Evaluation of Accuracy in Identification of ARIMA Models Based on Model Selection Criteria for Inflation Forecasting with the TSClust Approach

Yeni Rahkmawati*, I Made Sumertajaya*, Muhammad Nur Aidi *

* Department of Statistics, IPB University

DOI: 10.29322/IJSRP.9.09.2019.p9355
http://dx.doi.org/10.29322/IJSRP.9.09.2019.p9355

***Abstract*-** The issue in ARIMA modeling is the use of model selection criteria to choose the best ARIMA model. The use of different model selection criteria can produce the best ARIMA models that are different so it can be difficult to choose the best model to be used. Therefore, a study was conducted on the characteristics of the model selection criteria through evaluating the accuracy in identifying the ARIMA model by using simulations on several generation data conditions with different models, parameter values and variances. Furthermore, the best model selection criteria are used in forecasting inflation of animal food and processed products commodities using the TSClust approach. Based on the simulation results, BIC is a model selection criteria that has the highest identification accuracy in the simulation. While in the application of real data, modeling using TSClust is more effective and feasible to use compared to modeling per commodity. Clustering of time series in the inflation data of animal and processed food commodities produces seven groups that have different characteristics.

Keywords: model selection criteria, ARIMA model, clustering time series, Piccolo distance

I. INTRODUCTION

Forecasting is an activity of predicting several events in the future. Forecasting is widely used in various fields, such as business, industry, government, economics, and finance, so we need statistical techniques for forecasting that involve the use of time-series data. One forecasting technique that can be used is autoregressive integrated moving average (ARIMA (p, d, q)) model developed by Box, et.al [1]. The ARIMA model (p, d, q) uses the past and present values of the variables to produce accurate short-term forecasting [2].

The issue in ARIMA modeling is the selection of the best ARIMA model. The best ARIMA models can be obtained based on the model selection criteria. Model selection criteria commonly used in the selection of ARIMA models, namely: Akaike's information criterion (AIC) revealed by Akaike [3], Akaike's information criterion bias-corrected (AICC) by Hurvich and Tsai [4], Bayesian information criterion (BIC) by Schwartz [5], mean absolute percentage error (MAPE), and root mean squared error (RMSE). Reddy SK [6], Ningtyas [7], Paul and Hoque [8] used the ARIMA model for forecasting with several model selection criteria, but each of the model selection criteria produced the best ARIMA model which was different and the model selection criteria which produced the most appropriate model was not known. Therefore, we need a study of the characteristics of the model selection criteria. One characteristic that can be observed from the selection criteria for the model is the accuracy in identifying the ARIMA model. In this study, these characteristics will be evaluated by using simulations in several generation data conditions with different models, parameter values and variations.

The ARIMA model (p, d, q) will be used in forecasting inflation of animal and processed food commodities in DKI Jakarta Province with clustering time-series (TSClust) approach, is a clustering that takes into account the dynamic nature of a time series data. Clustering time-series can be used as an alternative in handling forecasting for data that has many objects. According to Liao [9], the distance used in clustering time series is divided into three types, namely: raw data, feature data, and distance based on time series models. Distance based on time series models were first introduced by Piccolo [10] who took advantage of the similarity of structures in the ARIMA model. Each time series was made an ARIMA model based on the selection criteria of the model. The selection criteria for the model to be used are model selection criteria that have the best model ARIMA identification at the simulation stage.

Based on the description above, the purpose of this study is to evaluate the accuracy in identifying the ARIMA model based on the model selection criteria and forecasting the inflation of animal food and processed products commodities in DKI Jakarta Province with the TSClust approach using the best selection criteria for simulation results.

II. MATERIALS AND METHODS

2.1 Model Selection Criteria

Model selection criteria are criteria used to choose the best model to be used. There are several criteria that are often used in the selection of the ARIMA model, namely: AIC, AICC, BIC, RMSE and MAPE.

i. *Akaike's information Criterion (AIC)*

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2k \quad (1)$$

ii. *Akaike's Information Criterion Bias Corrected (AICC)*

$$AICc = AIC + \frac{2(k+1)(k+2)}{n-k-2} \quad (2)$$

iii. *Bayesian information Criterion (BIC)*

$$BIC = -2\ln(\mathcal{L}) + k\ln(n) \quad (3)$$

iv. *Mean Absolute Percentage Error (MAPE)*

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y_t}}{y_t}x100\right| \quad (4)$$

v. *Root Mean Squared Error (RMSE)*

$$RMSE = \left(\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y_t}|^2\right)^{\frac{1}{2}} \quad (5)$$

where: $\mathcal{L}$ = the maximum value of the likelihood function, k = the number of parameters in the model, n = the number of observations, y_t = t-time observations, and $\hat{y}_t$ == estimated values of t-time observations

2.2 Data

There are two data used in this study, namely: generation data and real data.

1. Generation data used in the simulation are time series with AR(1), AR(2), MA(1), MA(2) and ARMA(1,1) models. Data generated with five parameters and two different kinds of variances. The parameters used are the generation parameters that adequate the data stationary requirements. The parameter requirements that must be adequate so that the data generated are stationary can be seen in Table 1.

Table 1 Stasionary condition on several ARIMA model

Time Series Model	Stationary Condition
AR(1)	$\|\phi\| < 1$
AR(2)	$\phi_1 + \phi_2 < 1;\ \phi_2 - \phi_1 < 1;\ \|\phi_2\| < 1$
MA(1)	$\|\theta\| < 1$
MA(2)	$\theta_1 + \theta_2 < 1;\ \theta_2 - \theta_1 < 1;\ \|\theta_2\| < 1$
ARMA(1,1)	$\|\phi\| < 1;\ \|\theta\| < 1$

Each series has a length of observation period of 100 time points. Time series data are generated with a mean of zero and assuming $\varepsilon_{it} \sim N(0, \sigma^2)$.

2. Real data is secondary data obtained from the publication of BPS in DKI Jakarta, namely: Jakarta Consumer Price Index and Inflation. The data collected in this study are monthly inflation data from January 2012 to December 2018. The data will be divided into 2 (two), namely 78 training data (January 2012-June 2018) and 6 test data (July 2018-December 2018). The objects used in this study were 45 animal food and processed products commodities (percent).

2.3 Data Analysis Procedure

Stages of analysis in this study are generally divided into two stages, namely: simulation with generation data and application to real data. The following stages are carried out in this study:

I. Simulation Using Generation Data
   1. Generating simulation data.
   2. Modeling time series data with ARIMA models.
      a. Identify the model and estimate the parameters with the maximum likelihood of the candidate model. Model candidates are a combination of models with a maximum order of p = 5, d = 2 and q = 5 in the ARIMA model (p, d, q) so that there are 108 model candidates. Model identification is done by selecting the minimum value for the model selection criteria from various model candidates.
   3. Repeat steps 1 and 2 100 times.
   4. Evaluate the accuracy of the model selection criteria in identifying the model with the following steps:
      a. Calculate the percentage accuracy of the model selection criteria in identifying ARIMA models in each simulation condition.
      b. Calculate the average percentage of accuracy in 50 simulation conditions.
      c. Choosing a model selection criteria that has good accuracy in identifying the model is the selection criteria that has the largest average percentage accuracy.

II. Application to real data

At this stage, real data is used for forecasting models with time series data clustering methods that use the best model selection criteria from simulation results. The steps at the stage of applying real data are as follows:

1. Data Exploration
2. ARIMA modeling at the individual level uses the best model selection criteria based on simulation results.
3. Applying clustering time series using Piccolo distance and average linkage methods.
4. Calculate the data that represents each cluster by using a prototype average value.
5. ARIMA modeling of cluster level.
6. Evaluate modeling without clustering and use clustering by comparing the RMSE values.
7. Forecasting based on the best ARIMA model and interpret the forecast results.

III. RESULTS

3.1 Evaluation of Accuracy in Identification of ARIMA Models Based on Model Selection Criteria

The time series data for each condition is identified based on the model selection criteria. Then the identification model is matched with generation data model. Thus, out of 100 replications in one simulation condition, the magnitude of accuracy in identifying the ARIMA model can be calculated based on the model selection criteria. For example in Figure 1 presents a bar diagram of the percentage of models identified based on the criteria for selecting the model in the arising condition data model AR (1) with parameters ϕ_1 = -0.78831 and σ_e ^ 2 = 0.0005. Based on Figure 1, the accuracy of identification of each model selection criteria in the simulation conditions is 52% based on AIC, 92% based on BIC, 58% based on AICC, 0% based on RMSE and 0% based on MAPE.

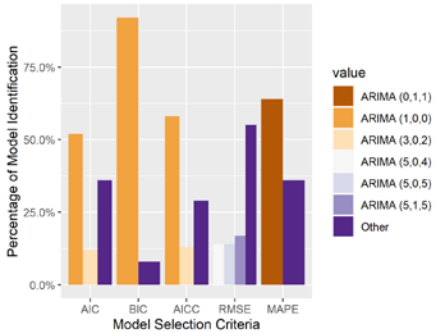


Figure 1 The results of the model identification on the simulation condition of the AR(1) with parameter $\phi_1 = -0.78831$ and $\sigma_e^2 = 0.0005$

There are 50 conditions of the simulation carried out so obtained the results of accuracy for each of these conditions. Figure 2 shows box plot of the accuracy of the identification of the ARIMA model in 50 simulation conditions. BIC has a considerable diversity compared to the others which can be seen based on Figure 2 where BIC has the largest box width. The high diversity of the BIC criteria shows that the accuracy values in 50 simulation conditions have quite varied values. Meanwhile, the AIC and AICC criteria have almost the same diversity because the width of the box in the box diagram is almost the same and the width of the box is narrower than the BIC which shows that the accuracy value is no more diverse than the BIC. Unlike the other criteria, RMSE and MAPE have very little diversity, because the accuracy value is almost the same in every 50 simulation conditions which is around the value of 0%.

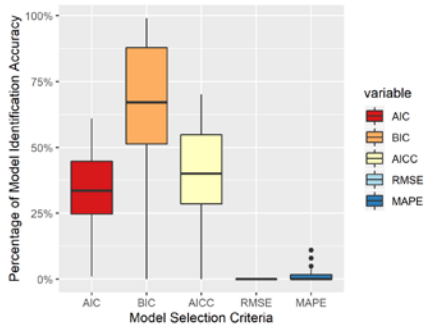


Figure 2 The results of accuracy of the model selection criteria in 50 simulation conditions

In 50 conditions of the simulation carried out, the BIC criteria had the highest average accuracy compared to other criteria with an average value of 63.36%. AIC was only able to correctly identify the model with an average accuracy of 33.66%. Likewise, AICC can correctly identify the model with an average accuracy of 39.98%. While RMSE and MAPE have a low average accuracy of identification, namely 0% and 1.44%, respectively.

Table 2 Summary of statistic on the accuracy of the model selection criteria

Statistic	AIC	BIC	AICC	RMSE	MAPE
Min.	1.00	0.00	0.00	0.00	0.00
Median	33.50	67.00	40.00	0.00	0.00
Mean	33.66	63.36	39.98	0.00	1.44
Max.	61.00	99.00	70.00	0.00	11.00
Standart Deviation	14.13	27.98	16.52	0.00	2.60

So, based on the evaluation results above, BIC becomes the criteria for selecting a model that has the best accuracy in identifying the model among the other criteria used in this simulation.

3.2 Modeling of Animal food and Processed Products Commodities Inflation with Clustering Time Series

3.2.1 Data Exploration

Plots of data were carried out to visually see patterns of inflation in animal food and processed products commodities in DKI Jakarta Province. The following is a plot of inflation data for several commodities presented in Figure 3. Inflation plot of broiler chicken meat has a certain pattern in the same period intervals with considerable fluctuations, but the pattern between periods is almost the same until the 60th period (December 2016), after December 2016 inflation in broiler chicken commodities did not experience large fluctuations until June 2018. Almost similar to chicken meat commodities, the plots of pomfret fish commodity inflation experienced similar things but large fluctuations occurred in almost every period until the 78th period (June 2018). If noted, the two plots have almost the same plot, although between the two commodities have a different range of inflation values. Furthermore, the catfish commodity inflation plot has a pattern of not too fluctuating but there is a surprise value at the 25th period (January 2014) to 29th (May 2014). Likewise, the inflation plot of sweetened condensed milk commodity has a similar pattern to catfish commodities, but the value shock occurred at the 24th period (December 2013) to the 27th (March 2014). A similar pattern can be an indication that the two commodities can be in the same group at the modeling stage with the groups which will be discussed in the next section.

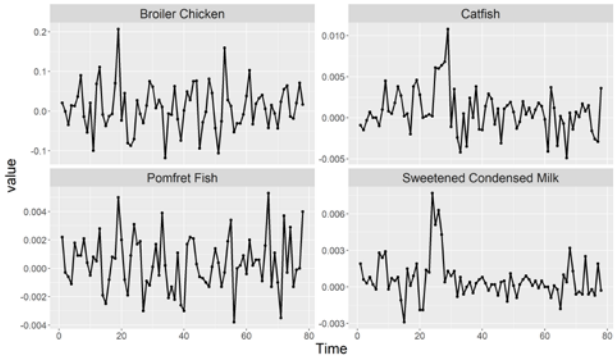


Figure 3 Plot inflation of broiler chicken commodity (top left), catfish commodity (top right), pomfret commodity (bottom left) and sweetened condensed milk commodity (bottom right)

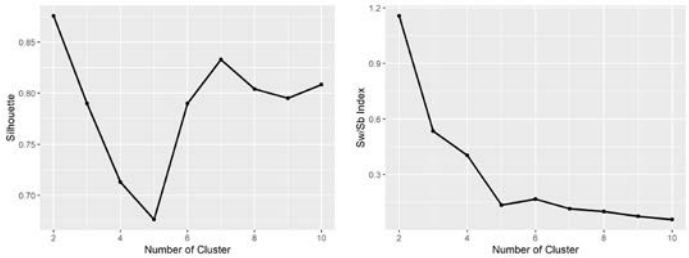
3.2.2 Inflation Modeling of Individual Level Data

The next step is modeling inflation at the individual level. Individual level modeling is modeling done on 45 animal food and processed product commodities. The model used is the ARIMA model by using the BIC model selection criteria. The ARIMA model that is identified in all commodities is not so varied, given

the results of the time series plot of inflation data show a similar pattern between commodities. For example, the commodities of catfish (Y18), tilapia fish (Y20), goldfish (Y27), and sweetened condensed milk (Y36) have the AR(1) model. Then, there are four commodities identified by the MA(1) model, namely: chicken meat commodity (Y3), carp (Y19), selar fish (Y21), and mackerel fish (Y30). Animal food and processed product commodities inflation data has a lot of objects so that it will become inefficient if modeling all commodities so that it is more efficient to do modeling with the TSClust approach.

3.2.3 Clustering Time Series

Animal food and processed product commodities inflation data can be grouped into 2 to 44 clusters. So, the next step in order to obtain the best cluster is to determine the optimal number of clusters that can be obtained based on the maximum value of the silhouette coefficient and the minimum value of the diversity ratio within the group and between groups in the number of groups 2 to 10 presented in the plot in Figure 4.



(a) (b)

Figure 4 Silhouette coefficient (a) The diversity ratio within the group and between groups (b)

Based on the silhouette coefficient and the ratio of diversity in clusters and between clusters in Figure 4 shows that many optimal clusters chosen are k = 7. Thus, inflation data for animal and processed food commodities will be clustered into 7 groups named as A, B, C, D, E, F and G, which are presented in Table 3.

Table 3 The results of the cluster of animal and processed food commodities inflation

Cluster	Members
A	Live chicken, meatballs, broiler chicken, pork, canned meat, beef, chicken liver, beef sausage, chicken nuggets, milkfish, pomfret, squid, gourami fish, red snapper fish, long jawed mackerel, Spanish mackerel fish, tuna fish, shrimp, milkfish presto, salted squid, canned fish, basket fish, anchovy, three spot gourami, cheese, baby food, milk for pregnant women, free-range chicken eggs, quail eggs, milk for elderly bones, packaged liquid milk, low-fat milk
B	Free-range chicken meat, catfish, goldfish, mackerel, snakehead fish, salted long jawed mackerel, sweetened condensed milk
C	Lamb
D	Tilapia Fish
E	Milk powder
F	Milk for toddlers, milk for babies
G	Broiler chicken eggs.

After identifying the membership of each cluster, it will proceed with the calculation of the prototype of each cluster. The prototype time series plot and several commodities per cluster are shown in Figure 5. The time series patterns between prototypes have different characteristics. This is because commodities that have a pattern of inflation are almost similar to being in one cluster and differing between groups, so the prototype produced by each group has different characteristics.

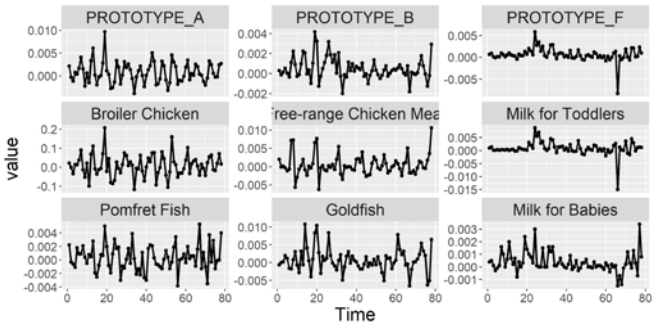


Figure 5 Plots of several commodities and prototypes per cluster

The prototype of each cluster is also called the cluster level data and the ARIMA modeling process for this data is called the ARIMA modeling of cluster level. The next stage in ARIMA modeling at the cluster level will be described in the next section.

3.2.4 Modeling of Cluster Level

Inflation modeling of cluster level is modeling that using prototype data on each cluster. The model used at this stage is the ARIMA model so that each prototype in each cluster is carried out iterative stages of ARIMA model. The ARIMA model and the formal test of cluster level are shown in Table 4. The ARIMA model of the resulting cluster level is different except for cluster A and cluster B which both have ARIMA models (0,0,1)

Table 4 ARIMA models and cluster level formal tests

Cluster	Model	p-value	
		KPSS test	Ljung-Box test
A	ARIMA(0,0,1)	0.1000	0.7572
B	ARIMA(0,0,1)	0.1000	0.6452
C	ARIMA(0,1,1)	0.0422*	0.6555
D	ARIMA(1,0,0)	0.0756	0.5908
E	ARIMA(0,1,2)	0.0119*	0.9963
F	ARIMA(1,0,1)	0.0832	0.7476
G	ARIMA(2,0,2)	0.1000	0.6041

*Significant on $\alpha = 0.05$

3.2.5 Evaluation of the Individual Model and the Cluster Model

Based on Table 5, the RMSE value without using clustering method in prediction evaluation and forecast evaluation has a smaller value than the method using the clustering, but the difference of the RMSE value between the two methods is not too large so it needs to be analyzed by mean difference test on 45 commodities. Based on the results of the mean difference test, the RMSE value between modeling without clustering and using clustering was not significantly different. Therefore, the method using clustering is very feasible to use considering this method is more efficient and the results obtained are not much different from the method without clustering.

Table 5 The average RMSE value of the results of prediction evaluations and forecasts between modeling without clustering and using clustering

Modeling Method	RMSE	
	Prediction	Forecast
Without clustering	0.003747815	0.005025309
Using clustering	0.003985997	0.005069382

3.2.6 Forecasting and Interpretation

The application of ARIMA modeling with time series clustering is to forecast each cluster with a length of forecasting time of 9 months (January 2019-September 2019).

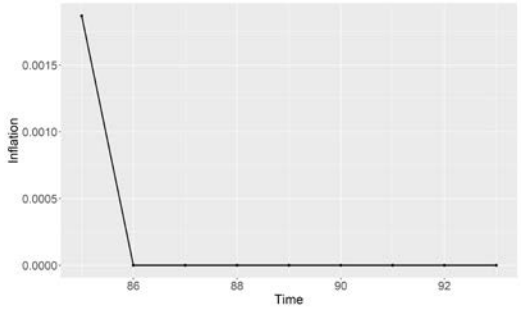


Figure 6 Result of inflation forecast Cluster A

Figure 6 presents the forecasting results for several clusters. Cluster A has an ARIMA model (0,0,1) which has a forecast value in the first forecasting period of 0.0018690%, while for the next period it has a forecast value of zero. All commodities in cluster A have forecast results in accordance with forecast results on the cluster.

IV. CONCLUSION

BIC model selection criteria become the best model selection criteria in the simulation, because it has the highest average accuracy of ARIMA model identification. The BIC criterion has a greater penalty than the others so that it is better at fulfilling the parsimony nature in identifying the model.

Monthly inflation rates for animal food and processed products commodities have varied volatility. Based on evaluation, Model of inflation forecasting using clustering is more effective and feasible. Animal and processed food commodities are grouped into seven groups which have different patterns.

ACKNOWLEDGEMENTS

This work is fully supported by Kemenristek DIKTI (Kementerian Riset Teknologi dan Pendidikan Tinggi) of Indonesia.

REFERENCES

[1] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Third Edit. New Jersey: Prentice-Hall, Inc., 1994.

[2] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, Third Edit. New Jersey: John Wiley & Sons. Inc ., 2008.

[3] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” 1973.

[4] C. M. Hurvich and C.-L. Tsai, “Regression and Time Series Model Selection in Small Samples,” *Biometrika*, vol. 76, no. 2, p. 297, Jun. 1989.

[5] G. Schwartz, “Estimating the Dimension of a Model,” *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[6] B. A. Reddy SK, “Exchange Rate Forecasting using ARIMA, Neural Network and Fuzzy Neuron,” *J. Stock Forex Trading*, vol. 04, no. 03, 2015.

[7] D. I. Ningtyas, “Peramalan Curah Hujan dengan Model ARIMA (Autoregressive Integrated Moving Average) dan VECM (Vector Error Correction Model),” Institut Pertanian Bogor, 2016.

[8] J. C. Paul and S. Hoque, “Selection of best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square Pharmaceutical Ltd.,” *Glob. J. Manag. Bus. Res.*, vol. 13, no. 3, pp. 15–26, 2013.

[9] W. T. Liao, “Clustering of time series data-a survey,” *Pattern Recognit.*, 2005.

[10] D. Piccolo, “A Distance Measure For Classifying ARIMA Models,” vol. 11, no. 2, pp. 153–164, 1990.

.

AUTHORS

First Author – Yeni Rahkmawati, S.Mat, college student, Department of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and yenirahkmawati@gmail.com

Second Author – Dr. Ir. I Made Sumertajaya, M.S., Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and imsjaya.stk@gmail.com

Third Author – Dr. Ir. Muhammad Nur Aidi, M.S., Lecturer, Department of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and Nuraidi@yahoo.com

Correspondence Author - Dr. Ir. I Made Sumertajaya, M.S., Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and imsjaya.stk@gmail.com