Clustering of Commodity Inflation Pattern based on Estimated ARIMA Model

T Hendrawati$^{1,2}$, A H Wigena$^2$, I M Sumertajaya$^2$, and B Sartono$^3$

$^1$Department of Statistics, Padjadjaran University, Bandung, Indonesia
$^2$Department of Statistics, IPB University, Bogor, Indonesia

*E-mail: triyani.hendrawati@unpad.ac.id

Abstract. Clustering is the stage that is carried out before further data analysis. There are many approach method that can be used for clustering time series data, one of which is the model-based approach. In this study, we clustering inflation data used the ARIMA model. The cluster model is carried out after obtaining the clusters. The similarity between time series is measured using the development of Piccolo distance. Furthermore, the Ward hierarchical method is used for clustering. The Silhouette averaging method is used to determine the optimal number of clusters. The cluster model can be used to represent individual models. The cluster model is more effective than creating all individual models.

1. Introduction
Time series data clustering developed study today. There are many applications of time series data clustering in various fields, for example in science, economics, and medicine. Several methods can be used for time-series data clustering, one of which is the model-based approach. One of the advantages of clustering is that makes it easier to analyse time-series data, especially large ones. Clustering is done as a first step before further analysis or modelling is carried out [1, 2, 3].

Piccolo [4, 5] introduced the distance between time series data using the ARIMA model approach. The selection of the ARIMA model is based on the smallest AIC or BIC criteria. The development of the Piccolo distance method using multi models based on the five criteria for selecting the model has been carried out. It has obtained a better clustering accuracy rate than the basic Piccolo distance [6].

The increase in inflation rate is one of the problems faced by many countries today, especially for developing countries such as Indonesia. The Central Statistics Agency (BPS) annually publishes the value of the national inflation/deflation rate in Indonesia [7, 8]. Inflation represents an increase in prices while deflation is for a decrease in prices [9]. The number of categories / subcategories of national inflation data is large, of course making individual models takes time. The use of the cluster model certainly makes work more efficient. The purpose of this research is to group the national inflation data in Indonesia with the Piccolo distance development method and to make a cluster model with the ARIMA model.

2. Theory

2.1. The development of Piccolo distance
The development method of the piccolo distance used five models with different selection criteria [6]. The distance between series \(X_t\) and \(Y_t\) is:

\[
d(X_t, Y_t)_{new} = [(\bar{\rho}_x - \bar{\rho}_y)(\bar{\sigma}_x - \bar{\sigma}_y)]^\frac{1}{2} = \left( \sum_{j=1}^{p} (\bar{\rho}_{j,x} - \bar{\rho}_{j,y})^2 \right)^{\frac{1}{2}}
\]

(1)

where \(\bar{\rho}_x = (\bar{\rho}_{1,x}, \bar{\rho}_{2,x}, \ldots, \bar{\rho}_{p,x})\) and \(\bar{\rho}_y = (\bar{\rho}_{1,y}, \bar{\rho}_{2,y}, \ldots, \bar{\rho}_{p,y})\) is the average parameters estimator of the autoregressive model (AR(\(p\))) for time series data \(X_t\) and \(Y_t\). The average parameters estimator is obtained by averaging the five autoregressive (AR(\(p\))) model parameters obtained based on the AIC, BIC AICC, MAPE, and RMSE criteria.

### 2.2. ARIMA Model

ARMA Model is mixed of autoregressive (AR(\(p\))) and moving average (MA(\(q\))) model. Autoregressive moving average model (ARMA(\(p, q\))) is expressed by the following equation:

\[
Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}
\]

(2)

while Autoregressive Integrated Moving Average Model (ARIMA(\(p, d, q\))) is non-stationary ARMA(\(p, q\)) model which has been differencing with order \(d\). The ARIMA model is written as backshift operator:

\[
\phi(B)(1-B)^d Y_t = \theta(B) e_t
\]

(3)

which:

\(\sigma^2_e > 0\); \(e_t\) is a random component that is mutually independent assuming \(e_t \sim N(0, \sigma^2_e)\); \(\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\); \(\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\); \(p, \ d, \ q = \) order of AR, differencing, and MA [10].

### 2.3. Model Selection Criteria

In this study, the criteria used for selecting the best model used five types namely Akaike’s Information Criterion (AIC), Bayesian Information Criterion (BIC), Akaike’s Information Criterion Bias Corrected (AICc), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). The formula of AIC is written as:

\[
AIC = -2 \log(\hat{\sigma}^2_e) + 2k
\]

(4)

The BIC formula is:

\[
BIC = -2 \log(\hat{\sigma}^2_e) + k \log(n)
\]

(5)

The AICc is a bias parameter corrected AIC, formula is written as:

\[
AICc = AIC + \frac{2(k+1)(k+2)}{n-k-2}
\]

(6)

The MAPE is formulated as:

\[
MAPE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t| \times 100}{n}
\]

(7)

The RMSE is formulated as:

\[
RMSE = \left( \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n} \right)^{1/2}
\]

(8)
where: $n$ is the number of observations, $k$ is the number of parameters, $\hat{\sigma}^2_e$ is the maximum likelihood estimator for the variance of the error, $y_t$ is the observation at time-$t$, and $\hat{y}_t$ is the value of the fit of observation at time-$t$.

2.4. Ward method

The clustering method used is the Ward method. This method is a hierarchical clustering method. Ward method aimed to join elements into clusters so that the variance within clusters is minimized. Each step combines the object pairs with the smallest increase in error sum of squares (ESS).

$$\text{ESS} = \sum_{j=1}^{n} (x_j - \bar{x})'(x_j - \bar{x})$$

where: $x_j$ is a j-th object and $\bar{x}$ is an average of objects [10, 13, 15, 16].

2.5. Determination of the number of clusters

The number of clusters is determined using the average silhouette index. The average silhouette index is the average distance from a data point to all other data points in its cluster and all data points in the nearest neighbor cluster.

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

where $a(i)$ is the distance of the i-th object to its own cluster, $b(i)$ is the distance of the i-th object to its nearest neighbor cluster [13, 18]. The maximum value of the average Silhouette is used to determine the optimal number of clusters.

2.6. Prototype

The cluster model requires time-series data that represents the cluster. This time-series data is called a prototype. In this study, the determination of the prototype uses the average value of the cluster data.

$$\text{prototype} = \text{average}\{y_{1,i}, y_{2,i}, \ldots, y_{c,i}\} \forall c \in C$$

which $i = 1,2,\ldots,N$ [1].

3. Methodology

The data used in this study are secondary in the form of national monthly inflation data according to sub-categories of 24 commodities in Indonesia from 2013 to 2019 published by the Central Statistics Agency (BPS) [7, 8]. The name of the commodity object used namely: (Y1) Grains, tubers, and their products; (Y2) Meat and their products; (Y3) Fresh fish; (Y4) Preserved fish; (Y5) Eggs, milk, and their products; (Y6) Vegetables; (Y7) Fruits; (Y8) Seasonings; (Y9) Fat and oil; (Y10) Processed Foods; (Y11) Non-alcoholic drinks; (Y12) Tobacco and alcoholic beverages; (Y13) Housing costs; (Y14) Household items; (Y15) Women's clothing; (Y16) Men's clothing; (Y17) Children's clothing; (Y18) Personal items and other clothing; (Y19) Health services; (Y20) Medicines; (Y21) Physical care services; (Y22) Physical care and cosmetics; (Y23) Recreation; (Y24) Sports.

The inflation data was clustered used the development of Piccolo distance, then made a cluster model with the ARIMA model. The algorithm is as follows:

1. Data exploration.
2. Perform clusters, the process carried out in clustering namely:
   a) Make five ARIMA models of inflation data with different model selection criteria, namely using the smallest AIC, BIC, AICc, RMSE and MAPE.
   b) Determine the coefficient $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_p)$ of the autoregressive (AR ($p$)) model; the parameters estimator are $\hat{\pi}_{\text{AIC}}, \hat{\pi}_{\text{BIC}}, \hat{\pi}_{\text{AICc}}, \hat{\pi}_{\text{RMSE}},$ and $\hat{\pi}_{\text{MAPE}}$.
   c) Calculate the average of parameters estimator [9]
\[
\vec{n} = \left[ \vec{n}_1 \ldots \vec{n}_p \right] = \left[ \frac{1}{5} \left( \hat{r}_{1,AIC} + \hat{r}_{1,BIC} + \hat{r}_{1,AICC} + \hat{r}_{1,RMSE} + \hat{r}_{1,MAPE} \right) \ldots \frac{1}{5} \left( \hat{r}_{p,AIC} + \hat{r}_{p,BIC} + \hat{r}_{p,AICC} + \hat{r}_{p,RMSE} + \hat{r}_{p,MAPE} \right) \right]
\] (12)

d) Calculate the distance between series uses the development of Piccolo distance with formula (1).

e) Clustering with the Ward hierarchical clustering method uses the formula in equation (9).

f) Determine the optimal number of cluster use the average of Silhouette index with formula (10).

3. Determine the prototype data for each cluster use formula (11).

4. Create the ARIMA model for cluster level use BIC as selection criteria.

5. The results of clustering were evaluated by compared the RMSE values of cluster-level models and individual-level models.

6. Forecasting the cluster model for the next six-period [19].

4. Results and Discussion

The box plot of national inflation data is presented in Figure 1. In Figure 1, it can be seen that most of the commodities have low variance, only a few commodities have large variances, such as Y2, Y5, Y6, and Y8. The Y8 commodity has the highest variance of inflation values.

![Figure 1. The box plot of national inflation data](image)

The results of the clustering illustrated by a dendrogram as in Figure 2. Based on Figure 2, it can be seen by several possible clusters formed, but we do not have the option to determine the number of clusters. The Silhouette method is used to determine the optimal number of clusters. Figure 3 contains the average of Silhouette for one to 10 cluster. The highest average of Silhouette is 0.2884988. This occurs in five clusters, so the optimal number of clusters is five clusters.
Figure 2. The dendrogram of national inflation data

Figure 3. Plot the average of Silhouette

The results of clustering and its members can be seen in table 1. It can be seen that five clusters are formed which are named clusters A, B, C, D, and E. Cluster B has only two member, while cluster E has the most members i.e. 11.
Table 1. The clusters and members

| Cluster | Number of members | Cluster’s members          |
|---------|------------------|----------------------------|
| A       | 3                | Y1, Y2, Y4                 |
| B       | 2                | Y17, Y18                   |
| C       | 3                | Y8, Y9, Y10                |
| D       | 5                | Y3, Y5, Y6, Y7, Y16        |
| E       | 11               | Y11, Y12, Y13, Y14, Y15, Y19, Y20, Y21, Y22, Y23, Y24 |

Figure 4 is shown the plot of the prototype and the selected members. Commodities Y2 and Y4 are members of cluster A. Commodities Y6 and Y7 are members of cluster D. It appears that on the same group has a similar pattern. If in a cluster there is one commodity with high inflation, it can be estimated that other commodities in the cluster are also high. The same thing will happen, if there is low inflation for a commodity in a group then other commodities in that group are also low. This information is useful for predicting which commodities are estimated to have high or low inflation due to the influence of the inflation of other commodities in the same cluster.

Figure 4. Plot prototype and the selected members

Prototype A          Prototype D

Y2                      Y6

Y4                      Y7
Table 2. The model for clusters

| Cluster | Cluster Model \(\text{Equation Model} \) |
|--------|----------------------------------|
| A      | ARIMA(2,0,2) \(Y_t = e_t + 0.9268Y_{t-1} - 0.9290Y_{t-2} + 0.5188e_{t-1} - 0.7042e_{t-2}\) |
| B      | ARIMA(0,0,1) \(Y_t = e_t - 0.4766e_{t-1}\) |
| C      | ARIMA(3,0,1) \(Y_t = 1.2215Y_{t-1} - 0.5497Y_{t-2} + 0.3282Y_{t-3} + 0.9914e_{t-1} + e_t\) |
| D      | ARIMA(0,0,1) \(Y_t = e_t - 0.6989e_{t-1}\) |
| E      | ARIMA(0,1,1) \(\nabla Y_t = e_t + 0.7953e_{t-1}\) |

Table 3. The average of RMSE prediction and forecast for model with and without clustering

| Model                  | RMSE Prediction | RMSE Forecast |
|------------------------|-----------------|---------------|
| With clustering        | 1.473169        | 0.966297      |
| Without clustering     | 1.583138        | 0.969892      |

The ARIMA model for each cluster is obtained as in Table 2. The model of cluster A is ARIMA(2,0,2); cluster B and D have the same model i.e. ARIMA(0,0,1) but different in coefficient of model; cluster C is ARIMA(3,0,1); and cluster E is ARIMA(0,1,1). Evaluation of cluster is carried out by comparing the RMSE model with clustering and without clustering. If the results are not significantly different, then the cluster model can be used to represent the individual models in one cluster.

Table 3 shows the average RMSE prediction and forecast value for models with or without clustering. We used the mean difference test to compare similarity between RMSE models with or without clustering at \(\alpha = 0.05\). The \(p\)-value of RMSE prediction is 0.201 while the \(p\)-value of RMSE forecast is 0.9118. The both of \(p\)-value is greater than 0.05, in other words there is no difference between the mean of RMSE models with or without clustering. So the cluster model can be used to represent individual models in the same cluster.

Figure 5 shows the forecast of clusters A and D for the next six periods. In cluster A model, the forecast value at point 85 to 90 are 0.7572592, -0.3964426, -1.0839907, -0.6426319, 0.4178497, and -0.9961906. The forecast for cluster D model at point 85 is 0.5103478, while for point 86 to 90 the value is 0.
Figure 5. Plot forecasting for the next six periods (a) cluster A and (b) cluster D

5. Conclusion
In this study, we clustering national monthly inflation data in Indonesia using the development of Piccolo distance. The optimal number of clusters is five. The model of cluster A is ARIMA (2, 0, 2); cluster B and D have the same model i.e. ARIMA (0, 0, 1) but different in coefficient of model; cluster C is ARIMA (3, 0, 1); cluster E is ARIMA (0, 1, 1). The RMSE of the individual model was not different from the RMSE cluster model at the level of confidence $\alpha = 0.05$, so the cluster model could be used to represent individual models in the same cluster. The cluster model more efficient than the individual model. The data on the same cluster has a similar pattern. It has consequences, if there is an increase in inflation for a commodity in a group, then other commodities in that group are also increase.
References

[1] Aghabozorgi S, Shirkhorshidi S, and Wah TY 2015 Time-series clustering - A decade review Information Systems 53 16–38
[2] Liao TW 2005 Pattern Recognition 38 (11) pp 1857–1874
[3] Montgomery DC, Jennings CL, and Kulachi M 2008 Introduction to Time Series Analysis and Forecasting (New Jersey: John Wiley & Sons Inc)
[4] Piccolo D 1990 A Distance Measure for Classifying Arima Models Journal of Time Series Analysis 11 153–164
[5] Piccolo D 2010 The Autoregressive metric for comparing time series models Statistica 70 pp 459–480
[6] Hendrawati T, Wigena AH, Sumertajaya IM, Sartono B 2020 A new approach to clustering time series data using the ARIMA model uncertainty Commun. Math. Biol. Neurosci. Article ID 52
[7] BPS Provinsi DKI Jakarta 2018 Indeks Harga Konsumen dan Inflasi DKI Jakarta (Jakarta: BPS Provinsi DKI Jakarta)
[8] BPS Provinsi DKI Jakarta 2019 Indeks Harga Konsumen dan Inflasi DKI Jakarta (Jakarta: BPS Provinsi DKI Jakarta)
[9] Schofield NC and Bowler T 2011 Trading the Fixed Income Inflation and Credit Markets Hoboken (NJ: John Willey & Sons Ltd.) pp 2–5
[10] Cryer J D and Chan K 2008 Time Series Analysis with Application in R second edition (New York: Springer)
[11] Akaike 1973 Information theory and an extension of the maximum likelihood principle 2nd International Symposium on Information Theory pp 267–281
[12] Schwarz 1978 Estimating the dimension of a model Ann. Statist. 6 pp 461–464
[13] Charrad M, Ghazzali N, Boiteau V and Niknafs A 2014 Journal of Statistical Software 61 Issue 6
[14] Jain AK and Dubes RC 1988 Algorithms for Clustering Data (New Jersey: Prentice Hall)
[15] Eszergár-Kíss D and Caesar B 2017 Definition of user groups applying Ward’s method Transportation Research Procedia 22 pp 25–34
[16] Everitt B S, Landau S, Leese M, Stahl D 2011 Cluster analysis 5th edition (Wiley)
[17] Murtagh F and Legendre P 2014 Journal of Classification 31 274-295
[18] Rousseeuw PJ 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis Journal of computational and applied mathematics 20 pp 53–65
[19] Hyndman RJ and Khandakar Y 2008 Automatic Time Series Forecasting: The forecast Package for R Journal of Statistical Software Volume 27 Issue