International Journal of Scientific & Engineering Research Volume 8, Issue 12, December-2017
ISSN 2229-5518

Development of Forecasting Model of Cooking Oil Price with Time series Clustering Approach

Widyayanto Adinugroho, Utami Dyah Syafitri, I Made Sumertajaya

Abstract— Cooking oil is one of commodities that contributed to rural inflation in Indonesia. Time series data can be used to model for forecasting. Time series data can have pattern resemblance. This resemblance can be used to form clusters. This research aimed to compare the accuracies of ARIMA model of cooking oil price developed using hierarchical agglomerative clustering (HAC) time series method with ARIMA of individual level. As the results, modeling for 32 provinces individually resulting an average MAPE of 2.56%, while ARIMA model development with DTW as the distance measures for HAC and employing 2 or 3 clusters, gave a more efficient on modeling process and only giving effect on increased MAPE just by average of 6 to 7%. They were compared with models without clustering. Those clusters with DTW gave a lower MAPE in 11 provinces. The DTW cophenetic coefficient was 0.772.

Index Terms— *ARIMA, CPO, Clustering, Hierarchical, Inflation*

1 Introduction

High or unstable inflation rate affecting the welfare achievement. Since 2005, the task of stabilizing and controlling the inflation rate is in Inflation Control Team (TPI, Tim Pengendalian Inflasi). TPI works as a central government institution. In 2008, the controlling coordination is extended to the regional scope; the Regional Inflation Control Team (TPID, Tim Pengendalian Inflasi Daerah) was established in Regencies / Municipals / Provinces level [1].

In 2015, 23 of the 34 provinces have the percentage of rural population of more than 50% [2]. Therefore, the prices control by the TPID is important in order to control rural inflation. In running its responsibility, TPID coordinates with regional representatives of the BPS- Statistics of Indonesia (BPS, Badan Pusat Statistik).

Cooking oil is one of the important commodities to be monitored its inflation levels because it belongs to the strategic commodity group, that is, the most widely consumed commodities. Cooking oil also belongs to the group of volatile foods [3]. Based on the results of the National Socio-Economic Survey (Susenas) in September 2015, the average per capita consumption of cooking oil in rural area was as much as 0.220 liters per week, while in 2016 rose to 0.226 liters per week (11.75 liters per year).

Price control policy formulation, data and short-term market projection, such as the future consumption and production estimates and short-term prices prediction are very necessary [4]. Statistics Office provides data prices with monthly period, e.g. rural consumer price, which includes cooking oil [5]. The fulfillment of price projection is done through forecasting.

Previous researches on price forecasting in Indonesia have been done. Susila and Munadi [6] used ARIMA for modeling the price of cooking oil. Sumaryanto [7] apply modeling with ARIMA and ARCH of basic food price, which including cooking oil at aggregation level. Ati [8] performed basic food price modeling in the Central Java Province using ARIMA and GARCH. [9] done modeling with ARIMA on palm production of PT. Nusa Indah Borneo Plantations. Whereas Munandar [10] done ARIMA modeling of tin price at futures market. Related to the issue of aggregation, it is advisable to use the data with a more specific coverage (e.g. provinces) [7]. In addition, modeling the price at the level of the province means of providing more specific information, so the price control policy formulation becomes more precise.

Data collected in time sequence called time series data. Time series data collection can be done in several locations (multi locations). Time series data of multi locations are very likely to have a resemblance in patterns. Previous researches of clustering time series has been done. Ardiansyah [11] done clustering of third-party funds of 30 provinces by using autocorrelation distance, Complexity-Invariant Dissimilarity (CID) distance, periodogram distance and Dynamic Time Warping (DTW). Whereas Veti [12], done clustering the value of Indonesia's exports to 20 countries of destination using autocorrelation distance, CID and DTW distance. Both, after the clustering process, didn't do modeling.

This paper aimed to build ARIMA models for each location (individual level) which aimed to provide specific information, developing ARIMA individual level to be ARIMA of cluster level using clustering timeseries and to compare the accuracy of ARIMA individual level with its development of ARIMA with a clustering time series process.

2 Literature Riviews

2.1 Autoregressive Integrated Moving Average (ARIMA)

Yt is a mixed model of autoregressive integrated moving average if at d differencing $W_t = \nabla^d Y_t$ is a stationary ARMA process. If Wt follows an ARMA(p,q) then Yt is an ARIMA(p,d,q) process [13]. And, when d=1 then :

$$W_t=\phi_1 W_{t-1}+ \phi_2 W_{t-2}+\ldots+ \phi_p W_{t-p}+e_t- \theta_1 e_{t-1}- \theta_2 e_{t-2}-\ldots- \theta_q e_{t-q}$$

with $W_t = Y_t - Y_{t-1}$, and when d=2

$$Z_t=\phi_1 Z_{t-1}+ \phi_2 Z_{t-2}+\ldots+ \phi_p Z_{t-p}+e_t- \theta_1 e_{t-1}- \theta_2 e_{t-2}-\ldots- \theta_q e_{t-q}$$

with $Z_t = W_t - W_{t-1}$. Both with order of p and q in it d order stationary.

2.2 Dynamic Time Warping (DTW) distance

DTW is dissimilarity distance that measures minimum distance of two set of points with a mapping process. DTW distance is defined as follows [14]:

$$d_{DTW}(X_T, Y_T) = \min_{r \in M} \left(\sum_{i=1,\ldots,m} | X_{ai} - Y_{bi} | \right)$$

International Journal of Scientific & Engineering Research, Volume 8, Issue 12, December -2017
ISSN 2229-5518

2.3 Hierarcycal Agglomerative Clustering (HAC) with Average Linkage

Method of clustering which at first clustering n object in n groups, and grouping a similar objects into same cluster, until all object become just one group. One of the linkages can be used is group average linkage as defined by Sokal and Michener in [15] is:

$$d(R,Q) = \frac{1}{|R|.|Q|}\sum_{\substack{i\in R \\ j\in Q}} d(i,j)$$

2.4 Cluster Representatives

Cluster representative is defined as a sequence of points in time series data. A method to obtain this representative is with calculating average of each point in timeseries [16].

3 Methods

3.1 Data

Data used in this paper available at BPS-Statistics of Indonesia website in published materials compiled as publication "Rural Consumer Price Statistics: Group of Foods", from 2009 to 2016. The commodity being focused is cooking oil price in 32 provinces in Indonesia. The data in these publications obtained from monthly surveys conducted by the office in all provinces on which its samples were selected purposively. As the testing dataset are the 12 data in 2016 of each province.

3.2 Method of Analysis

The steps used in analyzing the data are:

1. Data explorations
2. Modeling ARIMA in individual level (each province)
   Building ARIMA model using Box-Jenkins iterative procedures as follows:
   a. Model identification
      The stage of identifying the order of ARIMA in stationary time series data by carefully checks the pattern in autocorrelation plot and partial autocorrelation plot. A few tentative models can be identified in this stage.
   b. Parameter estimation
      Parameter estimation used is maximum-likelihood estimation. This stage is conducted for all tentative models.
   c. Diagnostic checking
      Checks for independence in residual with Ljung-Box test. The hypothesis is:
      H_0 : the residual is independent
      H_1 : the residual isn't independent
      The test statistic used is as follows:
      $$Q^* = n(n+2)\left(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-2} + \ldots + \frac{\hat{r}_K^2}{n-K}\right)$$
      Model with independence in residual, significant in parameters and most parsimonious will be used for forecasting.
3. Modeling ARIMA in cluster level
   The stages consist of:
   a. Calculating dissimilarity matrix with DTW distance
   b. Clustering the provinces in groups with hierarchical agglomerative clustering method with average linkage
   c. Calculating silhouette coefficients (SC) for each number of clusters (k) from $2 \leq k \leq 10$
   d. Choosing k with SC with SC ≥ 0.50 as in [15]
   e. Identifying cluster members
   f. Calculating cluster representatives with average sequence of the set method
   g. Using cluster representatives as inputs for step 2.a – 2.c to build ARIMA for each cluster representative.
4. Measure MAPE from ARIMA individual models (stage 2) and ARIMA cluster (stage 3). MAPE for developed ARIMA are calculated with comparing the forecast with actual data as mentioned in section 3.1
5. Compare MAPE.
6. Finished.

- *Widyayanto Adinugroho is currently pursuing master degree in Applied Statistics in Bogor Agricultural University, Indonesia. E-mail: widyayanto.a@gmail.com*
- *Utami Dyah Syafitri is lecturer at Department of Statistics, Bogor Agricultural University, Indonesia. E-mail: utamids@gmail.com*
- *I Made Sumertajaya is lecturer at Department of Statistics , Bogor Agricultural University, Indonesia. E-mail: imsjaya@gmail.com*

4 Result and Discussion

4.1 Data Explorations

Cooking oil price in the western region of Indonesia showed a much lower price than the eastern region. Most of the time, the price in Nusa Tenggara, Maluku, North Maluku, West Papua and Papua always become the most expensive among 32 provinces. Cooking oil price in 32 provinces is mainly to have trend over time as showed in Fig 1.

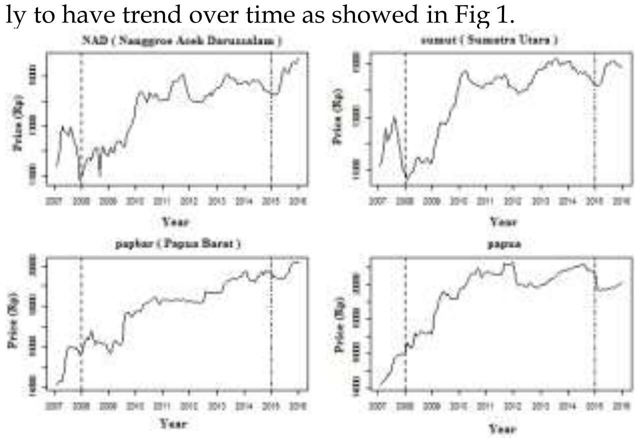

Fig 1 Time series plot cooking oil price in 4 provinces

International Journal of Scientific & Engineering Research, Volume 8, Issue 12, December -2017
ISSN 2229-5518

4.2 Clustering Timeseries

Using the hierarchical agglomerative clustering method with average linkage, the number of cluster can be formed is stretched from 32 to 1. 32 is when all the provinces is as a cluster of its own. While 1 is when all 32 provinces merge as one group.

The silhouette coefficient in forming the group is as illustrated in Fig 1. Fig 2 shows the number of cluster which has the strong anda make sense classification. The number of cluster was 2 and 3 respectively. The silhouette coefficient for strong classification was 0.755, while for the make sense classification was 0.543.

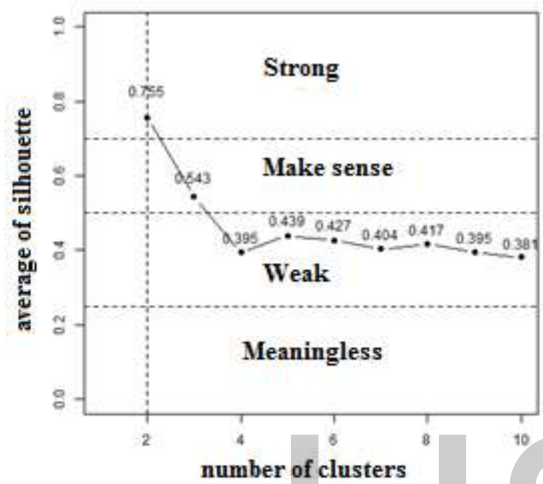

Fig 1 Silhouette coefficients plot with various number of clusters (*k*)

The dendogram shows the step by step process of forming cluster from all object are separated, and in the end become one group. The members of 2nd group when the number of cluster is 2 are East Nusa Tenggara, East Kalimantan, North Maluku, West Papua dan Papua. When the number of cluster is 3 then North Maluku, West Papua and Papua belonged to the 3rd group. For more consice explanation, a dendogram is showed in Fig 2

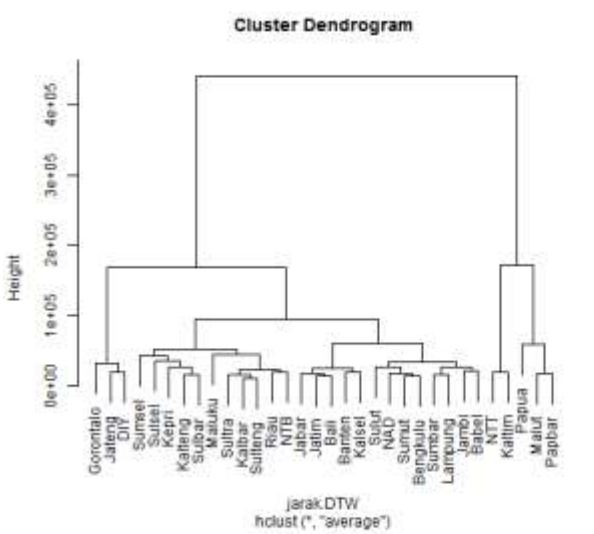

Fig 2 Dendogram of DTW

4.3 Cluster Representatives

Based on the cluster formed, cluster (or gerombol) representatives can be calculated. When the number of cluster is 2, then the name of each representative namely cluster representative A dan B. And for the number of cluster is 3, then the name of each representative are cluster representative C, D and E. The membership of each cluster (A and B, C, D and E) can be derived from the dendogram in Fig 3. The summary of cluster membership is available in Table 2.

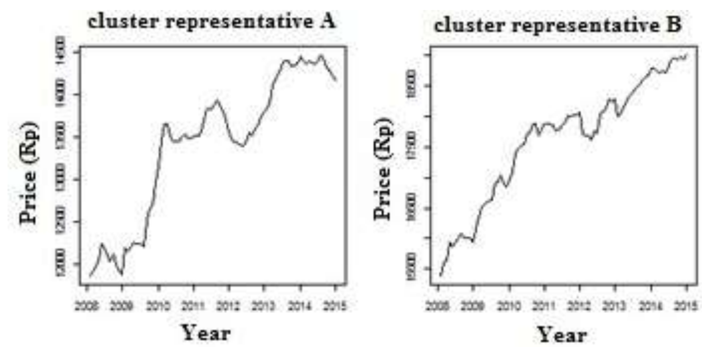

Fig 3 Cluster representatives A and B

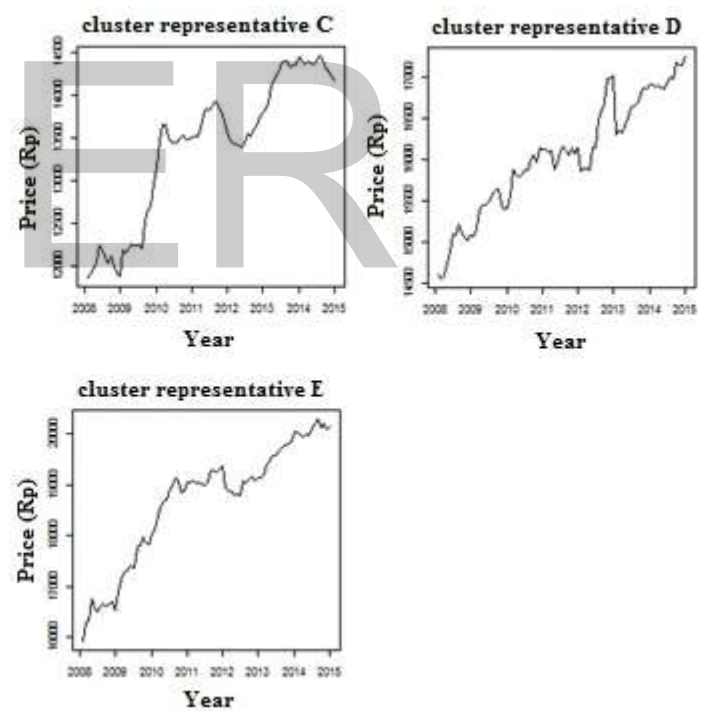

Fig 4 Cluster representatives C, D and E

As for illustration, the calculation of cluster representative B was made by calculating the average in each t from 1 to 84 of North Maluku, West Papua and Papua provinces. These values then to be sequenced by using its *t* order. The calculation for the remainder cluster representatives follows the same procedure. Time series plot for these cluster representatives showed in Fig 3 and Fig 4.

International Journal of Scientific & Engineering Research, Volume 8, Issue 12, December -2017
ISSN 2229-5518

TABLE 1
Membership of each cluster representatives

Province	DTW distance		Province	DTW distance	
	2 cl	3 cl		2 cl	3 cl
NAD	A	C	NTB	A	C
Sumut	A	C	NTT	B	D
Sumbar	A	C	Kalbar	A	C
Riau	A	C	Kalteng	A	C
Jambi	A	C	Kalsel	A	C
Sumsel	A	C	Kaltim	B	D
Bengkulu	A	C	Sulut	A	C
Lampung	A	C	Sulteng	A	C
Babel	A	C	Sulsel	A	C
Kepri	A	C	Sultra	A	C
Jabar	A	C	Gorontalo	A	C
Jateng	A	C	Sulbar	A	C
DIY	A	C	Maluku	A	C
Jatim	A	C	Malut	B	E
Banten	A	C	Papbar	B	E
Bali	A	C	Papua	B	E

cl : clusters

4.4 Best ARIMA Models

The Box-Jenkins procedures being used in modeling ARIMA in 32 provinces and 5 cluster representatives resulting the use of orders p, d and q as showed in Table 2 and Table 3.

The models in Table 2 are used to forecast 12 points ahead individually for each province. While the models in Table 3, used to forecast as many as 12 points too, but these points used by every province of its member

TABLE 2
Best ARIMA Models for 32 Provinces

No	Province	ARIMA			No	Province	ARIMA		
		p	d	q			p	d	q
1	NAD	0	1	0	17	NTB	0	1	4
2	Sumut	1	1	0	18	NTT	0	1	0
3	Sumbar	0	1	0	19	Kalbar	0	1	1
4	Riau	0	1	0	20	Kalteng	0	1	3
5	Jambi	0	1	1	21	Kalsel	2	1	0
6	Sumsel	0	1	0	22	Kaltim	2	1	1
7	Bengkulu	0	1	0	23	Sulut	1	1	2
8	Lampung	0	1	0	24	Sulteng	0	1	0
9	Babel	1	1	0	25	Sulsel	0	1	0
10	Kepri	0	1	4	26	Sultra	1	1	0
11	Jabar	0	1	1	27	Gorontalo	0	1	6
12	Jateng	1	1	0	28	Sulbar	0	1	0
13	DIY	1	1	1	29	Maluku	0	1	0
14	Jatim	1	1	0	30	Malut	0	1	0
15	Banten	3	1	0	31	Papbar	0	1	0
16	Bali	0	1	0	32	Papua	0	1	0

TABLE 3
Best ARIMA Models for 5 Cluster Representatives

No	Cluster representative	ARIMA			Note
		p	d	q	Distance (Number)
	Using DTW distance				
1	A	0	1	2	DTW.2 (1)
2	B	1	1	0	DTW.2 (2)
3	C	0	1	2	DTW.3 (1)
4	D	0	1	0	DTW.3 (2)
5	E	1	1	0	DTW.3 (3)

The forecast values used along with their respective actual data (testing datasets) to calculate a statistic called Mean Absolute Percentage Error.

4.5 Accuracy comparison

As stated earlier, in measuring the accuracy of forecasting for both methods, MAPE were used. The general formula is as in [17].

Fig 5 shows the substract between MAPE from cluster using DTW and MAPE from model without clustering. A bar above the zero axis means that the particular provinces gain advantages from the use of clustering before ARIMA modeling.

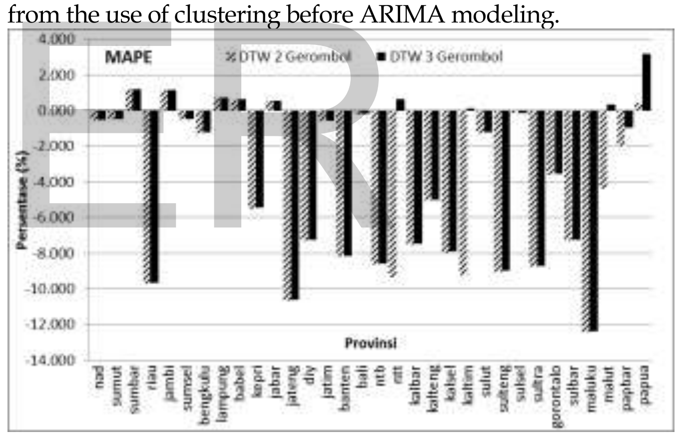

Fig 5 Substract of MAPE clustered with DTW distance

Figure 5 also shows a few provinces gained this extra accuracy from the introduction of clustering before ARIMA modeling. With 2 clusters, 6 provinces gained lower MAPE. These provinces are West Sumatra, Jambi, Lampung, Bangka Belitung Islands, West Java and Papua. Most of the province consisted of 26 provinces suffered of lacking accuracy from the use of 2 clusters clustering technique. The lost of accuracy are in range of 1% - 16%.

As the number of cluster increased from 2 to 3, 23 provinces failed getting lower MAPE. Although, as many as 9 provinces succed on improving their accuracy. They are West Sumatra, Jambi, Lampung, Bangka Belitung Islands, West Java, Papua, East Nusa Tenggara, East Kalimantan and North Maluku.

Table 4 shows the details of MAPE of each 32 provinces resulting from ARIMA without using clustering and MAPE from ARIMA with clustering time series using DTW distance.

International Journal of Scientific & Engineering Research, Volume 8, Issue 12, December -2017
ISSN 2229-5518

TABLE 4

MAPE Comparison of ARIMA individual level and ARIMA cluster level by Provinces

No	Province	Training data	MAPE in testing dataset Without clustering	DTW 2 clusters	DTW 3 clusters
1	Nanggroe Aceh Darussalam	1.245	4.862	5.869	5.869
2	North Sumatra	0.925	3.782	4.265	4.265
3	West Sumatra	0.886	3.684	2.480	2.480
4	Riau	0.944	2.390	12.047	12.047
5	Jambi	1.018	3.304	2.150	2.150
6	South Sumatra	0.786	3.269	3.762	3.762
7	Bengkulu	1.051	2.965	4.205	4.205
8	Lampung	0.872	2.618	1.899	1.899
9	Bangka Belitung Islands	0.849	3.155	2.549	2.549
10	Riau Islands	0.724	1.023	6.377	6.377
11	West Java	0.725	3.071	2.536	2.536
12	Central Java	1.011	5.298	15.924	15.924
13	D. I. Yogyakarta	1.081	6.019	13.260	13.260
14	East Java	0.865	3.799	4.364	4.364
15	Banten	0.868	1.615	9.787	9.787
16	Bali	0.896	2.893	3.098	3.098
17	West Nusa Tenggara	0.871	2.191	11.356	11.356
18	East Nusa Tenggara	0.646	1.451	10.999	1.070
19	West Kalimantan	0.561	1.866	9.348	9.348
20	Central Kalimantan	0.634	1.569	6.595	6.595
21	South Kalimantan	0.478	1.367	9.289	9.289
22	East Kalimantan	0.803	1.185	10.414	0.768
23	North Sulawesi	0.934	0.931	2.216	2.216
24	Central Sulawesi	0.778	1.804	10.797	10.797
25	South Sulawesi	0.783	0.784	0.920	0.920
26	South-East Sulawesi	0.603	0.703	9.419	9.419
27	Gorontalo	1.664	4.127	7.695	7.695
28	West Sulawesi	0.743	1.230	8.486	8.486
29	Maluku	0.633	0.504	12.871	12.871
30	North Maluku	0.697	1.992	6.155	1.651
31	West Papua	0.763	1.585	3.353	2.539
32	Papua	0.772	4.818	4.181	1.660
	mean	0.847	2.558	6.833	5.977
	std dev	0.220	1.449	4.045	4.311
	minimum	0.478	0.504	0.920	0.768
	Q1	0.725	1.430	3.289	2.414
	median	0.825	2.291	6.486	4.315
	Q3	0.903	3.278	9.943	9.365
	maximum	1.664	6.019	15.924	15.924

4 Conclusion

ARIMA individual level, modeling of 32 provinces still give the best accuracy rather than ARIMA modeling with clustering (both with clusters of 2 and 3). The introduction of clustering in time series data and continued with modeling gave practicality in building forecasts for lots of provinces.

Appendix

The figures and tables used abbreviations referring to provinces name in Indonesia. The abbreviations used are as follow:

Abbreviation	Name of Province
NAD	Nanggroe Aceh Darussalam
Sumut	North Sumatra
Sumbar	West Sumatra
Riau	Riau
Jambi	Jambi
Sumsel	South Sumatra
Bengkulu	Bengkulu
Lampung	Lampung
Babel	Bangka Belitung Islands
Kepri	Riau Islands
Jabar	West Java
Jateng	Central Java
DIY	D. I. Yogyakarta
Jatim	East Java
Banten	Banten
Bali	Bali
NTB	West Nusa Tenggara
NTT	East Nusa Tenggara
Kalbar	West Kalimantan
Kalteng	Central Kalimantan
Kalsel	South Kalimantan
Kaltim	East Kalimantan
Sulut	North Sulawesi
Sulteng	Central Sulawesi
Sulsel	South Sulawesi
Sultra	South-East Sulawesi
Gorontalo	Gorontalo
Sulbar	West Sulawesi
Maluku	Maluku
Malut	North Maluku
Papbar	West Papua
Papua	Papua

Acknowledgment

The first authors wish to thank BPS-Statistics of Indonesia for allowing author to be part of 2nd Batch of BPS-APBN Master Scholarship Program.

International Journal of Scientific & Engineering Research, Volume 8, Issue 12, December -2017
ISSN 2229-5518

REFERENCES

[1] [Bappenas] Badan Perencanaan Pembangunan Nasional. Rancangan Awal Rencana Pembangunan Menengah Nasional 2015-2019, Buku II Agenda Pembangunan Bidang. Jakarta (ID) : Bappenas. 2014.

[2] [BPS] Badan Pusat Statistik. Proyeksi Penduduk Indonesia 2010-2035. Jakarta (ID) : Badan Pusat Statistik. 2013.

[3] [BPS] Badan Pusat Statistik. Laporan Bulanan Data Sosial Ekonomi, Desember 2016. Jakarta (ID) : Badan Pusat Statistik. 2016.

[4] [Kemendag] Kementrian Perdagangan. Kajian Pengembangan Model Proyeksi Jangka Pendek Pasar Komoditas Pangan Pokok. Jakarta (ID) : Kementrian Perdagangan. 2009.

[5] [BPS] Badan Pusat Statistik. Pedoman Pencacahan Survei Harga Konsumen Perdesaan . Jakarta (ID) : Badan Pusat Statistik. 2017.

[6] Susila WR, Munadi E. Peramalan Harga Eceran Minyak Goreng dengan Model ARIMA. Jurnal Ekonomi. 3(7):40-53. 2008.

[7] Sumaryanto. Analisis Volatilitas Harga Eceran Beberapa Komoditas Pangan Utama dengan Model Arch/GARCH.JAE.27(2):135-163. 2009.

[8] Ati ENL. Analisis Volatility Forecasting 9 Bahan Pokok menggunakan Metode GARCH dengan Program R [skripsi]. Semarang (ID) : Universitas Negeri Semarang. 2015.

[9] Elvani SP, Utary AR, Yudaruddin. Peramalan Jumlah Produksi Tanaman Kelapa Sawit dengan Menggunakan Metode ARIMA. Jurnal Manajemen. 8(1):95-112. 2016.

[10] Munandar AI. Prilaku Harga Komoditas Timah di Pasar Bursa Komoditas. Jurnal Ilmiah Manajemen dan Bisnis. 2(1): 129-137. 2016.

[11] Ardiansyah M. Penggunaan Jarak Dynamic Time Warping (DTW) pada Analisis cluster Data Deret Waktu (Studi Kasus pada Dana Pihak Ketiga Provinsi se-Indonesia). Jurnal Mahasiswa Statistik Universitas Brawijaya. 2(4):277-280. 2014.

[12] Veti AP. Kajian Penggerombolan Data Deret Waktu dengan Menggunakan Beberapa Ukuran Jarak [skripsi].Bogor (ID): Institut Pertanian Bogor. 2015.

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

[14] Montero P, Vilar JA. Tsclust : An R Package for Time Series Clustering. Journal of Statistical Software, November 2014. 62(1):1-43. 2014.

[15] Kaufman L, Rousseeuw PJ. Finding Groups in Data : An Introduction to Cluster Analysis. New Jersey (US) : John Wiley and Sons Inc. 1990.

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

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