - 0.671Y_{c2,t-1} + 0.432Y_{c2,t-2} - 0.459Y_{c2,t-3} - 0.738Y_{c2,t-4}$ $+ 0.625\alpha_{c2,t-1} - 0.513\alpha_{c2,t-2} + 0.625\ \alpha_{c2,t-3} + \alpha_{c2,t-4}$ $+ \alpha_{c2,t-12} + \alpha_{c2,t}$	134.043
3	$Y_{c3,t} = 0.226 + 0.169Y_{c3,t-1} + \alpha_{c3,t}$	33.448
4	$Y_{c4,t} = 0.403 + 0.166\alpha_{c4,t-1} + \alpha_{c4,t}$	120.435

3.4. VARIMA Inflation Modeling based on Sub-category of Commodity between Clusters

VAR is a system of dynamic equations, with the prediction of a variable in a given period depending on the movement of the variables and other variables involved in the system in the previous period [13]. In the modeling, VAR only applied the endogenous variables between inflation of the sub-category of commodity. The application of endogenous variables means that the double time series model was performed simultaneously because inflation data movements of a commodity sub group took place along with or follow the movement of inflation data of other subcategory of commodity. After model identification and overfitting, VAR (2) is the best model supported by minimum AIC score and the result of portmanteau test was not significant up to lag 10. GARCH test result shows that p-value is bigger than 0.05 showing homogenous variety. There were many outliers resulting in abnormalities of residuals.

Table 4: VAR (2) model using Piccolo measures of dissimilarity

Cluster	Model
1	$Y_{c1,t} = 0.318 + 0.101Y_{c1,t-1} + 0.187Y_{c2,t-1} - 0.234Y_{c3,t-1} - 0.171Y_{c4,t-1}$ $+ 0.05Y_{c1,t-2} + 0.186Y_{c2,t-2} + 0.114Y_{c3,t-2} - 0.175Y_{c4,t-2} + \alpha_{c1,t}$
2	$Y_{c2,t} = 0.189 + 0.093Y_{c1,t-1} - 0.180Y_{c2,t-1} + 0.3Y_{c3,t-1} + 0.057Y_{c4,t-1}$ $+ 0.013Y_{c1,t-2} - 0.155Y_{c2,t-2} + 0.127Y_{c3,t-2} - 0.088Y_{c4,t-2} + \alpha_{c2,t}$
3	$Y_{c3,t} = 0.173 + 0.044Y_{c1,t-1} - 0.058Y_{c2,t-1} + 0.194Y_{c3,t-1} + 0.016Y_{c4,t-1}$ $- 0.019Y_{c1,t-2} - 0.049Y_{c2,t-2} - 0.010Y_{c3,t-2} + 0.062Y_{c4,t-2} + \alpha_{c3,t}$
4	$Y_{c4,t} = 0.343 - 0.018Y_{c1,t-1} - 0.036Y_{c2,t-1} + 0.231Y_{c3,t-1} + 0.114Y_{c4,t-1}$ $- 0.046Y_{c1,t-2} + 0.120Y_{c2,t-2} - 0.091Y_{c3,t-2} - 0.041Y_{c4,t-2} + \alpha_{c4,t}$

3.5. VARIMAX Inflation Model based on sub-category of commodity between clusters

To obtain VAR model that has better forecast precision, exogenous variables were added and used in the modeling as they affect endogenous variables. VAR model added by exogenous variables is known as *Vector Autoregressive Exogenous (VARX)* model. Exogenous variables contributing to inflation in Jakarta are rupiah exchange rate [14] and BI Rate [15].

After model identification and overfitting, VARX model was established by adding dummy variables namely VARX (2,0) model having AIC score -7.35 and portmanteau test result was insignificant up to lag 12. GARCH test result shows p-value bigger than 0.05 showing homogenous variety.

Although dummy variables were included, VARX(2,0) model residuals were not normally distributed. Outliers were observed in each cluster's residuals. The result of parameter estimation on VAR(2) model is presented in Table 4.3.

3.6. ARIMAX Inflation Model based on sub-category of commodity between clusters.

To obtain a model with better accuracy, exogenous variables were added to ARIMA cluster-level model. ARIMA modeling by adding exogenous variables on each cluster (ARIMAX) resulted in models with smaller AIC and RMSE compared to ARIMA cluster-level model.

Table 5: VARX Model(2,0) using Piccolo measures of dissimilarity

Cluster	Model
1	$Y_{c1,t} = 0.042 + 0.036Y_{c1,t-1} + 0.04Y_{c2,t-1} - 0.084Y_{c3,t-1} - 0.193Y_{c4,t-1} + 0.148Y_{c1,t-2}$ $+ 0.065Y_{c2,t-2} + 0.271Y_{c3,t-2} - 0.028Y_{c4,t-2} + 0.27X_{2,t} + 1.26D_1$ $+ 0.004D_2 + 0.04D_3 - 0.171D_4 + \alpha_{c1,t}$
2	$Y_{c2,t} = 0.142 + 0.092Y_{c1,t-1} - 0.2Y_{c2,t-1} + 0.273Y_{c3,t-1} + 0.072Y_{c4,t-1} + 0.017Y_{c1,t-2}$ $- 0.14Y_{c2,t-2} + 0.183Y_{c3,t-2} - 0.078Y_{c4,t-2} - 0.066X_{2,t} - 0.003D_1$ $+ 0.527D_2 + 0.133D_3 + 0.156D_4 + \alpha_{c2,t}$
3	$Y_{c3,t} = 0.178 + 0.044Y_{c1,t-1} - 0.096Y_{c2,t-1} + 0.118Y_{c3,t-1} - 0.004Y_{c4,t-1} - 0.014Y_{c1,t-2}$ $- 0.062Y_{c2,t-2} + 0.021Y_{c3,t-2} + 0.073Y_{c4,t-2} - 0.279X_{2,t} - 0.035D_1$ $+ 0.118D_2 + 0.309D_3 + 0.117D_4 + \alpha_{c3,t}$
4	$Y_{c4,t} = 0.285 - 0.016Y_{c1,t-1} - 0.054Y_{c2,t-1} + 0.246Y_{c3,t-1} + 0.114Y_{c4,t-1} - 0.044Y_{c1,t-2}$ $+ 0.135Y_{c2,t-2} - 0.023Y_{c3,t-2} - 0.028Y_{c4,t-2}$ $+ 0.018X_{2,t} - 0.009D_1 + 0.403D_2 + 0.292D_3 + 0.137D_4 + \alpha_{c4,t}$

Table 6: Estimation of parameters on cluster-level ARIMAX model

Cluster	Model
1	$Y_{c1,t} = 0.073 + 0.0244X_{2,t} + 1.145D_1 + 0.038D_2 - 0.021D_3 - 0.068D_4 + \alpha_{c1,t}$ dimana, $\alpha_{c1,t} \sim ARIMA(0,0,0)(1,0,0)12$
2	$Y_{c1,t} = 0.2051 - 0.0272X_{2,t} + 0.021D_1 + 0.338D_2 + 0.011D_3 + 0.064D_4 + \alpha_{c2,t}$ dimana, $\alpha_{c2,t} \sim ARIMA(4,0,4)(0,0,2)12$
3	$Y_{c1,t} = 0.2015 - 0.251X_{2,t} - 0.017D_1 + 0.141D_2 + 0.303D_3 + 0.115D_4 + \alpha_{c3,t}$ dimana, $\alpha_{c3,t} \sim ARIMA(3,0,0)$
4	$Y_{c1,t} = 0.073 - 0.0244X_{2,t} - 1.145D_1 + 0.038D_2 - 0.021D_3 - 0.068D_4 + \alpha_{c4,t}$ dimana, $\alpha_{c4,t} \sim N$ ((0, Σ)

3.7. Model Evaluation

The best model is decided based on the model's cluster-level of accuracy to predict individual level inflation based on RMSEP score.

The best model should have the least RMSEB score. ARIMAX model resulted in the least RMSEP average on training data and testing data.

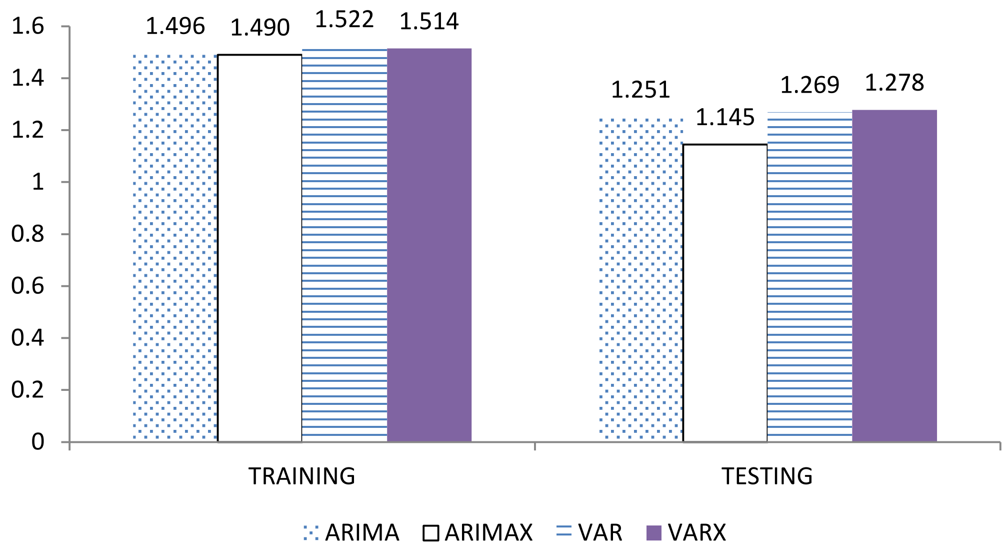

Figure 4: RMSEP of ARIMA, ARIMAX, VAR and VARX models according to Piccolo measures of dissimilarity to training data and testing data

4. Conclusion

The best dissimilarity measures used to cluster inflation based on commodities sub-category in Jakarta was the piccolo approach. Using piccolo approach, inflation by commodity sub group was categorized into 4 clusters. Each cluster had different values, variations and patterns, but cluster-level ARIMAX model made better forecast to 35 inflation variables compared to the other three models.

5. Recommendation

Further research might consider other explanatory variables affecting inflation to provide more models that are accurate. Besides that, a similar method can also be applied in other areas or to the inflation variables by commodity.

Acknowledgements

The authors would like to acknowledge the valuable comments and suggestions of the Editors. These led to a considerable improvement in the paper. This work is financially supported by National Statistical Office (BPS) scholarship program in collaboration with Department of Statistics at Bogor Agricultural University.

References

[1] [BI] Bank Indonesia. 2017. Pengenalan Inflasi [Internet]. [downloaded 2017 Maret 4]. Available on : http : // www.bi.go.id / id / moneter / inflasi / pengenalan / Contents / Pentingnya. aspx.

[2] [BI dan Kemendagri] Bank Indonesia, Kementrian Dalam Negeri, 2014. Buku Petunjuk TPID. Jakarta : Bank Indonesia.

[3] [BPS] Badan Pusat Statistik. 2012. Diagram Timbang Indeks Harga Konsumen. Jakarta:Badan Pusat Statistik.

[4] Cryer JD, Chan K. 2008. Time Series Analysis with Applications in R 2nd Ed. Iowa (US): Springer.

[5] Galeano P, Pena D (2000). \Multivariate Analysis in Vector Time Series." Resenhas do Instituto de Matematica e Estaistica da Universidade de Sao Paulo, 4(4), 383{403.

[6] Golay X, Kollias S, Stoll G, Meier D, Valavanis A, Boesiger P (2005). \A New Correlation-Based Fuzzy Logic Clustering Algorithm for fMRI." Magnetic Resonance in Medicine, 40(2), 249{260.

[7] Piccolo D (1990). \A Distance Measure for Classifying ARIMA Models." Journal of TimeSeries Analysis, 11(2), 153{164.

[8] Aghabozorgi S, Shirkhorshidi AS, Wah TY. 2015. Time-series clustering – A decade review. Information Systems. 2015 : 16-38.

[9] Cryer JD, Chan K. 2008. Time Series Analysis with Applications in R 2nd Ed. Iowa (US): Springer.

[10] Wei WWS. 2006. Time Series Analysis, Univariate and Multivariate Methods, Second Edition. New York (US): Pearson Education, Inc.

[11] SAS Institute Inc.2011.SAS/ETS 9.3 User’s Guide. Cary, NC: SAS Institute Inc. [Internet]. [downloaded 2017 April 28]. Available on: https: // support.sas.com / documentation / cdl /en/etsug/63348/HTML/default/viewer.htm#etsug_varmax_sect025.htm

[12] Cryer JD, Chan K. 2008. Time Series Analysis with Applications in R 2nd Ed. Iowa (US): Springer.

[13] Montgomery DC, Jennings CL, Kulahci M. 2008. Introduction to Time Series Analysis and Forecasting, Third Ed. New Jersey (US) : Jon Wiley & Sons. Inc.

[14] Enders W. 2004. Applied Econometric Time Series. Second edition. Canada (US): John Wiley and sons.

[15] Pratiwi, AP, Ferry. 2013. Determinan Inflasi di Indonesia: Analisis Jangka Panjang dan Pendek [thesis], Surabaya : Universitas Brawijaya.

[16] [BI] Bank Indonesia. 2017. Penjelasan BI Rate [Internet]. [downloaded 2017 Maret 4]. Available on : http : // www.bi.go.id / id / moneter / bi-rate / penjelasan / Contents / Default. aspx.