Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Implementation of Fuzzy C-Means Algorithm for Clustering Provinces in Indonesia Based on Micro and Small Industry Ratio in Village Areas

Frandito Rahmanesta[1], Zamahsary Martha[2‡], Dodi Vionanda[3], and Zilrahmi[4]

[1,2,3,4] Department of Statistics, Universitas Negeri Padang, Indonesia
‡corresponding author: zamahsarymartha@fmipa.unp.ac.id

Abstract

Post-economic crisis, the micro and small industries contribute the most labor compared to other industries. Regional development sourced from small micro industries is a strategic force in developing a country because the development of small micro industries leads to realizing equitable welfare to reduce income inequality. Development in village areas is an important factor for regional development, reducing inequality between regions, and alleviating poverty. However, based on the 2018 Potensi Desa survey, there are regional imbalances in Indonesia in the micro and small industries which is centralized on Java Island. Clustering were carried out using the Fuzzy C-Means algorithm to cluster 34 provinces in Indonesia based on the ratio of small micro industries in village areas in 2021, to see how the development of micro and small industries in village areas in each province in Indonesia. Fuzzy C-Means is one of the data clustering techniques that uses a fuzzy clustering model, where cluster formation is based on a membership degree value that varies between 0 and 1. The Fuzzy C-Means algorithm generates 4 clusters, cluster 1 and 2 represents provinces with high and very high micro and small industry development in village areas and cluster 3 and 4 represents provinces with medium and low category. Policies and strategies are needed by the Provincial government to develop small micro industries in village areas, especially in provinces that are categorized in the medium and low category. The Fuzzy C-Means algorithm produces a good cluster structure with a silhouette coefficient value of 0,6406

Keywords:Cluster Analysis, Fuzzy C-Means, Micro and Small Industry.

[*] Received: Nov 2024; Reviewed: Nov 2024; Published: Dec 2024

1. Introduction

The manufacturing industry is an industry that processes raw materials into finished or semi-finished goods by physical, chemical, or mechanical processes. The manufacturing industry produces products with high added value so the manufacturing industry contributes significantly to the Gross Domestic Product (GDP) in Indonesia and is the foundation of the Indonesian economy. The processing industry is classified into two based on the number of workers, namely Large and Medium Industries and Micro and Small Industries. Large and Medium Industries employ 20-100 workers, while Micro and Small Industries employ 1-19 workers. The National Industrial Development Master Plan (RIPIN) 2015-2035 states that micro and small industries are one of the contributors to the national economy. Micro and small industries are more adaptable to economic crises than large industries. In the post-economic crisis period, the micro and small industries contributed the most workers compared to other industries. As an industry that is relatively easy to establish, micro and small industries have an advantage in labor employment. The development of the micro and small industry is expected to make a long-term contribution that is stable and sustainable so that it has a good impact on the national economy (BPS Indonesia, 2020).

Regional development sourced from micro and small industries is a strategic force in the development of a country because micro and small industries lead to realizing equitable welfare to reduce income inequality. Development in village areas is an important factor for regional development, reducing inequality between regions, and alleviating poverty (Soleh, 2017). The development of micro and small industries in village areas is directed at increasing overall community income and expanding employment opportunities. However, the development of micro and small industries in village areas in Indonesia is still uneven. Based on the Village Potential Survey (PODES) carried out by BPS in 2018, there are inequalities between regions in Indonesia in small micro industries. The number of micro and small industries in the village area in PODES 2018 amounted to 136,776 micro and small industry units, where 52% of the number of micro and small industries were located on Java Island (BPS Indonesia, 2018). This indicates that the development of small micro industries in the village area is centralized in the Java Island area. Development in the processing industry, especially small micro industries in areas outside Java Island, which has not been optimized, is a problem in the development of village areas in Indonesia. Therefore, clustering and looking at the characteristics of the Province based on PODES data in the small micro industry sector so that development policies and strategies can be carried out properly and on target. This clustering process is carried out using the cluster analysis method.

Cluster analysis is an analytical technique that aims to cluster data that has similarity between one data and other data into one cluster, so that data in one cluster has a high level of similarity and data between clusters has a low level of similarity (Nishom, 2019). Cluster analysis usually involves at least three steps. The first step is the measurement of similarity or not between data. Next is the clustering process where the observations are partitioned into clusters. The last step is to interpret the cluster results (Hair et al., 2013). Cluster analysis has the advantages of being able to group relatively large amounts of observational data and can be used on ordinal, interval, and ratio data (Talakua et al., 2017). In general, cluster analysis is divided into two methods, namely hierarchical and non-hierarchical. The hierarchical method is used when the number of clusters is unknown, while the non-hierarchical method is used when the number of clusters is defined in advance (Halkidi et al, 2001).There are several algorithms included in the non-hierarchical method, one of the algorithms is the Fuzzy C-Means algorithm. The Fuzzy C-Means algorithm is a cluster analysis technique that uses a fuzzy clustering model, where the formation of clusters is based on membership degree values that vary between 0 and 1. The advantages of the Fuzzy C-Means algorithm are handling overlapping clusters and providing good information (Cebeci & Yildiz, 2015). Based on previous research, it is also found that the Fuzzy C-Means algorithm has the best results on data with outliers and overlapping (Li et al., 2008). In addition, Fuzzy C-Means can also be used for high-dimensional data (Pramana et al., 2018). This research wants to use the Fuzzy C-Means algorithm in clustering provinces in Indonesia based on PODES data in the micro and small industry sector with Silhouette Coefficient cluster validation. The data in this research has outlier values in several provinces, which means that Fuzzy C-Means can be used in this research. The results of this research can be used to see the clustering of provinces in Indonesia based on the development of small micro industries in village areas and the characteristics of each province. This research was expected to help the provincial government in making policies and strategies for the development of small micro industries in village areas.

2. Methodology

2.1 Data and Variables

This research uses secondary data obtained through BPS publications related to PODES statistics in 34 provinces in Indonesia in 2021 in the micro and small industry sectors. The research conducted was the application of the Fuzzy C-Means algorithm to cluster the Provinces in Indonesia based on PODES data on the micro and small industry sector. This research analyzes 15 variables which are the ratio of micro and small industries in village areas based on the main raw materials such as Leather Industry (X_1), Furniture Industry (X_2), Metal Goods Industry (X_3), Textile Industry (X_4), Apparel Industry (X_5), Pottery Industry (X_6), Wood Industry (X_7), Food Industry (X_8), Beverage Industry (X_9), Tobacco Processing Industry (X_{10}), Paper Industry (X_{11}), Printing and Recording Media Reproduction Industry (X_{12}), Transportation Equipment Industry (X_{13}), Craft Industry (X_{14}), Machine Repair and Installation (X_{15}).

2.2 Principal Component Analysis

Principal Component Analysis (PCA) is an analytical technique to transform the original variables that are still correlated with one another into a new set of variables that are no longer correlated. The purpose of PCA is to reduce the dimensionality of the original data from independent variables to principal components. PCA emerged as a solution for the data collection process where the data consists of a large number of variables so that new variables are obtained that are fewer in number but still able to explain the variance of the data (Johnson & Wichern, 2007). Before PCA is performed, the data needs to fulfill the multicollinearity assumption first, which is seen by the VIF value. To calculate the VIF value, the following equation can be used (Ghozali, 2016).

$$VIF = \frac{1}{(1 - R^2 j)} \tag{1}$$

Where :

VIF = Variance Inflation Factor (VIF)

j = Number of samples (j = 1, 2, … k)

$\mathrm{R}^2\mathrm{j}$ = The determination coefficient of the j-th independent variable with other variables.

If there is multicollinearity between variables, PCA is performed.The following are the steps in PCA based on correlation matrix (Johnson & Wichern, 2007):

a. Calculate the variance-covariance matrix S

b. Calculate the eigenvalues of the variance-covariance matrix

$$|\boldsymbol{S} - \lambda I| = 0$$

c. $\mathrm{ê}_\mathrm{i}$ is the eigenvector obtained from each eigenvalue λ_m that satisfies

$$(S - \lambda I).\hat{e}_i = 0$$

d. Calculate the proportion of variance

$$Total\ Variance = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_1} \times 100\%$$

e. Calculate the value of the cumulative proportion of the variance of the original data that can be explained by the k-th principal component

$$Total\ Cumulative\ Variance = \frac{\sum_{j=1}^{n} \lambda_j}{\sum_{i=1}^{m} \lambda_1} \times 100\% \tag{2}$$

2.3 Determination of K Clusters

To determine the optimal number of clusters, the Elbow Method is used by looking at the largest decrease in the Sum of Square Error (SSE) value between the number of clusters and then followed by a small decrease in value. The decrease in the SSE value can also be seen from the point that forms the elbow. To calculate the SSE value, the following equation can be used (Muningsih, 2017).

$$SSE = \sum_{i=1}^{k} \sum_{x_{ij} \epsilon V_{kj}} \left\| x_{ij} - v_{kj} \right\|^2 \tag{3}$$

Where :

k = Number of cluster

x_{ij} = i-th object on j-th variable

v_{kj} = Centroid of the i-th object on j-th variable

i = Province

2.4 Fuzzy C-Means

The main concept in Fuzzy C-Means algorithm is to determine the cluster center that will mark the average location for each cluster. Each data point has its membership degree value in each cluster. The membership degree value is scaled between 0 and 1 which represents how close the object is to the cluster center. The following are the steps in the Fuzzy C-Means algorithm (Kusumadewi & Purnomo, 2004):

a. Input data to be clusterized.
b. The number of clusters (k), the partition matrix's power (w) and the least anticipated error (ξ).
c. Generate randome integers μ_{ik}, where i = 1,2,....,n,; k = 1,2,....,c; as the elements of the initial partition matrix U.
d. Calculate the k-th cluster center for the j-th attribute: V_{kj}, with k = 1,2,....,c; and j = 1,2,....,m.

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik})^{w}.X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}} \tag{4}$$

e. Calculate the objective function at the iteration to-t.

$$P_t = \sum_{k=1}^{c}\sum_{i=1}^{n}\left(\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right](\mu_{ik})^{w}\right) \tag{5}$$

f. Calculate the change in the partition matrix.

$$\mu_{ik} = \frac{[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}} \tag{6}$$

g. Check the stop condition:
If ($|P_t - P_{t-1}| < \xi$) or (t > MaxIter) then stop
If no: Continue to step 4 with t = t+1.

2.5 Cluster Validation with Silhouette Coefficient

Cluster validation is applied to measure how well and accurately the clusters are formed. One method to measure the accuracy of clusters is the silhouette coefficient by measuring the average value of the distance between clusters. The following are the steps in calculating the silhouette coefficient (Handoyo & Nasution, 2014).

a. Calculate the average distance of an object, e.g. the i-th object, to all other objects in a cluster.

$$a(i) = \frac{1}{|A| - 1}\sum_{j\in A, j\neq i} d(i,j)$$

Where :
$\mathrm{a(i)}$ = The average difference of an object (i) to all other objects in A
A = Cluster (number of data in the A-th cluster)
$\mathrm{d(i,j)}$ = The distance between object i and j

b. Calculate the average of the ith object with all objects in the other clusters.

$$d(i,C) = \frac{1}{|C|}\sum_{j\in C} d(i,j)$$

Where :
$\mathrm{d}(\mathrm{i},\mathrm{C})$ = The average difference of an object (i) to all other objects in C
C = Number of other cluster data except cluster A and C ≠ cluster A

c. After calculate $\mathrm{d}(\mathrm{i},\mathrm{C})$ for all C, then the smallest value is taken using the following equation.

$$b(i) = min_{C \neq A} d(i, C)$$

d. Calculate the silhouette coefficient.

$$s(i) = \frac{(b(i) - a(i))}{max\, a(i), b(i)} \tag{7}$$

The average silhouette coefficient value is in the interval $-1 \leq s(\mathrm{i}) \leq 1$. The more the average silhouette coefficient value approaches 1, the better the clustering of data in one cluster. Otherwise, if the average value of the silhouette coefficient is close to -1, the better the clustering of the data in one cluster (Kauffman & Rousseeuw, 2005).

Table 1: Silhouette Coefficient Criteria

Silhouette Coefficient Value	**Criteria**
$\leq 0{,}25$	Poor cluster structure
$0{,}26 - 0{,}50$	Weak cluster structure
$0{,}51 - 0{,}70$	Good cluster structure
$0{,}71 - 1{,}00$	Strong cluster structure

3. Results

Before processing the data, a multicollinearity test will be carried out by looking at the Variance Inflation Factor (VIF) value. Table 2 below shows the VIF value of each variable used in this research.

Table 2: VIF Value.

Variable	**VIF**
Leather Industry	15,800
Furniture Industry	13,206
Metal Goods Industry	21,203
Textile Industry	4,119
Apparel Industry	2,275
Pottery Industry	13,398
Wood Industry	10,583
Food Industry	9,215
Beverage Industry	12,168
Tobacco Processing Industry	5,322
Paper Industry	13,317
Printing and Recording Media Reproduction Industry	24,944
Transportation Equipment Industry	3,535
Craft Industry	14,696
Machine Repair and Installation	19,934

Based on the multicollinearity test results shown in Table 2, it shows that there are several variables that have a VIF value > 10 which indicates the presence of multicollinearity between variables. Therefore, Principal Component Analysis (PCA) was conducted.

3.1 Principal Component Analysis

For the application of PCA requires the assumption of sample representation conducted with the Kaiser Mayer Olkin (KMO) test and the assumption of correlation conducted with the Bartlett test of Sphericity. Table 3 below shows the results of the KMO and Bartlett test of Sphericity on this research data.

Table 3: KMO and Bartlett Test of Sphericity

KMO	**0,8114899**
Bartlett Sphericity (Khi-Squared)	660,552
Degree of Freedom (df)	105
sig	$2{,}05 \times 10^{-81}$

Based on Table 3, it shows that the KMO value is 0.8114899. Based on the KMO criteria indicates that the data in this research has a correlation between variables that is sufficient to produce valid result. Next, the number of main components is determined by looking at the eigenvalues. Table 4 below shows the eigenvalue to determine how many main components will be used.

Table 4: Eigenvalue

Component	Eigenvalue	%Variance	%Cumulative
1	9,613	64,090	64,090
2	1,942	12,947	77,037
3	1,118	7,457	84,494
4	0,710	4,735	89,228
5	0,486	3,238	92,467
6	0,363	2,418	94,885
7	0,293	1,951	96,836
8	0,127	0,844	97,680
9	0,093	0,619	98,299
10	0,082	0,544	98,844
11	0,051	0,342	99,186
12	0,044	0,296	99,482
13	0,032	0,212	99,694
14	0,026	0,173	99,867
15	0,020	0,133	100,000

In Table 4, it can be seen that there are 3 components that have eigenvalues > 1, which indicates that the 3 components are considered to represent 15 variables in this research data. Table 4 also shows the cumulative variance value of the 3 components, which is 84.494%, which means that the 3 components have represented 84.494% of this research data.

The next step is to calculate the principal component score. This principal component score will represent each data. Table 5 below shows the principal component scores of the 34 provinces.

Table 5: Principal Component Score

Province	PC1	PC2	PC3
Aceh	-58,1725385	8,7344003	-3,52818067
Sumatera Utara	-40,3297458	9,6209737	-9,0136382
Sumatera Barat	76,0984403	5,2171638	-7,92065856
Riau	-8,640701	1,4119319	-3,15408645
Jambi	-16,5884619	8,3432475	-1,31742192
Sumatera Selatan	-31,7330183	4,2725994	5,59153519
Bengkulu	-35,5925459	3,6044487	6,31599185
Lampung	26,302626	-10,7956911	-2,47482912
Kepulauan Bangka Belitung	65,7193459	-28,6744505	11,60237238
Kepulauan Riau	25,403398	-26,3245509	3,21568999
DKI Jakarta	-0,8501869	35,3399096	-32,86914118
Jawa Barat	51,6394281	7,325061	-9,30900106
Jawa Tengah	62,0732839	9,6768628	-10,17061109
DI Yogyakarta	105,0880966	24,4537743	11,26953779
Jawa Timur	46,3877367	9,3122111	2,46953473
Banten	30,95475	1,8416638	-8,04751434
Bali	84,5578327	11,4764829	19,06261422
Nusa Tenggara Barat	40,1751407	3,5956047	28,28036588
Nusa Tenggara Timur	-13,3977324	23,0923755	6,75405827
Kalimantan Barat	-39,9121728	-1,9871207	11,80351649
Kalimantan Tengah	-40,671123	-9,8556149	10,27121963
Kalimantan Selatan	-13,7389541	-11,6843771	-14,963235
Kalimantan Timur	-14,1183871	-9,6277281	-8,44183456
Kalimantan Utara	-58,7894387	9,8403409	14,90220561
Sulawesi Utara	-29,0224595	-12,4595723	-14,88167813
Sulawesi Tengah	-13,3715663	-11,9821182	0,09261807
Sulawesi Selatan	19,4823516	0,9872437	-7,34803619
Sulawesi Tenggara	5,1060329	-20,9457288	-5,10995725
Gorontalo	29,9967114	-19,4663481	-9,49182841
Sulawesi Barat	-3,2611125	5,7249272	4,28852117
Maluku	-30,2740324	-16,5157293	-4,82867982
Maluku Utara	-27,7170586	-28,0545559	-0,61085859
Papua Barat	-93,1231185	11,0970111	7,57717388
Papua	-99,6808207	13,4053521	9,98423536

Based on Table 5, the Principal Component (PC) score is obtained for all provinces in Indonesia. This score will be used for clustering provinces in Indonesia.

3.2 Determination of K Clusters

Before cluster analysis, the optimal number of clusters was determined using the Elbow Method which can be seen in Figure 1.

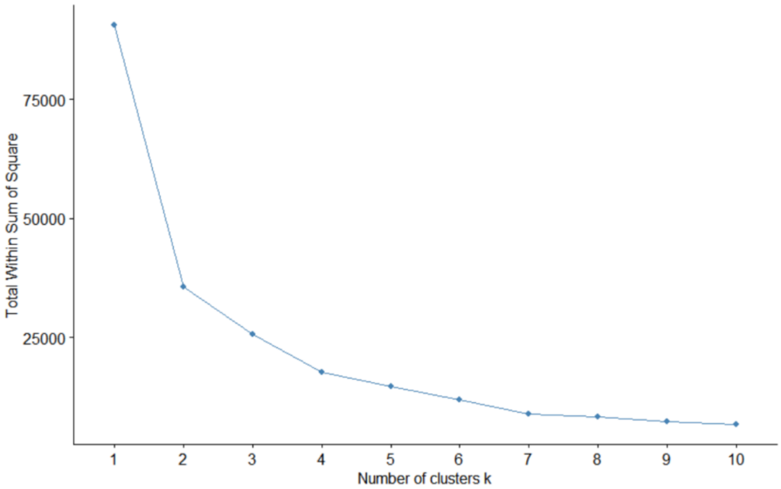

Figure 1: Elbow Method

Based on Figure 1, there is a point that forms an elbow between cluster 3 and cluster 4. The point that forms an elbow is characterized by a large decrease in value between two points followed by a small decrease in the next point. Therefore, the optimal number of clusters for clusterization using the Fuzzy C-Means algorithm is 4 clusters because there is a large enough decrease between cluster 3 and cluster 4 and followed by a small decrease in value in the next cluster.

3.3 Fuzzy C-Means

After determining the optimal number of clusters, namely 4 clusters, the clustering process will then be carried out with the Fuzzy C-Means algorithm. To see how the clustering of provinces in Indonesia based on the ratio of micro and small industries in the village area, the membership degree value in each cluster is used. Table 6 below shows the membership degree value at the last iteration.

Table 6: Membership Degree

Province	Membership Degree			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Aceh	0,019736	0,046315	0,249049	0,684901
Sumatera Utara	0,027047	0,074289	0,685704	0,212961
Sumatera Barat	0,934574	0,047410	0,012709	0,005306
Riau	0,027654	0,145150	0,789204	0,037992
Jambi	0,019665	0,076243	0,862218	0,041874
Sumatera Selatan	0,014779	0,045685	0,863454	0,076082
Bengkulu	0,018402	0,054212	0,806819	0,120567
Lampung	0,004123	0,990654	0,004279	0,000945
Kepulauan Bangka Belitung	0,524655	0,350569	0,089454	0,035322

Kepulauan Riau	0,088091	0,780805	0,105484	0,025620
DKI Jakarta	0,150480	0,328363	0,381804	0,139353
Jawa Barat	0,545646	0,375479	0,058871	0,020004
Jawa Tengah	0,805681	0,149543	0,032465	0,012310
DI Yogyakarta	0,771109	0,138681	0,059411	0,030799
Jawa Timur	0,455957	0,456504	0,065997	0,021542
Banten	0,061000	0,889329	0,039801	0,009870
Bali	0,877580	0,081123	0,028202	0,013096
Nusa Tenggara Barat	0,355961	0,474753	0,125700	0,043585
Nusa Tenggara Timur	0,068472	0,203326	0,604269	0,123933
Kalimantan Barat	0,025206	0,070762	0,698911	0,205122
Kalimantan Tengah	0,025995	0,074563	0,704944	0,194497
Kalimantan Selatan	0,030525	0,144895	0,774612	0,049968
Kalimantan Timur	0,017451	0,085269	0,867133	0,030147
Kalimantan Utara	0,017848	0,039961	0,182022	0,760169
Sulawesi Utara	0,020994	0,073440	0,835887	0,069679
Sulawesi Tengah	0,018677	0,092736	0,856970	0,031617
Sulawesi Selatan	0,039198	0,884486	0,064002	0,012314
Sulawesi Tenggara	0,064216	0,570845	0,320170	0,044768
Gorontalo	0,060200	0,873049	0,053556	0,013194
Sulawesi Barat	0,053693	0,285602	0,604009	0,056695
Maluku	0,017151	0,059779	0,860336	0,062734
Maluku Utara	0,038768	0,134165	0,716298	0,110769
Papua Barat	0,008082	0,015241	0,043202	0,933474
Papua	0,014828	0,027024	0,070938	0,887209

Based on Table 6, information is obtained from the membership degree value of each province in Indonesia which shows the tendency of a province to belong to a cluster. The greater the value of the membership degree will show the highest tendency of a province to become a member of a cluster. A summary of clustering results using the Fuzzy C-Means algorithm can be seen in Table 7.

Table 7: Fuzzy C-Means Algorithm Clustering Results

Cluster	**1**	**2**	**3**	**4**
Total Province	6	8	16	4

Based on Table 7, the clustering of provinces in Indonesia based on the ratio of micro and small industries in village areas shows that most provinces are in cluster 3. The characteristics of each cluster formed based on the average value can be seen in Table 8.

Table 8: Characteristics of Clusters Formed Based on Average Value

Cluster	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	**Category**

1	12,7	81,6	52,3	27,1	64,3	43,4	62,5	83,6	75,4	3,3	2,54	26,9	5,16	19,6	29,2	Very High
2	4,11	67,6	31	17,6	40,2	34,4	46,1	72,8	60,4	4,05	0,849	16,2	6,44	6,97	19,2	High
3	2,4	44	18,3	12,6	24,2	15,4	29,8	48,3	43,2	0,527	0,374	8,76	6,02	4,87	12,3	Medium
4	0,425	15,5	7,13	4,06	10	7,34	16,1	16,4	15,2	0,378	0,098	3,13	3,73	3,59	4,41	Low

Table 8 shows the characteristics of each cluster formed based on the averages. Cluster 1 is categorized as a province with a very high ratio of micro and small industries in village areas characterized by a very high average value. Cluster 4 has a low average value which can be interpreted as a province with a low ratio of micro and small industries in the village area. The distribution of clustering results with the Fuzzy C-Means algorithm can be seen in Figure 2.

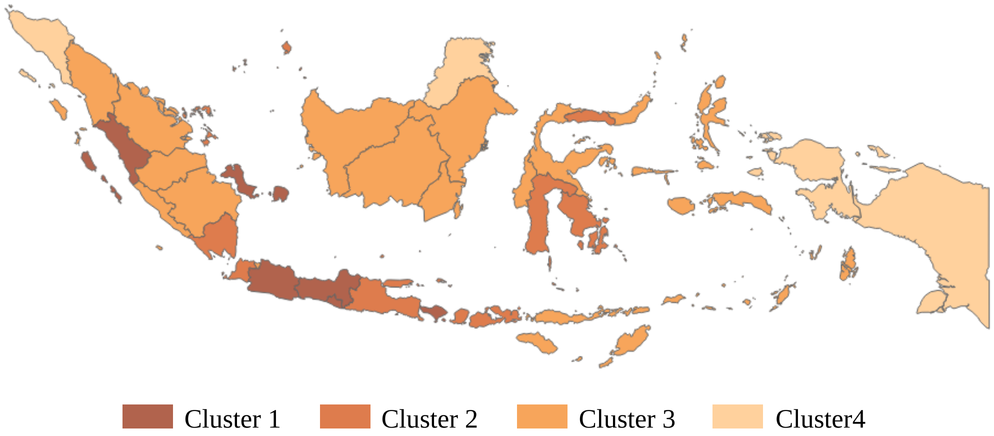

Figure 2: Provincial Clustering Distribution with Fuzzy C-Means Algorithm

Figure 2 shows the distribution of clustering of Provinces with the Fuzzy C-Means algorithm based on the ratio of small micro industries in the village area. Cluster 3 and 4 is dominated by provinces outside Java, namely all provinces on the island of Kalimantan, Maluku, and Papua, which can be interpreted that the development of micro and small industries in the village area is still not optimal in provinces that are in the ratio of micro and small industries with medium and low categories. Furthermore, cluster 1 and 2 contains provinces with a very high and high ratio of small micro industries in the village area. Almost all provinces on Java Island are classified into cluster 2 and there are several provinces such as West Sumatra, Riau Islands, Lampung, Bangka Belitung, Bali, West Nusa Tenggara, South Sulawesi, Southeast Sulawesi and Gorontalo. This shows that the development of small micro industries in the village area is still concentrated in provinces close to the center of government, namely Java Island, and there is still a need to develop small micro industries in provinces far from the center of government such as Kalimantan Island, Papua, Maluku, and several provinces on Sumatra Island and Sulawesi Island to create equality between regions.

To see how well and accurately the clustering is done with the Fuzzy C-Means algorithm, the silhouette coefficient value is measured. The results of cluster validation with silhouette coefficient can be seen in Table 9.

Table 9: Cluster Validation of Fuzzy C-Means Algorithm

Validation Index	**Fuzzy C-Means**
Silhouette Coefficient	0,6406

Table 9 shows the results of measuring cluster validation with the silhouette coefficient and obtained a value of 0.6406. Therefore, it can be interpreted that clustering with the Fuzzy C-Means algorithm produces a good cluster structure based on the silhouette coefficient value.

4. Conclusions

The Fuzzy C-Means algorithm produces a good cluster structure based on the silhouette coefficient value. Clustering Provinces in Indonesia based on the ratio of micro and small industries in rural areas using the Fuzzy C-Means algorithm obtained 4 optimal clusters calculated by the Elbow Method. Cluster 1 contains 6 provinces with a ratio of micro and small industries in village areas classified in the very high category. Cluster 2 contains 8 provinces with a very high ratio of micro and small industries in the village area category. Cluster 3 contains 16 provinces with a ratio of micro and small industries in village areas classified in the medium category. Cluster 4 contains 4 provinces with low ratio of micro and small industries in village areas. Based on the clustering results, it is necessary to develop the micro and small industry sectors in the village areas of provinces in Indonesia, especially in provinces that are categorized in the medium and low category. Therefore, policies and strategies are needed by the Provincial government in the development of small micro industries in village areas directed at increasing overall community income and expanding employment to reduce unemployment and poverty problems.

References

Badan Pusat Statistik. (2018). *Statistik Potensi Desa 2018*. Jakarta: Badan Pusat Statistik.

Badan Pusat Statistik. (2020). *Profil Industri Mikro dan Kecil 2019*. Jakarta: Badan Pusat Statistik.

Cebeci, Zeynel, & Figen Yildiz. (2015). “Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures”. *Journal of Agricultural Informatics* 6(3):13–23. doi: 10.17700/jai.2015.6.3.196.

Ghozali, I. (2016). *Aplikasi Analisis Multivariate dengan Program IBM SPSS 25 Edisi 9*. Semarang: Badan Penerbit Universitas Diponegoro.

Hair JR, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate Data Analysis*.7th ed. Edinburgh: Pearson.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(3), 107–145. https://doi.org/10.1023/A:1012801612483.

Handoyo, R., M, R. R. & Nasution, S. M., (2014). Perbandingan Metode Clustering menggunakan metode single linkage dan k-means pada pengelompokan dokumen. Volume 15, pp. 73-82.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis Sixth Edition*. New Jersey: Pearson Education,Inc.

Kaufman, L. & Rousseeuw, P.J. (2005) *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley.

Kusumadewi, S & Purnomo, P. (2004). *Aplikasi Logika Fuzzy Untuk Mendukung Keputusan*. Yogyakarta: Graha Ilmu.

Li MJ, Ng MK, Cheung Y, Huang JZ. (2008). Agglomerative Fuzzy K-means Clustering Algorithm with Selection of Number of Clusters. *IEEE Transactions on Knowledge and Data Engineering*, 20.

Muningsih, E. (2017). Optimasi jumlah cluster k-means dengan metode elbow untuk pemetaan pelanggan. *Pros. Semin. Nas. ELINVO*, 105-114.

Nishom M, (2019), Perbandingan Akurasi Euclidean Distance, Minkowski Distance dan Manhattan Distance Pada Algoritma K-Means Clustering Berbasis Chi-Square. *Jurnal Informatika : Jurnal Pengembangan IT*, 4(1), 20-24. doi: 10.30591/jpit.v4i1.1253.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., & Nooraeni, R. (2018). *Data Mining dengan R : Konsep serta Implementasi*. Bogor: In Media.

Soleh, A. (2017). Strategi pengembangan potensi desa. *Jurnal Sungkai*, 5(1), 32-52.

Talakua, M. W., Leleury, Z. A., & Taluta, A. W. (2017). *Analisis cluster dengan menggunakan metode k-means untuk pengelompokkan Kabupaten/Kota di provinsi maluku berdasarkan indikator indeks pembangunan manusia tahun 2014. BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 11(2), 119-128.