International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

The Clustering of the Aquaculture Fisheries Companies in Indonesia Using the K-Prototypes and Two Step Cluster (TSC) Algorithm

Sri Sulastri[a], Budi Susetyo[b*], I Made Sumertajaya[c]

[a,b,c]*Department of Statistics, IPB University, Bogor, 16680, Indonesia*

[a]*Email: sulastri_bps@apps.ipb.ac.id;* [b]*Email: budisu@apps.ipb.ac.id;* [c]*Email: imsjaya@apps.ipb.ac.id*

Abstract

Background: Fisheries subsector has an important role in the Indonesian economy, especially for the aquaculture fisheries companies. Each aquaculture fisheries companies has its own characteristics like in terms of technical, financial, staffing, or input and output structures. It is necessary to clustering 258 aquaculture fisheries companies to make it easier to identify the characteristics of these different companies based on the characteristics of their cluster. One of the method that can be used to grouping objects is cluster analysis. On this study, the clustering process was using the K-Prototypes and Two Step Cluster (TSC) algorithm because the data that used in this study was the mixed data type (13 numerical and 8 categorical variables). Then this study would choose the best algorithm by the smallest ratio between the standard deviation within the cluster (S_W) and the standard deviation between cluster (S_B). The smallest ratio means that the diversity within clusters is quite homogeneous, while the diversity between clusters is heterogeneous. Based on the comparison of the ratio between S_W and S_B from the k-prototypes and the TSC algorithm, the k-prototypes algorithm with 6 clusters was the best algorithm for clustering the aquaculture fisheries companies in Indonesia. The result showed that the cluster 5 was the best cluster and the cluster 6 was the worst cluster related to the condition of the aquaculture fisheries companies in Indonesia. Cluster 5 which is characterized by most of the central companies in the form of PT and do the enlargement of sea water fish in fishpond and has a high numerical variable value. Cluster 6 which is characterized by most of the central companies in the form of PT and CV and do the hatchery of land water fish in water tubs and has the lowest value compared to other clusters.

Keywords: Cluster; Fisheries; K-Prototypes; Mixed Data; Two Step Cluster.

---

* Corresponding author.

1. Introduction

Indonesia is one of the most fish production country in the world. The Food and Agriculture Organization of The United Nations (2020) stated that in 2018, Indonesia was ranked second as the most capture fisheries production country (after China), and was ranked third as the most aquaculture production country (after China and India) in the world [1]. Even according to The Statistics Indonesia (BPS), the fisheries subsector contributed 2.83 percent to Indonesia's Gross Domestic Product (GDP) and occupied the third largest position in the agriculture, forestry and fisheries sector in the second quarter of 2020. The agriculture, forestry and fisheries sector was also the second largest contributing sector to GDP Indonesia with a contribution of 15.46 percent [2]. In general, fisheries are divided into capture and aquaculture fisheries. It is noted that starting from 2010 until now, aquaculture fisheries has dominated fish production compared to capture fisheries in Indonesia. The latest data from The Ministry of Marine Affairs and Fisheries of Indonesia said that the aquaculture fisheries contributed 66.06 percent (as much as 3.83 million tons) of total national fish production in the fourth quarter of 2019, while the remaining 33.94 percent was contributed from the capture fisheries [3]. So, the government of Indonesia should pay attention to the importance of the fisheries subsector, especially the aquaculture fisheries. This is also in line with the 14th goal in the Sustainable Development Goals (SDGs), which is to improve the quality of management and marine resources. The most significant source of the aquaculture fisheries production is through the aquaculture fisheries companies. BPS noted that there were 258 aquaculture fisheries companies active in 2018 [4]. Each aquaculture fisheries companies has its own characteristics like in terms of technical, financial, staffing, or input and output structures. It is necessary to clustering the aquaculture fisheries companies in Indonesia to make it easier to identify the characteristics of these different companies based on the characteristics of their cluster. This is also necessary for the government (The Ministry of Marine Affairs and Fisheries) not to be mistaken in determining policies related to the aquaculture fisheries companies in Indonesia. In statistics, one of the method that can be used to grouping objects is cluster analysis. The grouping is done based on the level of similarity (homogeneity) and dissimilarity (heterogeneity) of all research objects. The author in [5] stated that there are two clustering methods in generally, namely the hierarchical clustering method and the nonhierarchical clustering method. The difference between the two methods is in the arrangement of determining the number of groups. Some examples of hierarchical clustering methods are single linkage, complete linkage, average linkage, centroid linkage, median likage, and Ward's method. While some examples of nonhierarchical clustering methods are k-means [6] and k-modes [7]. The classical method of hierarchical and nonhierarchical clustering is usually used for processing one type of data only, such as only numerical data types or only categorical data types. Therefore, it is needed to developing an algorithm that can process mixed data types (numerical and categorical data types). This study involved 258 aquaculture fisheries companies and a large number of variables (21 variables) that consisting of 13 numerical variables and 8 categorical variables. In this study, an appropriate processing algorithm is needed to clustering objects with mixed data types. There are several suitable algorithms, including K-Prototypes , Two Step Cluster [8], Cluster Ensemble [9], Gower Method [10], and Latent Class Cluster [11]. However, the constraints of this study is only use two algorithms (K-Prototypes and Two Step Cluster (TSC) algorithms) on clustering the aquaculture fisheries companies in Indonesia. K-Prototypes is the development of the k-means method (for data with numerical variables only) and the k-modes method (for data with categorical variables only), so that the K-Prototypes algorithm can be used

for clustering mixed data types (numerical and categorical variables). In addition, this method does not use complex algorithms and it better than other hierarchical algorithms. The author in [12] said that this algorithm basically divides the data set into a predetermined number of k groups and then estimates the center point of the cluster. This estimation of the center point of the cluster is the weakness of the K-Prototypes algorithm because the different center point can produce the different final clusters too . The k-prototypes algorithm has been commonly used to clustering objects, for example the author in [13] used this algorithm to classifying all villages and subdistrict in Indonesia, as well as the author in [14] used this algorithm to classifying the flow of tweets on Twitter based on their level of importance. The Two Step Cluster (TSC) algorithm was first introduced by Chiu and his colleagues where this algorithm can also be used for processing mixed data types. This algorithm is suitable for dealing with unequal data measurement scales and for clustering the relatively large number of observations. This algorithm is more complex than other methods and it is one of the weakness of the TSC algorithm. The processing includes two stages, namely the initial stage (preclustering stage) and the final clustering stage. This algorithm has also been commonly used in clustering objects, for example by the author in [15] to classifying patients suffering from polycystic syndrome based on the characteristics and symptoms experienced by the patient and as well as by the author in [16] to clustering 608 individual samples based on their respective characteristics. This study will evaluate the results of clustering the aquaculture fisheries companies in Indonesia using the K-Prototypes and Two Step Cluster (TSC) algorithms based on the variance's cluster values. Previous study about comparison between K-Prototypes and Two Step Cluster (TSC) algorithms have also been carried out by the authors in [17] with a different research object. They are clustering villages and subdistricts in East Nusa Tenggara Province of Indonesia by using several indicators of poverty in that area, where the results showed that the TSC algorithm was the best clustering algorithm than the K-Prototypes algorithm. That is because the results obtained that using the TSC algorithm have a smaller ratio between S_W and S_B.

2. Materials and Methods

The data that used in this study is secondary data from the 2018 Annual Report of Aquaculture Fisheries Companies that conducted by Statistics Indonesia (BPS) and involved 258 aquaculture fisheries companies (Table 1). This study compares two cluster analysis algorithms, namely K-Prototypes and Two Step Cluster algorithm. The author in [18] said that the degree of similarity between objects is determined by the distance from these objects. The greater distance between two objects, so the more different that two objects, and vice versa. According to the author in [7], the K-Prototypes algorithm uses a mixed data type distance measure with the following calculation formula:

$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 + \gamma \sum_{s=p+1}^{m} \delta(x_{is}, x_{js}) \qquad (1)$$

with:

d_{ij} = a mixed data type distance that obtained between the i-th object and the j-th object

$\sum_{k=1}^{p}(x_{ik} - x_{jk})^2$ = a distance for numerical variables

$\gamma \sum_{s=p+1}^{m} \delta(x_{is}, x_{js})$ = a distance for categorical variables

The gamma coefficient (γ) is the average standard deviation (σ) of all numerical variables that used in this study. Meanwhile, according to the author in [19], the Two Step Cluster (TSC) algorithm uses the log-likelihood distance measure with the following calculation formula:

$$d(i,j) = \xi_i + \xi_j - \xi_{(i,j)} \quad (2)$$

with:

$$\xi_i = -N\left(\sum_{k=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{ik}^2\right) - \sum_{k=1}^{K^B}\sum_{l=1}^{L_k}\frac{N_{ikl}}{N_i}log\left(\frac{N_{ikl}}{N_i}\right)\right) \quad (3)$$

$$\xi_j = -N\left(\sum_{k=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{jk}^2\right) - \sum_{k=1}^{K^B}\sum_{l=1}^{L_k}\frac{N_{jkl}}{N_i}log\left(\frac{N_{jkl}}{N_i}\right)\right) \quad (4)$$

$$\xi_{(i,j)} = -N\left(\sum_{k=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{(ij)k}^2\right) - \sum_{k=1}^{K^B}\sum_{l=1}^{L_k}\frac{N_{(ij)kl}}{N_i}log\left(\frac{N_{(ij)kl}}{N_i}\right)\right) \quad (5)$$

d(i,j) = the distance between the i-th cluster and the j-th cluster

N = many objects

Ni = the number of objects in the i-th cluster

Nikl = the number of objects in the i-th cluster for the k-th categorical variable with the l-category

$\hat{\sigma}_k^2$ = estimate of the variance in the k-th numerical variable for all objects

$\hat{\sigma}_{ik}^2$ = estimate of the variance in the k-th numerical variable for all objects in the i-th cluster

KA = the number of numerical variables

KB = the number of categorical variables

Lk = the number of categories for the k-categorical variable

The evaluation of clustering between two algorithms can be measured by the ratio between the standard deviation within the cluster (S_W) and the standard deviation between cluster (S_B). The smaller ratio between S_W and S_B means that the clustering process can be said to be optimal.

Table 1: Description of variables

Variable	The name of variable	Data type	Explanation
X1	Form of business entity	Categoric	1 = PN/PD/Persero/Perum
			2 = PT
			3 = CV
			4 = Firm
			5 = Cooperative
X2	Capital status	Categoric	1 = Foreign capital
			2 = Domestic capital
			3 = Others
X3	Company status	Categoric	1 = Branchless
			2 = Central
			3 = Branch
X4	Type of activity	Categoric	1 = Hatchery
			2 = Enlargement
X5	The main types of aquaculture	Categoric	1 = Sea water
			2 = Brackish water
			3 = Land water
X6	The main of aquaculture container	Categoric	1 = Fishpond
			2 = Floating net cage
			3 = Water tub
			4 = Span rope
			5 = Karamba
			6 = Others
X7	Aquaculture technology	Categoric	1 = Simple
			2 = Semi intensive
			3 = Intensive
X8	Export activities	Categoric	1 = Export
			2 = No
X9	The number of employees	Numeric	Unit = people
X10	The owned land area	Numeric	Unit = thousand m^2
X11	The land area does not owned it	Numeric	Unit = thousand m^2
X12	The expenditures for worker's wages	Numeric	Unit = million rupiah
X13	The expenditures for materials and services	Numeric	Unit = million rupiah
X14	The expenditures for fuel and lubricants	Numeric	Unit = million rupiah
X15	The expenditures for purchasing seeds	Numeric	Unit = million rupiah
X16	The expenditures for purchasing broodstock	Numeric	Unit = million rupiah
X17	The expenditures for electricity and gas	Numeric	Unit = million rupiah
X18	The expenditures for medicines	Numeric	Unit = million rupiah
X19	The expenditures for feed	Numeric	Unit = million rupiah
X20	The expenditures for fertilizer	Numeric	Unit = million rupiah
X21	Income	Numeric	Unit = million rupiah

3. Result and Discussion

3.1. Data Description

Based on Table 2, each aquaculture fisheries company has an average of 38 employees and controls 87.40

thousand m2 owned land area. It can be said that in 2018 all aquaculture fisheries companies experienced a surplus or obtained business profits based on the value of expenditure and income. The large standard deviation value also indicates that the business scale of aquaculture fisheries companies in Indonesia are diverse, some are small scale and some are large scale.

Table 2: Descriptive statistics of numerical variables (X9 to X21)

Variables	Average	Minimum	Maximum	Standard deviation
X9	38.49	1.00	438.00	58.22
X10	87.40	0.00	5,000.28	354.34
X11	85.97	0.00	4,720.00	458.41
X12	1,178.80	0.00	17,123.00	2,218.52
X13	506.83	0.00	18,353.99	1,571.36
X14	272.19	0.00	6,768.00	766.58
X15	630.70	0.00	16,065.10	1,664.53
X16	191.10	0.00	10,920.00	984.72
X17	48363	0.00	9,300.00	1,146.16
X18	218.09	0.00	23,458.89	1,513.14
X19	2,348.13	0.00	51,024.54	6,562.22
X20	47.24	0.00	2,133.80	195.33
X21	8,752.00	0.00	96,518.00	14,142.39

3.2. K-Prototypes Algorithm

It is necessary to set the set.seed so that the center point of the cluster will always be the same even though there is repeated processing, so that in this study the value of set.seed (100) is used. Based on the processing results, the weighting coefficient (γ) was same for every number of cluster from the number of cluster (k) = 2 to k = 10, which was 1.9839. The value of the weighting coefficient (γ) was determined by the number of numerical variables, categorical variables, and the number of objects that used in this study. Based on Figure 1, there is a fluctuation in the ratio between S_W and S_B when k = 2 to k = 10. Overall, the optimal number of cluster is obtained when k = 6 because it has the smallest ratio value. Therefore, the optimal clustering of the aquaculture fisheries companies using the k-prototypes algorithm was as many as 6 clusters with the distribution of members per cluster can be seen in Table 3.

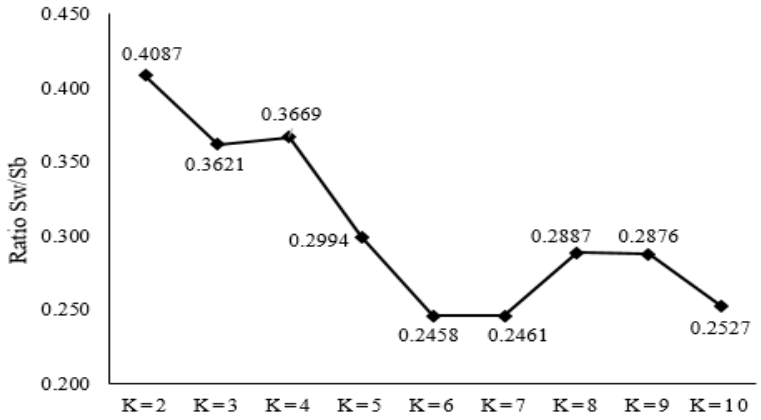

Figure 1: Ratio between S_W and S_B from K-Prototypes algorithm

Table 3: The distribution of members per cluster using K-Prototypes algorithm

Cluster	Members	Percentage *(%)*
1	8	3.10
2	65	25.19
3	19	7.36
4	75	29.07
5	14	5.43
6	77	29.85
Total	258	100.00

3.3. Two Step Cluster (TSC) Algorithm

The first step of the TSC algorithm is to determine the maximum number of cluster that can be formed through the BIC change ratio value that is closest to $c_1 = 0.04$. Therefore, the maximum number of cluster that can be formed in this study is 6 clusters with a BIC change ratio value of 0.002 (Table 4). The next step is to determine the optimal number of clusters based on whether or not there is a significant difference in the cluster change ratio value. The largest distance is seen when the number of clusters is 2 and 5 clusters. When the number of clusters is 2 clusters, the ratio of distance or $R(k_1)$ is 1.478 (the highest value), while the value of $R(k_2)$ is 1.429 for the number of clusters of 5 clusters (the second highest value). The ratio between $R(k_1)$ and $R(k_2)$ is 1.034 and this ratio is smaller than the value of c_2, which is 1.15. Therefore, the optimal number of clusters is obtained based on the maximum value of $\{k_1, k_2\}$, so that the optimal number of clusters using the TSC algorithm in this study is 5 clusters. Cluster 3 is the cluster with the most number of members which contain 72 aquaculture fisheries companies (Table 5). The next position based on the number of cluster members is occupied by cluster 5, cluster 4, cluster 2, and the last is cluster 1 (the cluster with the least number of members).

Table 4: BIC, change in BIC, BIC change ratio, and the distance ratio of each number of cluster that using the TSC algorithm

The number of clusters	BIC	Change in BIC	BIC change ratio	The distance ratio
1	6,150.803			
2	5,414.719	-736.084	1.000	1.478
3	4,997.436	-417.284	0.567	1.307
4	4,736.964	-260.472	0.354	1.421
5	4,627.625	-109.338	0.149	1.429
6	4,626.141	-1.485	0.002	1.308
7	4,683.868	57.728	-0.078	1.303
8	4,786.261	102.393	-0.139	1.047
9	4,895.259	108.998	-0.148	1.293
10	5,036.141	140.883	-0.191	1.044

Table 5: The distribution of members per cluster using TSC algorithm

Cluster	Members	Percentage *(%)*
1	19	7.36
2	39	15.12
3	72	27.91
4	59	22.87
5	69	26.74
Total	258	100.00

3.4. The Comparison of Clustering Method Using K-Prototypes and TSC Algorithm

The clustering results using the k-prototypes and the TSC algorithm resulted a different optimal number of clusters. For the k-prototypes algorithm, the optimal number of clusters is 6 clusters, while for the TSC algorithm 5 clusters. Figure 2 shows that most of the members of clusters 2, 4, and 6 of the k-prototypes algorithm are also members of clusters 5, 3, and 4 of the TSC algorithm. Members of the cluster 1 of the TSC algorithm are mostly members of cluster 1 and cluster 5 of the k-prototypes algorithm. Meanwhile, most of the member's cluster 2 of the TSC algorithm are members of cluster 3 and cluster 6 of the k-prototypes algorithm too. The clustering using the TSC algorithm also shows that most of the members of cluster 3 are members of cluster 4 of the k-prototypes algorithm. Furthermore, the TSC algorithm cluster 4 members are also mostly members of cluster 6 of the k-prototypes algorithm. The results also show that most of the TSC algorithm cluster 5 members are also predominantly members of cluster 2 of the k-prototypes algorithm.

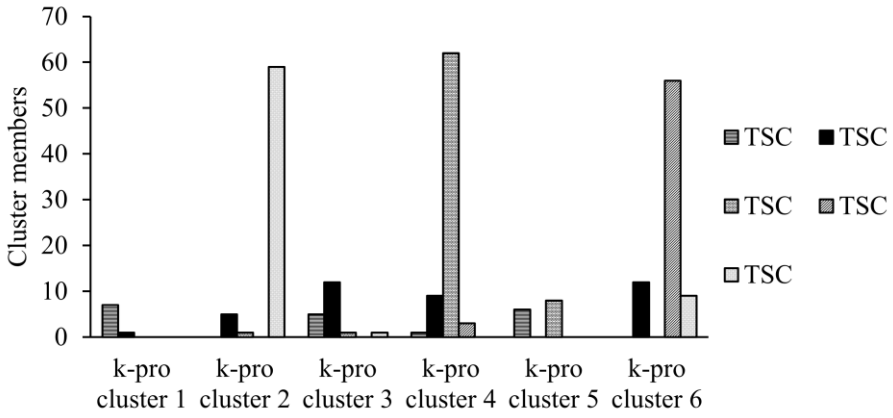

Figure 2: The comparison of cluster members using k-prototypes and TSC algorithm

3.5. Determination of The Best Clustering Algorithm

Table 6: Comparison of the ratio between S_W and S_B that using the k-prototypes and TSC algorithm

Variable	Data type	K-Prototypes algorithm with k = 6			Two Step Cluster algorithm with k = 5		
		S_W	S_B	S_W/S_B	S_W	S_B	S_W/S_B
X1	Categorical	0.5908	0.7239	0.8161	0.5788	1.2066	0.4797
X2	Categorical	0.4301	3.3833	0.1271	0.3450	4.2923	0.0804
X3	Categorical	0.4301	3.3833	0.1271	0.3450	4.2923	0.0804
X4	Categorical	0.2915	2.4912	0.1170	0.3240	2.5434	0.1274
X5	Categorical	0.6656	4.8182	0.1381	0.6686	5.3551	0.1248
X6	Categorical	1.3214	6.3378	0.2085	1.2958	7.3502	0.1763
X7	Categorical	0.7141	2.4611	0.2902	0.7416	2.2174	0.3345
X8	Categorical	0.3033	0.9955	0.3047	0.3082	1.0080	0.3058
X9	Numerical	32.4313	348.1419	0.0932	39.6109	344.2751	0.1151
X10	Numerical	337.0200	853.7744	0.3947	350.6043	540.3283	0.6489
X11	Numerical	448.3458	818.4828	0.5478	442.8058	1,048.5247	0.4223
X12	Numerical	1,324.4102	12,828.7653	0.1032	1,480.2978	13,327.7488	0.1111
X13	Numerical	1,408.8084	5,185.0019	0.2717	1,423.7093	5,517.2019	0.2580
X14	Numerical	641.0476	3,081.1364	0.2081	683.7921	2,860.4527	0.2391
X15	Numerical	1,509.2603	5,254.3225	0.2872	1,484.5729	6,214.1143	0.2389
X16	Numerical	796.5264	9,626.6390	0.0827	960.1761	1,997.1561	0.4808
X17	Numerical	887.9637	5,271.0445	0.1685	1,097.5543	2,865.5796	0.3830
X18	Numerical	1,418.2876	4,037.7217	0.3513	1,452.6593	3,692.5109	0.3934
X19	Numerical	4,297.7291	35,812.2205	0.1200	5,594.8453	28,051.9830	0.1994
X20	Numerical	177.8469	605.7295	0.2936	186.0261	512.3315	0.3631
X21	Numerical	8,821.0036	79,741.6988	0.1106	9,430.4090	85,002.7762	0.1109
Average		1,052.6394	7,785.2035	0.2458	1,172.9367	7,238.2499	0.2702

Determination of the best algorithm for clustering aquaculture fisheries companies is shown by the smallest ratio between the standard deviation within the cluster (S_W) and the standard deviation between clusters (S_B). Overall, the average of the ratio between standard deviation within the cluster (S_W) and standard deviation between clusters (S_B) that generated using the k-prototypes algorithm is smaller than the TSC algorithm. Therefore, the best aquaculture fisheries companies clustering algorithm is using the k-prototypes algorithm with an optimal number of clusters of 6 clusters. The distribution of cluster members per province of the k-prototypes algorithm can be seen in the Table 7.

Table 7: The distribution of cluster members per province of the k-prototypes algorithm

Province	Cluster						Total
	1	2	3	4	5	6	
North Sumatera	1	1	-	5	1	1	9
Lampung	1	3	1	11	6	2	24
Bangka Belitung Islands	-	2	-	1	-	-	3
Riau Islands	-	-	-	1	-	-	1
DKI Jakarta	-	1	1	-	-	7	9
West Java	-	1	1	3	1	2	8
Central Java	-	1	-	2	2	1	6
East Java	1	29	1	29	3	29	92
Banten	-	1	-	1	-	6	8
Bali	-	4	2	2	1	5	14
West Nusa Tenggara	1	6	2	13	-	4	26
East Nusa Tenggara	1	1	4	2	-	2	10
West Kalimantan	-	4	-	3	-	-	7
Central Kalimantan	-	-	-	-	-	1	1
East Kalimantan	-	-	-	-	-	1	1
North Kalimantan	-	1	-	-	-	6	7
South Sulawesi	1	6	-	1	-	9	17
Southeast Sulawesi	-	1	1	-	-	1	3
Gorontalo	-	-	1	-	-	-	1
Maluku	1	3	1	1	-	-	6
North Maluku	-	-	2	-	-	-	2
West Papua	1	-	2	-	-	-	3
Indonesia	8	65	19	75	14	77	258

3.6. Characteristics of The Clustering Result Based on The Categorical Variable

The relationship between categorical variables can be seen through the association test that using the chi-square test. The chi-square test shows that all categorical variables that used in this study have an association with the best clustering results that have been obtained (Table 8). Through X1 variable, we can see the distribution of

aquaculture fisheries company's cluster members based on the form of business entity. Aquaculture fisheries companies with the form of PT dominate in all clusters formed, especially in cluster 4. The form of PN/PD/Persero/Perum, CV, and Firm are mostly found in cluster 6. Meanwhile for the aquaculture fisheries companies with the form of cooperative only formed in clusters 2 and 3 with 1 member of each cluster. Then, based on their capital status (X2), cluster 1 and 3 are dominated by aquaculture fisheries companies with foreign capital. Meanwhile, clusters 4, 5, and 6 are dominated by aquaculture fisheries companies with domestic capital. Furthermore, for cluster 2, the most cluster members are filled by aquaculture fisheries companies with other sources of capital. The distribution of cluster members based on company status (X3) showed that clusters 1 and 3 are dominated by aquaculture fisheries companies with company status of branchless. Meanwhile, clusters 4, 5, and 6 are dominated by central aquaculture fisheries companies. Furthermore, for cluster 2, the mostly members is filled by aquaculture fisheries companies with a branch company status. The further information about characteristics of the clustering result based on the categorical variable can be seen in the section 3.8.

Table 8: Chi-square test between categorical variables and the clustering result of k-prototypes algorithm (k=6)

Variable	Chi-square	Degree of freedom	p-value
X1	38.511	20	0.038
X2	270.192	10	0.000
X3	270.192	10	0.000
X4	152.945	5	0.000
X5	143.362	10	0.000
X6	297.209	25	0.000
X7	54.902	10	0.000
X8	45.607	5	0.000

3.7. Characteristics of The Clustering Result Based on The Numerical Variable

Before analyzing the clusters formed based on their numerical variables, the assumption of homogeneity of variance is first tested on these variables. The test results show that the data is not homogeneous (Table 9), so it is necessary to do the Welch test as an alternative test for analysis of variance on the data. Based on the Welch test (Table 10), it was found that almost all numerical variables that used in this study were significantly different for the results of cluster formation. There is only one numerical variable (variable X11). In addition, there is one numerical variable (variable X16) which cannot produce the Welch test statistic value because there is one category that has zero variance.

Table 9: Homogeneity test of variance

Variable	Levene statistic	Degree of freedom 1	Degree of freedom 2	p-value
X9	41.174	5	252	0.000
X10	12.906	5	252	0.000
X11	10.439	5	252	0.000
X12	32.415	5	252	0.000
X13	25.832	5	252	0.000
X14	42.681	5	252	0.000
X15	18.554	5	252	0.000
X16	90.335	5	252	0.000
X17	60.705	5	252	0.000
X18	33.187	5	252	0.000
X19	60.897	5	252	0.000
X20	41.354	5	252	0.000
X21	24.143	5	252	0.000

Table 10: Welch test

Variable	Welch statistic	Degree of freedom 1	Degree of freedom 2	p-value
X9	18.892	5	40.082	0.000
X10	5.671	5	38.344	0.001
X11	1.351	5	47.122	0.260
X12	15.292	5	39.800	0.000
X13	3.889	5	39.372	0.006
X14	7.426	5	39.088	0.000
X15	2.901	5	39.462	0.025
X16	-	-	-	-
X17	9.222	5	39.928	0.000
X18	3.769	5	43.255	0.006
X19	13.274	5	39.236	0.000
X20	6.065	5	44.892	0.000
X21	16.991	5	39.999	0.000

The visualization of each cluster based on its numerical variables can be seen by radar chart (Figure 3). The data that used of making the radar charts is normalized data. This is done to solve the problem of the unit differences of each numerical variable, so that the data can be compared. Overall, cluster 1 and 5 have 6 numerical variables with the highest value compared to other clusters, then followed by cluster 3. Based on the radar chart, cluster 2, 4, and 6 have the smaller numerical variable than the other three clusters.

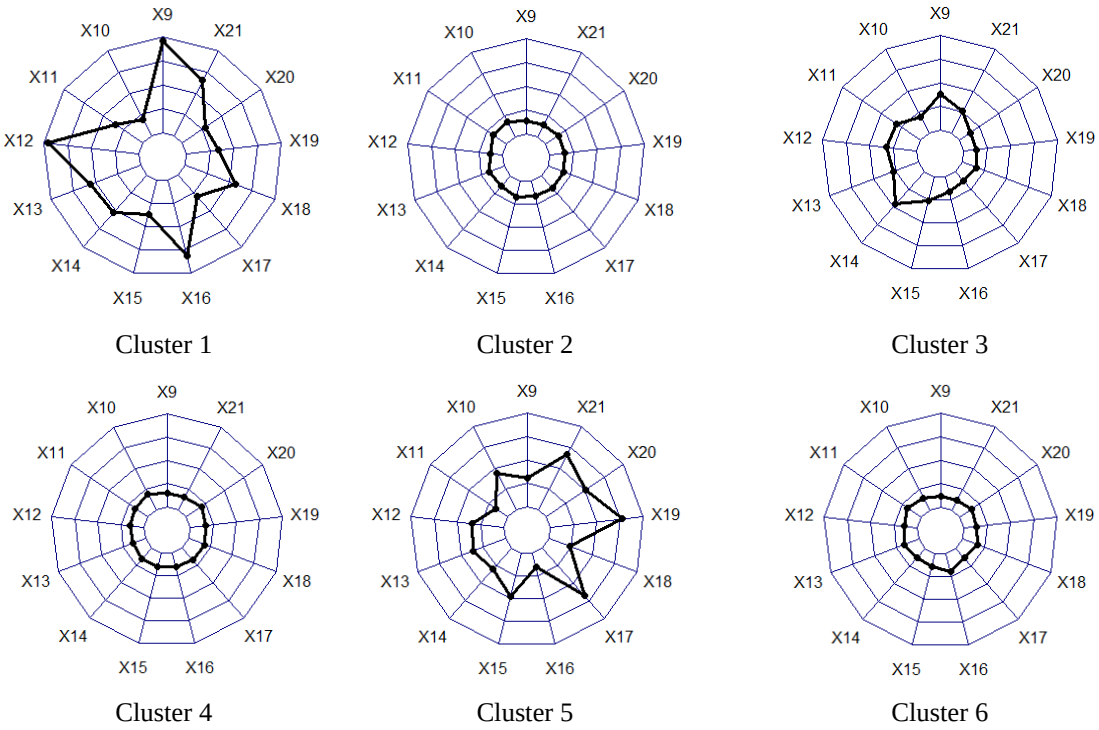

Figure 3: Radar chart of each cluster based on its numerical variables

3.8. The Summary of Cluster's Characteristics

Based on the characteristics that obtained from the categorical and numerical variables, and the distribution of its members, the summary of each cluster's characteristics that using the k-prototypes algorithm (k = 6) can be seen in Table 11. It also shows that the order of the clusters that related to the condition of the aquaculture fisheries companies starting from the best to the worst is starting from cluster 5, cluster 1, cluster 3, cluster 2, cluster 4, and cluster 6. This means that the aquaculture fisheries companies in cluster 6 need more attention, especially from the government, so that these companies can further improve the quality and quantity of their output. It also means that the aquaculture fisheries companies in cluster 5 can be used as role models for the best aquaculture fisheries companies in Indonesia.

Table 11: The summary of each cluster's characteristics

Category	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster members	8 companies in 8 provinces	65 companies in 16 provinces	19 companies in 12 provinces	75 companies in 14 provinces	14 companies in 6 provinces	77 companies in 14 provinces
Most distribution location	The amount is same for 8 provinces	East Java Province	East Nusa Tenggara Province	East Java and West Nusa Tenggara Province	Lampung	East Java and South Sulawesi Province
Form of business entity	PT	PT and Cooperative	PT and Cooperative	PT	PT	PT and CV
Company status	Branchless	Branch	Branchless	Central	Central	Central
Capital status	Foreign capital	Others	Foreign capital	Domestic capital	Domestic capital	Domestic capital
Type of activity	Enlargement	Enlargement	Enlargement	Enlargement	Enlargement	Hatchery
The main types of aquaculture	Land water	Sea water	Land water	Sea water	Sea water	Land and brackish water
The main of aquaculture container	Others	Fishpond	Span rope	Fishpond	Fishpond	Water tub
Aquaculture technology	Intensive	Simple	Simple and intensive	Intensive	Intensive	Semi intensive and intensive
The dominant numerical variables	X9, X11, X12, X13, X14, X16, and X18	Fairly low compared to clusters 1, 3, and 5 but on the average is still above clusters 4 and 6	X9 and X14	Fairly low compared to clusters 1, 2, 3, and 5 but on the average is still above cluster 6	X10, X15, X17, X19, X20, and X21	Have the lowest value compared to other clusters
Position compared to other clusters	2nd position	4th position	3rd position	5th position	1st position	6th position

4. Conclusion

On the basic findings of the data analysis, it can be concluded as follows:

1. Based on the comparison of the ratio between the standard deviation within cluster (S_W) and the standard deviation between clusters (S_B) from the k-prototypes algorithm (k = 6) and the TSC algorithm (k = 5), it can be concluded that the k-prototypes algorithm (k = 6) is the best algorithm for clustering aquaculture fisheries companies in Indonesia. This is because the k-prototypes algorithm (k = 6) has the smallest average ratio value and it indicates that the diversity within clusters is quite homogeneous, while the diversity between clusters is increasingly heterogeneous.
2. Cluster 1 has 8 members, cluster 2 has 65 members, cluster 3 has 19 members, cluster 4 has 75 members, cluster 5 has 14 members, and cluster 6 has the most members (77 members).
3. Furthermore, based on the characteristics of each cluster, then the order of the clusters that related to the condition of the aquaculture fisheries companies starting from the best to the worst is starting from cluster 5, cluster 1, cluster 3, cluster 2, cluster 4, and cluster 6.

5. Suggestions

1. This study only involves two clustering algorithms, so it would be better if other algorithms were added as a comparison in clustering the aquaculture fisheries companies in Indonesia.
2. Suggestions for the government also with this study are to pay more attention to the condition of aquaculture fisheries companies in cluster 6 and to provide the right policies, so that these companies can further increase their productivity. The efforts that can be made include providing guidance to related businesses, providing loan assistance programs with low interest rates, simplifying the system for extending business licenses, etc.

Acknowledgement

We would like to extend our gratitude to IPB University as a place to carry out this study and also to Statistics Indonesia (BPS) for the data that we used.

References

[1]. Food and Agriculture Organization of The United Nations. The State of World Fisheries and Aquaculture Sustainability in Action. Rome: Food and Agriculture Organization of The United Nations, 2020.
[2]. Statistics Indonesia. Economic Indicators for July 2020. Jakarta: Statistics Indonesia, 2020.
[3]. The Ministry of Marine Affairs and Fisheries of Indonesia. Marine and Fisheries Figures in 2018. Jakarta: The Ministry of Marine Affairs and Fisheries of Indonesia, 2018.
[4]. Statistics Indonesia. Statistics of Fishery Establishment 2018. Jakarta: Statistics Indonesia, 2019.
[5]. G. Gan, C. Ma, and J. Wu. Data Clustering Theory, Algorithms, and Applications. Virginia: American Statistical Association (ASA), 2007.

[6]. J. Macqueen. “Some methods for classification and analysis of multivariate observations”. In Proceedings of The 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.

[7]. Z. Huang. “Clustering large data sets with mixed numeric and categorical values”. In Proceeding of the First Pacific Asia Knowledge Discovery and Data Mining Conference, 1997, pp. 21–34.

[8]. T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris. “A robust and scalable clustering algorithm for mixed type attributes in large database environment”. In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, pp. 263-268.

[9]. A. Strehl and J. Ghosh. “A knowledge reuse framework for combining multiple partitions”. Journal of Machine Learning Research. vol. 3, pp. 583-617, 2002.

[10]. J.C. Gower. “A general coefficient of similarity and some of its properties”. International Biometric Society. vol. 27, pp. 857-871, 1971.

[11]. P.F. Lazarsfeld and N.W. Henry. Latent Structure Analysis. New York: Houghton Mifflin, 1968.

[12]. D.T. Pham, M.M.S. Alvarez, and Y.I. Prostov. “Random search with k-prototypes algorithm for clustering mixed datasets”. In Proceedings of The Royal Society A: Mathematical, Physical, and Engineering Sciences, 2011, pp. 2387-2403.

[13]. R. Nooraeni, J. Suprijadi, and Zulhanif. “K-prototype for clustering the mixed data type”. Journal of Theoretical and Applications Statistics: Biomedics, Industry, Business, and Social Statistics. vol. 13, pp. 9-16, 2019.

[14]. S.R. Ahire and L. Landge. “K-prototype clustering with efficient summarization for topic evolutionary tweet stream clustering”. International Journal of Science and Research (IJSR). vol. 6, pp. 769-774, 2015.

[15]. O. Pasin and H. Ankarah. “Comparison of EM and two step cluster method for mixed data: an application”. International Journal of Medical Science and Clinical Inventions. vol. 4, pp. 2768-2773, 2017.

[16]. M. Kayri. “Two step cluster analysis in researches: a case study”. Eurasian Journal of Educational Research (EJER). vol. 7, pp. 89-99, 2007.

[17]. A.D. Munthe, I.M. Sumertajaya, and U.D. Syafitri. “The clustering of villages and subdistricts based on poverty indicators by applying the TSC and k-prototypes algorithm”. Indonesian Journal of Statistics and Its Applications. vol. 2, pp. 63-76, 2018.

[18]. B.S. Everitt, S. Landau, M. Leese, and D. Stahl D. Cluster Analysis 5th Edition. London, UK: John Wiley and Sons Ltd, 2011.

[19]. J. Bacher, K. Wenzig, and M. Vogler. “SPSS Twostep cluster-a first evaluation”. Lehrstuhl fur Soziologie Arbeits- und Diskussionpapiere. vol. 2, pp. 1–20, 2004.