DKI Jakarta vegetable food commodity inflation modeling with tsclust approach using k-error method

I M Sumertajaya¹, M N Aidi¹, E Safitri¹*, Y Rahkmawati¹

¹Department of Statistics, IPB University, Bogor, 16680, Indonesia

*Corresponding author: emeyliasa@gmail.com

Abstract. Inflation is a critical indicator in showing economic symptoms about prices in a region. The Province of DKI Jakarta is the nation's capital city which plays an important role in the national economy including national inflation. An interesting commodity group is foodstuffs because it contains commodities that play a significant role in inflation in DKI Jakarta. TSClust approach was able to reduce the model by first bundling the commodities together. K-Error is a clustering method that incorporates the error element related to data to be clustered into the clustering step. The purpose of this study is to forecast inflation in Jakarta's agricultural food commodities with the K-Error. Clustering with the K-Error resulted in a lower pseudo F value as the number of clusters increases. Next, the selection of optimum number of clusters was based on a stable pseudo F value, such as the value k = 4 to k = 8. Based on the RMSE value, the optimum number of clusters selected is 5 clusters with unique models. Based on the RMSE value, it can be seen that forecasting at the individual level produce RMSE values that were not significantly different from RMSE level of the cluster for each commodity.

1. Introduction

Cluster analysis is one of the method in the analysis of multiple variables that can cluster n objects in m cluster with m<n based on the similarity between objects. Objects that have similar time series data movement patterns will be in one cluster. One method of cluster analysis that can be applied to time series objects is K-Error. Development in this method is done by inserting error related data (measurement error) into the clustering algorithm. K-Error can be applied for clustering time series. The basic error based clustering is similar to the concept of Piccolo distance and Maharaj distance [1]. Wang [2] evaluated 9 distances measures and DTW is the most accurate and Łuczak [3] constructed a distance function, combining DTW and the derivative DTW. Another distance that can be used for time series data is CID distance but if the two objects being compared have the same complexity, then the CID distance will be the euclidean distance [4]. K-Error method use Mahalanobis distance to include the error of parameter estimations.

Previous research by Ariski [5] used cluster analysis as pre-processing before time series modeling. The clustering method by entering measurement errors related to data can produce better cluster results and different from the results of the classic cluster analysis. When objects have different models, special handling in individual level modeling is needed so that each object has the same dimensions. One approach taken is to use the AR(p) model which refers to Piccolo theory. Each object will be modeled directly into the AR(p) model. Previous research by Safitri [6] used K-Means method as pre-processing for forecasting inflation in Jakarta's agricultural food commodities with 2
approaches such as original model and AR(p) model in individual modeling. Based on that study the AR(p) approach produced more efficient model with lower RMSE. The selection of p value is based on the highest AR order on the original model of all objects.

The K-Error method uses the ARIMA model for each time series data object and then measures the dissimilarity between the corresponding models using parameters. The ARIMA model is one of the time series modeling techniques commonly used for forecasting. Previous research by Hartati [7] conducted forecasting national inflation using simple method such as exponential smoothing and ARIMA. Previous research by Sen [8] conducted forecasting energy consumption and GHG emission with ARIMA. Forecasting with the TSClust approach using the K-Error method will be applied to the problem of inflation. Inflation is a problem that always arises in a country, especially developing countries. National Inflation is a reflection of regional inflation [9]. The increase in the price of daily necessities from time to time indicates inflation. The Province of DKI Jakarta is the country's capital city which is the center of the country's economic activities. DKI Jakarta’s inflation value has the same pattern as national inflation. Inflation is calculated based on several commodity groups. The commodities that are often the main contributors to core inflation in DKI Jakarta are food commodities, especially the agricultural sector, so the estimation of the value of inflation is quite important. The data used is secondary data, namely the monthly inflation data for Jakarta's agricultural commodities. The data used is from January 2012 to December 2018. The data is sourced from the Jakarta BPS publications, the Jakarta Consumer Price Index and Inflation [10]. The data will be divided into 2 (two), which is 78 training data (January 2012-June 2018) and 6 test data (July 2018-December 2018).

Direct ARIMA modeling of all food commodities in the agricultural sector is ineffective due to the large number of commodities. TSClust approach is able to reduce the number of commodity models to as many as number of clusters. Previous research by Adinugroho [11] conducted a TSClust approach for forecasting the price of cooking oil in Indonesian provinces. Previous research by Utami [12] conducted a TSClust approach for forecasting inflation based on subcategories of commodity. Based on that study the TSClust approach produced more efficient model with lower number of clusters. The concept of TSClust approach is to cluster time series objects into several clusters by first changing the time series objects into ARIMA parameters. The ARIMA parameter basically has an element of estimation error, the K-Error method takes this into account in calculating the center of the cluster and the distance. Paul [13] conducted research on the selection of the best ARIMA model using AIC, AICC, BIC, AME, MAPE, RMSE. Therefore, the purpose of this study is to forecast inflation in Jakarta's agricultural food commodities using the TSClust approach with the K-Error method.

2. Research Method

Data Analysis used software R Studio 3.5.2. The steps of analysis will be by the following:

1. Data exploration.
2. Dividing data into training data and test data.
3. Individual modeling in the initial stages will identify the model for each object based on the smallest BIC criteria of all model candidates. The general form of ARIMA (p,d,q) is as follows:

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

(1)

\(B\) is called a backshift operator, \(BY_t = Y_{t-1}\) and \((1 - B) = \nabla d\)
\(\phi\) is autoregressive parameter; \(\theta\) is moving average parameter; \(a_t\) is the error value at t-time.

4. Compile data containing parameters and variance-covariance matrixes for all 43 DKI Jakarta agricultural food commodities.
5. Perform a time series clustering with the K-Error method.
6. Calculating prototype [14]. Select some number of clusters based on stable pseudo F values.
7. Perform cluster level modeling.
   a. Model identification.
This process begins by checking the stationary data, which performs the Augmented Dickey-Fuller (ADF) test.

b. Estimating the model parameters.
The candidate model Bayesian information Criterion (BIC) value will be obtained. Determination of the best model of all candidates is the model with the smallest BIC value.

c. Model evaluation is by testing the residuals models obtained.
Good model is a model that has a random residuals. Significance of the remaining ACF and PACF tests was performed using the Ljung-Box test.

8. Evaluate cluster level and individual level modeling based on RMSE values.
9. Do forecasting for 9 periods.

3. Results and Discussion

3.1. Description of The Monthly Inflation of Jakarta's Agricultural Food Commodities

Figure 1 shows that the distribution pattern of commodity inflation is not symmetric with many outliers on large values and small values. In addition, the width of box-plot is also not the same which illustrates that the variance of inflation rate data between commodities is not homogeneous. Commodities Y41 (red chillies), Y36 (shallots) have a greater variety of data than other commodities. Most commodities have a narrow-line box diagram, such as the fruit group commodity, Y25 to Y35.

![Figure 1](image)

**Figure 1.** Data distribution of inflation rate according to commodity in January 2012-June 2018

3.2. Determination of Optimum Number of Clusters
The K-Error method is a type of non-hierarchical clustering. This method requires the initial determination of the number of clusters (k). K-Error is the development of K-Means method. K-means algorithm delivers better efficiency for clustering huge amounts of data [15]. K-Error is the distribution-based clustering method. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult [16]. The number of clusters tried was k = 2 to k = 10. Figure 2 presents the pseudo F values for the K-Error method. The determination of the number of clusters based on the Sw / is starting to stabilize. The overall pseudo F values appear to be sloping, steep decreases are seen on the way to k = 3 and k = 5. Therefore, the k value chosen is from k = 4 to k = 8.
3.3. Cluster Level Modeling with ARIMA

Table 1 presents the RMSE values for several k values. Determination of the best model of all candidates for each cluster is the model with the smallest BIC value [17]. The RMSE value is the average of the RMSE in each group. The first, second, and third lowest RMSE values namely when k = 4, 7, 8. The cluster level model that is produced is not unique because there are 2 clusters that have the same model, namely ARIMA (0,0,0) with a zero intercept value. The fourth lowest RMSE value is when the number of cluster (k) is 5 clusters. The resulting cluster level model is unique because each group has a different model. Therefore, the number of cluster (k) = 5 becomes the optimum cluster in the clustering of food commodities from the province of DKI Jakarta.

| Number of clusters (k) | Average RMSE |
|------------------------|--------------|
| 4                      | 0.00742      |
| 5                      | 0.00871      |
| 6                      | 0.00992      |
| 7                      | 0.00801      |
| 8                      | 0.00804      |

3.4. Optimum Cluster Member Identification

Based on Table 2, group A consists of 7 commodities. The commodities of concern are from the chilli group, namely red chili, cayenne pepper, and green chili. Based on Figure 3, in cluster A, the red chilli commodity has a pattern very similar to the prototype produced. Green chilli commodity is seen to have a pattern similar to prototype cluster A which in the 50th period and so on. In cluster C, the commodity looks quite similar to the prototype which is in the 40th period to the end.

| Cluster | Model | Members of cluster                      |
|---------|-------|----------------------------------------|
| A       | 7     | Red chili, Cayenne, Green chili, Young jackfruit, Petai, Potato, Cabbage. |
| B       | 1     | Avocado.                               |
| C       | 12    | Rice, Leeks, Cassava Leaves, Kale, Cauliflower, Chayote, Carrot, Mung Beans, Apples, Oranges, Pears, Pepper. |
| D       | 10    | Tree Cassava, Green Mustard, White Mustard, Cucumber, Papaya, Banana, Grapes, Salak, Coconut, Coriander. |
| E       | 13    | Spinach, Beans, Long Beans, Bean Sprouts, Vegetable Tomatoes, Fruit Tomatoes, Sweet Corn, Peanuts, Garlic, Shallots, Candlenuts, Melons, Watermelons. |
3.5. **Optimum Cluster Member Identification**

Cluster level modeling on optimum clusters is done by identifying the cluster representatives (prototypes) for each cluster. Furthermore, each prototype will be tested for stationary use using the ADF test. The test results are presented in Table 3, all stationary clusters at 5% significance level, based on p-value. ARIMA level models on the optimum number of cluster groups are presented in Table 3. Figure 4 presents a prototype plot with a cluster-level ARIMA model. It can be seen that clusters A, B, and C have prototypes with the same pattern as the ARIMA model fitting results. This means that the model built is sufficient to represent the prototype or data to be modeled. The resulting ARIMA model fitting values look underestimated. However, there are several periods that result in the overestimated value of the fitted ARIMA model, that is, when the value of inflation on the prototype decreases, the p-value of all clusters in the Ljung-Box test is greater than the actual level used which is 5%. That considerably. Next, the residuals assumptions of the diagnosed model will be examined. Based on Table means, in all clusters, the independent residuals assumption is fulfilled.

![Figure 3. Prototype plots with several objects in the same cluster](image)

![Figure 4. Prototype plots with cluster level fitting ARIMA model](image)

| Cluster | Model     | Equation                                | ADF test | Ljung box-test |
|---------|-----------|-----------------------------------------|----------|----------------|
| A       | ARIMA(0,0,1) | $Y_t = -0.42e_{t-1} + e_t$              | 0.01     | 0.98           |
| B       | ARIMA(0,0,2) | $Y_t = -0.28e_{t-1} - 0.58e_{t-2} + e_t$ | 0.01     | 0.32           |
| C       | ARIMA(1,0,0) | $Y_t = 0.003 - 0.38Y_{t-1} + e_t$       | 0.03     | 0.74           |
| D       | ARIMA(0,0,0) | $Y_t = 0.0006 + e_t$                    | 0.01     | 0.59           |
| E       | ARIMA(0,0,0) | $Y_t = 0 + e_t$                         | 0.02     | 0.11           |

* : Model has intercept >0
### 3.6. Forecasting Inflation Using The ARIMA Model at Cluster Level

Figure 5 presents the forecasting results for several clusters. Cluster A has the ARIMA model (0,0,1). Forecasting value generated only in the first period, for the next period has a forecast value of intercept (average). Cluster A is a model with an interception of zero, so for forecasting in the 2nd period (period 86) and so on will produce a forecast of zero. Furthermore, the C cluster has an ARIMA model (1,0,0). The resulting forecast value has a decreasing pattern. Significant decrease shows in the 86th period, namely February 2019. From the next period until the end, a decrease shown that occurs more sloping.

### 3.7. Evaluation of Forecasting Cluster Level Models

Cluster level modeling is to reduce the number of models. The cluster level modeling has an RMSE that does not differ greatly in value from the individual level. There is one commodity whose RMSE value is very large at the cluster level. In addition, a formal z-test was also conducted to test the differences between RMSE at the individual level and the cluster level. The resulting p-value is 0.92, the value is not significant at the 5% level of significance. Therefore, it can be said that there is no difference between individual RMSE and cluster level RMSE. This means that the cluster level modeling is quite good and representative.

![](image)

**Figure 5.** Forecasting with cluster-level ARIMA models for cluster A (a) and cluster C (b)

### 4. Conclusion

The K-Error method produces DKI Jakarta food commodity inflation having an optimum number of clusters of 5 clusters. Each clusters formed has a different ARIMA model. Cluster level modeling has forecast accuracy that is not much different from individual level models based on the resulting RMSE values.

### References

[1] Montero P, Vilar J A 2014 *J. Stat. Soft.* 62 1  
[2] Wang X, Mueen A, Ding H, Trajcevski P, Scheuermann P, Keogh E 2013 *J. Data. Min. Knowl. Disc* 26 275  
[3] Łuczak M 2016 *J. Expert Syst. Appl.* 62 116  
[4] Batista G E A A, Keogh E J, Tataw O M et al. 2014 *J. Data. Min. Knowl. Disc.* 28, 634  
[5] Ariski N, Sumertajaya I M, Aidi M N 2017 *J. Sci. Eng. Res.* 8 2244  
[6] Safitri E, Sumertajaya I M, Aidi M N 2019 *J. Sci. Res.* 9 448  
[7] Hartati 2017 *J. Math. Sci. Tech.* 18 1  
[8] Sen P, Roy M and Pal P 2016 *J. Ene.* 116 103  
[9] Bank of Indonesia 2016 *Statistik Ekonomi Keuangan Indonesia.* (Jakarta: Bank of Indonesia)  
[10] Central Bureau of Statistics DKI Jakarta 2018 *Indeks Harga Konsumen dan Inflasi DKI Jakarta* (Jakarta: Central Bureau of Statistics DKI Jakarta)  
[11] Adinugroho W, Sumertajaya I M, Syafitri U D 2017 *J. Sci. Eng. Res.* 12 1975  
[12] Utami B, Wijayanto H, Sumertajaya I M 2018 *J. Sci. Basic Appl. Res.* 39 21
[13] Paul J C, Hoque S 2013 *Glob. J. Manag. Bus. Res.* **13** 15
[14] Aghabozorghi S, Shirzhorsidi A S, Wah T Y 2015 *J. Inf. Syst.* **53** 16
[15] Shaik I, Nitella S S, Hiwarkar T, Nalla S 2019 *J. Innov. Tech. Explor. Eng.* **8** 1910
[16] Ghuman S S 2016 *J. Comp. Sci. Mob. Comp.* **5** 524
[17] Rahkmawati Y, Sumertajaya I M and Aidi M N 2019 *J. Sci. Res.* **9** 439