BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1617–1630
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1617-1630

CLUSTERING OF STATE UNIVERSITIES IN INDONESIA BASED ON PRODUCTIVITY OF SCIENTIFIC PUBLICATIONS USING K-MEANS AND K-MEDOIDS

Ermawati[1*], Idhia Sriliana[2], Riry Sriningsih[3]

[1]*Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Alauddin Makassar, Makassar, 92118, Indonesia*
[2]*Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Bengkulu Bengkulu, 38371, Indonesia.*
[3]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Padang Padang, 25131, Indonesia*

*Corresponding author's e-mail: *ermawati@uin-alauddin.ac.id*

Article History:
Received: 03rd May 2023
Revised: 8th August 2023
Accepted: 16th August 2023

Keywords:
Clustering;
K-Means;
K-Medoids;
Scientific publications

ABSTRACT

Scientific publication is a measure of the performance of a university. Universities that are owned and operated by the government and whose establishment is carried out by the President of Republic Indonesia are state universities (PTN). One of the efforts that can be made to determine the quantity and quality of state university scientific publications is to conduct PTN clustering based on the productivity of scientific publications. This clustering aims to see the position of state universities in Indonesia into 3 categories, namely "high", "medium", and "low". One of the clustering methods that can be used is cluster analysis. The cluster analysis used in this study is k-means and k-medoids with Silhoutte's validity. Based on the results of the analysis, it was found that the Silhouette k-means value (0.8018) was higher than the Silhouette k-medoids value (0.7281). Therefore, in this case, it can be concluded that the k-means method is better than the k-medoids. The results of cluster analysis using K-Means are 1) PTN with high productivity of scientific publications, namely ITB, ITS, UGM, and UI. The four PTNs are PTN as Legal Entity (PTN-BH) located in Java, 2) PTN with medium scientific publication productivity consists of 16 PTN which were dominated by PTN-BH and PTN as Public Service Board (PTN-BLU) with the largest location in Java, and 3) PTN with low scientific publication productivity consisted of 102 PTN which were dominated by PTN as general state financial management (PTN-Satker) with most locations outside Java.

How to cite this article:

Ermawati, I. Sriliana and R. Sriningsih., "CLUSTERING OF STATE UNIVERSITIES IN INDONESIA BASED ON PRODUCTIVITY OF SCIENTIFIC PUBLICATIONS USING K-MEANS AND K-MEDOIDS," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1617-1630, September, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

State Universities hereinafter referred to as "PTN" is a university that is owned and operated by the government and its establishment is carried out by the President of Republic Indonesia [1]. There are 122 PTNs in Indonesia located in Java and outside Java. The PTN consists of PTN with Legal Entity (PTN-BH), PTN as Public Service Board (PTN-BLU), and PTN as general state financial management (PTN-Satker) [2]. The difference between each PTN lies in the management of its finances and resources.

PTN-BH is the highest level because it has full autonomy in managing finances and resources, including lecturers and staff. This PTN operates similarly to a state-owned company (BUMN). PTN-BLU is an institution with a second level in terms of autonomy. The management of this institution is similar to a state-owned hospital. All non-tax revenues are managed autonomously and reported to the state. For PTN-Satker, all income, including student tuition fees, must be entered into a state account (Ministry of Finance) before being used [2].

As a university that is managed by the government, PTN is obliged to contribute to the productivity of Indonesian scientific publications. The productivity of scientific publications is one of the outputs that must be achieved by universities towards the World Class University and World Class Research Institution. The productivity of scientific publications is seen from the quantity and quality of scientific publications. The quantity is seen from the number of scientific publications on a national and international scale. Quality is seen from how much the publication is used by other academics by citing the writings produced. The more scientific publications in national and international reputable journals will be able to determine the position of the Higher Education (PT) nationally and internationally. This will encourage Indonesia to be able to compete with other countries.

The productivity of scientific publications can be seen from the performance of research conducted by lecturers. One of the data sources that can be used to see the productivity of lecturers' scientific publications is SINTA (Science and Technology Index). SINTA is the largest web-based index, citation, and expertise center in Indonesia that offers quick, easy, and comprehensive access to measure the performance of researchers and institutions based on publications produced and journal performance based on the number of articles and citations produced. SINTA has several features such as citation (Scopus and Google citations), research output (journals, articles, published books), and scores (Scopus H-Index and Sinta Score). The data on this feature can be used as a variable to see how the conditions of the productivity of higher education scientific publications in Indonesia.

Based on data processed from Scopus and Web of Science (Thompson) in 2010 until April 2016, the number of indexed Indonesian scientific publications is still low, even below Malaysia and Thailand [3]. Furthermore, from the initial exploration of SINTA data, there are still many Indonesian PTNs that have a few publications and low quality. Not only that, but there was also an imbalance in the number and quality of scientific publications between Indonesian PTNs located in Java and outside Java. The number of scientific publications produced by PTNs in Java is more than PTNs outside Java, whereas the number of PTNs located in Java is less than PTNs located outside Java. Data also found outliers caused by superior universities, where the number and quality of scientific publications differed significantly from other universities.

Therefore, the government through the ministry of research and higher education technology continues to encourage lecturers to increase the quantity and quality of their scientific publications. The target of scientific publications that is currently the focus of the government is scientific publications in Scopus or Thompson Reuter indexed journals or reputable international journals characterized by Q1, Q2, Q3, and Q4 as one of the output targets that must be achieved by PT, especially PTN [4] and [5].

To achieve this target, the government needs to cluster PTN based on the productivity of its scientific publications. Therefore, it requires a clustering analysis of PTN based on the quantity and quality of scientific publications. The aim is to determine which PTN is included in the high, medium, and low clusters based on the quantity and quality of scientific publications. Thus the government can make the right policies for each PTN cluster to increase the productivity of its scientific publications.

One of the statistical methods used to clustering objects is cluster analysis. Cluster analysis is one of the techniques in multivariate analysis. The main goal is to clustering objects of observation into several clusters based on their characteristics. In cluster analysis, objects that are most closely similar are in the same cluster and have similarities with one another [6].

Among the many existing clusters analyzes, two types of cluster analysis that have interrelated algorithms, namely K-Means and K-Medoids clustering. Both methods are partitioning clustering methods that can be used for large data sizes. K-Means is a distance-based clustering method that divides data into some clusters and this algorithm works only on numerical attributes. Several studies that discuss k-means include [7], [8], [9], [10], and [11]. Meanwhile, K-Medoids is an algorithm that is still related to the K-Means algorithm. K-Medoids is a general version of the K-Means algorithm which works by measuring distances and is more computationally intensive. The advantages of this method can overcome the weaknesses of the K-Means method which is sensitive to outliers and can reduce noise. However, the complexity of each iteration in the k-medoids algorithm is computationally more complicated, especially if the data set size is large. Several studies that have been conducted include [12], [13], [14], and [15].

Both methods are partitioning methods and attempt to minimize squared error, the distance between the labeled point in the cluster, and the point designated as the center of the cluster. The difference between K-Medoids and K-Means is K-Medoids choose the data point as the center [16]. Several studies that have been conducted to compare the performance of k-means and k-medoids include studies that clustered Betta fish into 3 categories, namely "color features", "shapes", and "textures". The results of this study indicate that clustering with the K-Medoids algorithm is more accurate with a faster running time value than the K-Means algorithm [16], another study discusses the performance of lecturers and clusters them into three categories, namely "satisfactory", "good" and "bad". The results obtained also show that the K-Medoids algorithm has better performance than the K-Means which is reviewed from the Davies Bouldin Index (DBI) score [17]. Furthermore, another study comparing the K-means and K-medoids algorithm were [18], [19], and [20]. The results obtained by this study concluded that the K-medoids algorithm is better than the k-means algorithm. Its ability to be better at scalability on a larger dataset and more efficient than K-means. K-Medoids are better in terms of execution time, non-sensitive to outliers and that reduce noise as compared to K-Means as it minimizes the sum of dissimilarities of datasets.

The differences between this study and previous studies are 1). Data and variables used. In this study, the data used are SINTA data with 9 numerical variables. Clustering is carried out at PTN based on the productivity of its scientific publications, 2). The clustering produced by the two methods is not only seen from its numerical variables but also based on the location and status of PTN, and 3). The comparison of the two methods is seen from the Silhouette value.

Therefore, this research will apply the k-means and k-medoids methods in clustering PTNs in Indonesia based on the productivity of their scientific publications. This clustering is useful to find out whether there is a significant difference in the productivity of scientific publications between PTN groups and to see the position of each PTN against other PTNs in the formed cluster. Furthermore, the comparison is carried out based on the validation of the cluster results to get the best method.

2. RESEARCH METHODS

2.1 K-Means Clustering

The K-Means algorithm is the most popular clustering algorithm and is widely used in the industrial world [21]. This algorithm is structured based on a simple idea, where the first step is to determine the number of clusters to be formed. The first element in the cluster can be selected to serve as the centroid point cluster. The K-Means algorithm will then repeat the following steps until stability occurs (no objects can be moved). According to Rencher, the K-means cluster method allows objects to move from one cluster to another, unlike the hierarchical method. The algorithm of this method is as follows [22].

1. Determine the number of clusters (k) and the centroid of each cluster
2. Calculate the distance between each object and each centroid
3. Grouping objects based on the closest distance to the centroid
4. Determine the new centroid (mean value) for the formed cluster
5. Repeating the second to fourth steps until there are no more moving objects between clusters.

The determination of the first centroid can use a wide variety of ways, such as determining randomly. Furthermore, the measure of similarity between objects used is a measure of distance. There are several types of distance measurements, including Euclidean, City-Block, Minkowsky, and Mahalanobis. Because the K-

Means method is not sensitive to the similarity between objects, the simplest and most frequently used type of measure is the Euclidean distance.

If there are two points, namely x and y points, then the distance between x and y is written as $d(x,y)$. The Euclidean distance formula can be written as follows:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \tag{1}$$

Two objects/clusters with the closest distance will form a new cluster, then the centroid is determined. The centroid is the average of all the members in the cluster. The centroid of the new cluster formed is obtained by the formula:

$$\bar{x} = \frac{n_A\bar{x}_A + n_B\bar{x}_B}{n_A + n_B} \tag{2}$$

where:

n_A is the number of objects in cluster A

n_B is the number of objects in cluster B.

Each object/cluster that has been combined in a cluster will be recalculated new centroid in the cluster, therefor any occurrence of additional members, centroid will also change. With this method, every time a new cluster occurs, the centroid will be recalculated until a fixed group is formed.

2.2 K-Medoids Clustering

K-Medoids is a non-hierarchical cluster method that is a variant of the K-Means method. The difference between K-medoids and K-means is K-Medoids uses the representative object (medoid) as the cluster centroid for each cluster, while K-Means uses the average value (mean) as the centroid of the cluster. The advantages of this method are the ability to overcome the weaknesses of the K-Means method which is sensitive to noise and outliers, where an object with a large value may substantially deviate from the data distribution [22].

The K-Medoids method is a method that uses a representative object called the medoid as the center point or centroid. The K-Medoids method partitions by minimizing the amount of dissimilarity between each ith object and the closest representative object. Each remaining object is grouped with the most similar representative object where the distance calculation is calculated by the distance among each data [20].

K-Medoids Clustering Algorithm

K-Medoids is an iterative clustering procedure with the following steps [23]:

1. Select K objects randomly from a set of N objects as representative objects (medoid).
2. Calculate the Euclidean distance between each object and each medoid by the following equation:

$$d\left(x_{ij}, o_{mj}\right) = \sqrt{\sum_{j=1}^{N}\left(x_{ij}, o_{mj}\right)^2}, \quad m = 1,2,...,K;\ j = 1,2,...,N$$

where

$d\left(x_{ij}, o_{mj}\right)$ = the distance between the ith object and the mth medoid in jth variable

x_{ij} = the ith object in jth variable

o_{mj} = the mth medoid in jth variable

K = the number of clusters

N = the number of objects

3. Each object is entered into the closest cluster based on Euclidian distance.
4. Select objects randomly as new medoids candidates in each cluster.

5. Each object in each cluster is calculated its distance from the new medoid.
6. Calculate the total deviation (S) by calculating the value of total new distance - total old distance. If S<0 is obtained, exchange the object with the data cluster to create a new set of K objects as medoids.
7. If at least one medoid has changed, return to step 4. Otherwise, end the algorithm.

2.3 Cluster Validation

One of the evaluation methods that can be used to see the strength and quality of a cluster is the silhouette method. This method uses an average value approach to estimate the quality of the clusters formed. The higher the average value is the better cluster [24]. The silhouette width (SW) formula for the *i*th data given by

$$sw_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \ i = 1, 2, \ldots, n \tag{4}$$

The SW value of a *k*th cluster is obtained by calculating the average sw_i of all data that joins the cluster using the equation:

$$SW(k) = \frac{1}{n_k} \sum_{x_i \in C_k} sw_i \tag{5}$$

Furthermore, the global SW is obtained by calculating the average $SW(k)$ for all clusters based on the following equation:

$$SW = \frac{\sum_{k=1}^{K} n_k SW(k)}{\sum_{k=1}^{K} n_k}, \ k = 1, 2, \ldots, K \tag{6}$$

where

a_i = the average distance between the *i*th data and all data in the same cluster
b_i = the average distance between the *i*th data and all data in a different cluster
sw_i = *Silhouette Width* for the *i*th data
$SW(k)$ = *Silhouette Width* for the *k*th cluster
SW = Global *Silhouette Width*
x_i = the *i*th data object
C_k = cluster k
n_k = the number of data in cluster k

3. RESULTS AND DISCUSSION

3.1 Characteristics PTN in Indonesia

State Universities in Indonesia based on their financial management can be divide into three, namely State Universities with Legal Entities (PTN-BH), State Universities with general state financial management (PTN-BLU), and State Universities with general state financial management (PTN-Satker). A comparison of the number of state universities with the status of PTN-BH, PTN-BLU, and PTN-Satker in Indonesia, can be seen in **Figure 1** below:

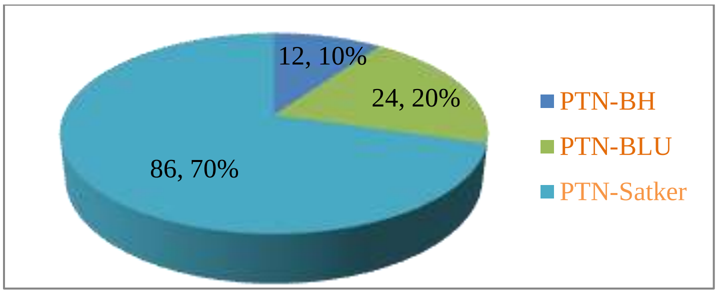
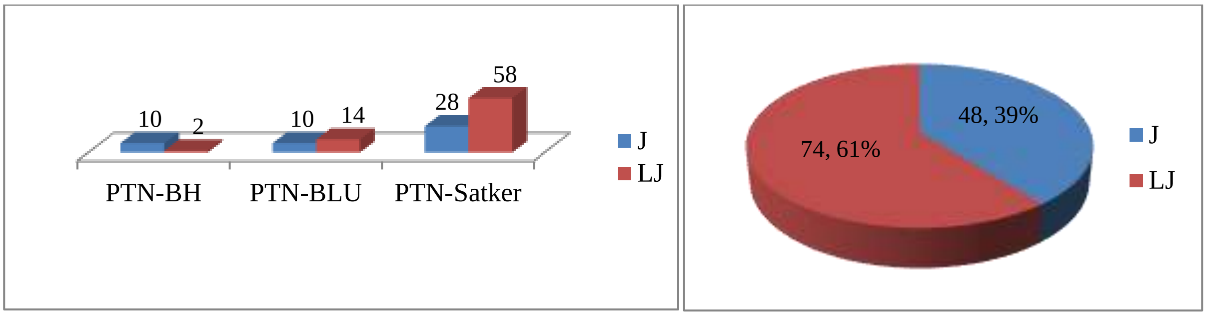


Figure 1. Percentage of Total State Universities based on PTN Status

Based on **Figure 1** above, it can be seen that of the 122 PTNs in Indonesia, the highest percentage is PTN-Satker, and the least is PTN-BH. Based on the location of higher education, it is divided into PTN in Java (J) and PTN outside Java (LJ). The number of universities in Java and outside Java can be seen in **Figure 2** below:



(a) (b)

Figure 2. The Number of PTN based on Location

Based on **Figure 2** (a) above, it can be seen that the number of PTNs based on the status of PTN both in Java and outside Java, where PTN-BH is dominant in Java, while PTN-BLU and PTN-Satker are predominantly outside Java. In **Figure 2** (b) it can be seen that the number of PTN outside Java is more than the PTN in Java, which is 74.61% as many as 84 PTN outside Java and there are 48.39% or 48 PTN in Java.

3.2 Correlation of Variables

In this case, there are 11 mixed variables (nine numeric variables and two categorical variables) used in the clustering of Higher Education, namely Scopus H-index $\left(X_3\right)$, Google H-index $\left(X_2\right)$, Sinta Score All $\left(X_3\right)$, Sinta Article $\left(X_4\right)$, Scopus Article $\left(X_5\right)$, Scopus Conference $\left(X_6\right)$, Scopus Other $\left(X_7\right)$, Scopus Citation $\left(X_8\right)$, Google Citation $\left(X_9\right)$, PTN Status $\left(X_{10}\right)$, and Location of PTN $\left(X_{11}\right)$. Therefor correlation testing is only carried out between numerical variables. The correlations between numerical variables $X_1, X_2, ..., X_9$ have presented in **Table 1**.

Table 1. Correlation Matrix between Numerical Variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1								
X_2	0.939	1							
X_3	0.994	0.944	1						
X_4	0.900	0.976	0.904	1					
X_5	0.981	0.927	0.984	0.879	1				
X_6	0.917	0.835	0.911	0.789	0.857	1			
X_7	0.878	0.775	0.888	0.711	0.843	0.878	1		
X_8	0.975	0.877	0.981	0.830	0.952	0.909	0.912	1	
X_9	0.908	0.930	0.936	0.910	0.902	0.830	0.815	0.901	1

If the correlation value between two variable close to 1 it's mean that they have strong relationship. The correlation matrix in Table 1 merely only shows the correlation between two variables but can not show the multivariate correlation of variables so correlation testing is needed using the Bartlett Test. The hypothesis testing for Bartlet test is

H_0: no correlation between variables

H_1: correlation between variables

Table 2. Barlett Test Spericity

	Value
Chi-square	3078.57
p-value	0.00
Degree of freedom	36

From Table 2 Bartlett Test spericity shows that p-value <0.000000 is less than 0.05 then H_0 is rejected so there is a multivariate correlation between variables. Because of the correlation between variables, an intermediate method (PCA) is needed before performing a cluster analysis. Principal component analysis (PCA) is a statistical technique by performing linear transformations in groups of origin variables (there is a correlation between groups of variables) into groups of variables that are not correlated and can represent a lot of information on groups of origin variables. The results of variable transformation with PCA are then used in cluster analysis.

Table 3. Percentage Cumulative Contribution of Variance in PCA

Eigen Value	Cumulative Percentage of Variance
8.1803	90.8924
0.4236	95.5995
0.1432	97.1908
0.1277	98.6102
0.0759	99.4538
0.0314	99.8023
0.0136	99.9534
0.0037	99.9943
0.0005	100.0000

Based on percentage cumulative contribution of variance in Table 3 we conclude to reduce nine variables into two dimensions because the contribution of cumulative percentage of variance is 95.5995% so that it is sufficient for further modelling.

3.3 Clustering of State Universities using K-Means

The cluster analysis process will be carried out by determining the number of clusters as many as 3, where PTN will be clustering into three categories of scientific publication productivity, namely PTN with High, Medium, and Low Productivity. Based on the results of data analysis, the number of PTNs in each cluster can be seen in Figure 3 below:

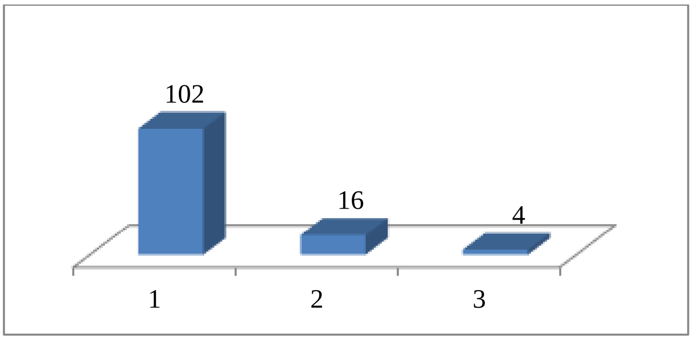


Figure 3. The Number of PTN in Each Cluster using K-Means

In Figure 3 Above it can be seen that there are 102 PTNs in cluster 1, 16 PTNs in cluster 2, and 4 PTNs in cluster 3. The results of clustering PTN using K-means cluster can be seen in Figure 4 below:

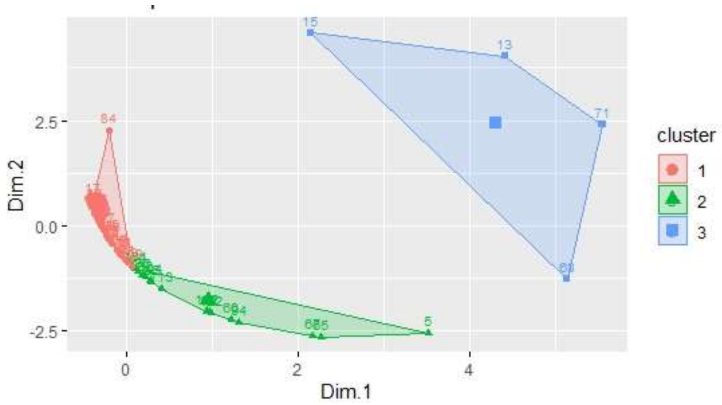


Figure 4. Plot Cluster

Based on Figure 4 above, it can be seen that there are PTNs in each cluster, where in cluster 3 there are 4 PTN members, namely ITB, ITS Surabaya, UGM, and UI. Cluster 2 consists of 16 PTN members consisting of IPB, Airlangga University, UNIBRAW, Diponegoro University, UNHAS, Jember University, Lampung University, Malang State University, Semarang State University, Yogyakarta State University, Padjadjaran University, Ganesha Education University, UPI, Sam University Ratulangi, University of North Sumatra, and Udayana University. The remaining 102 PTNs are included in cluster 1.

a. Characteristics of each cluster based on the numerical variables involved

The characteristics of each cluster can be seen from the mean or median value of each cluster. The characteristics of each cluster based on the mean of the numerical variables involved can be seen in Figure 5 below:

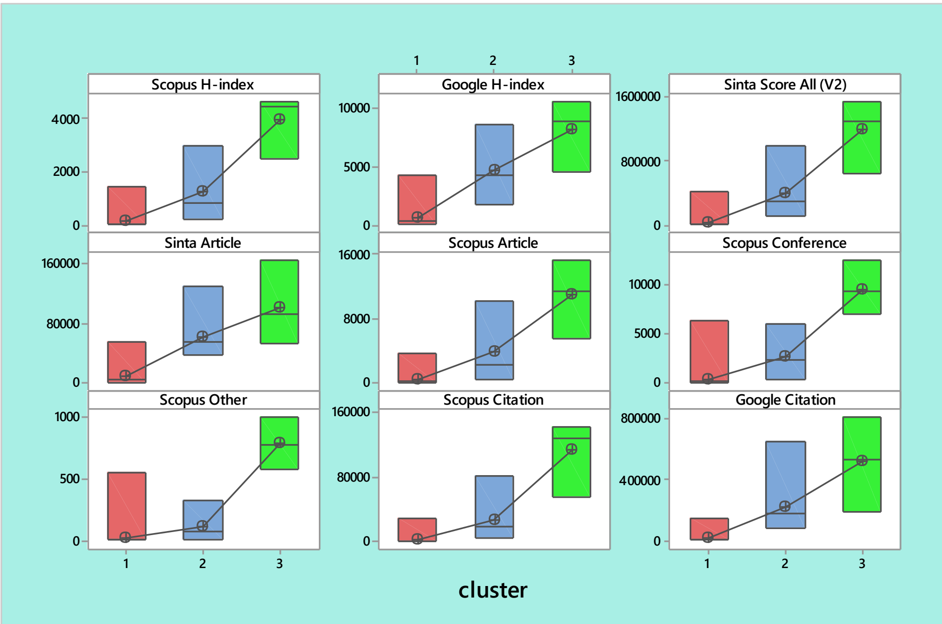


Figure 5. Characteristics Cluster Based on Numerical variabel

Based on Figure 5 above, the highest mean value for all variables is in cluster 3, therefor PTNs with high productivity of scientific publications are cluster 3 consisting of ITB, ITS, UGM, and UI. The second highest mean value of all variables is in cluster 2, so it can be said that cluster 2 is a PTN cluster with medium productivity of scientific publications consisting of 16 PTNs. While the lowest mean value for all variables is in cluster 1. The PTN with low productivity of scientific publications are in cluster 1 which consists of 102 PTNs.

b. *Characteristics of Each Cluster Based on Categorical Variables*

Category variables in this study involve two variables, namely the PTN status variable consisting of PTN-BH, PTN-BLU, and PTN-Satker. The second is a variable location that is divide into state universities in Java and outside Java. Based on the PTN status, the characteristics of each cluster can be known using coss tabulation as in Table 3 below:

Table 3. Cross Tabulation Based on PTN Status

Cluster	PTN Status						Total/ Percentage
	PTN-BLU		PTN-BH		PTN-Satker		
1	17	70,83%	1	8,33%	84	97,67%	102
	16.67%		0.98%		82.35%		100%
2	7	29,17%	7	58,33%	2	2,33%	16
	43.75%		43.7%		12.5%		100%
3	0	0%	4	33,34%	0	0%	4
	0%		100%		0%		100%
Total/ Percentage	24	100%	12	100%	86	100%	

where

Percentage based on PTN Status

Percentage based on cluster

Based on Table 3 above, cluster 1 is dominated by PTN-Satker, cluster 2 is dominated by PTN-BLU and PTN-BH, while cluster 3 is entirely PTN-BH. Based on the status of PTN, PTN-BH status is still dominant in cluster 2 (PTN with medium scientific publication productivity), which is 58.33%. However, cluster 1 is only occupied by PTN-BH as much as 33.4% of all PTN-BH. The status of PTN-BLU is still dominant in cluster 2 (PTN with low scientific publication productivity), the rest are in cluster 2 (PTN with moderate scientific publication productivity). The status of PTN-Satker is still dominant in cluster 3 (PTN with low scientific publication productivity) as much as 97.67% of PTN-Satker.

The second categorical variable is the location of PTN were consists of PTN in Java and PTN outside Java. By using cross-tabulation, characteristics of the cluster can be seen in Table 4, below:

Table 4. Cross Tabulation Based on Location of PTN

Cluster	Location of PTN				Total
	Java		Outside Jawa		
1	34	70,84%	68	91,89%	102
	33.33%		66.67%		100%
2	10	20,83%	6	8,11%	16
	62.5%		37.5%		100%
3	4	8,33%	0	0%	4
	100%		0%		100%
Total	48	100%	74	100%	

where

Percentage based on Location of PTN

Percentage based on cluster

In Table 4 above, it can be seen that cluster 1 is dominated by PTN outside Java, cluster 2 is dominated by PTN in Java, while cluster 3 consists of PTN in Java. If reviewed based on the location of PTN, PTN in Java is still more productive than PTN outside Java, because the percentage of PTN in Java for clusters 2 and 3 is still higher than PTN outside Java.

3.4 Clustering of State Universities using K-Medoids

The clustering process using K-Medoids method begins by determining the optimal number of clusters to be formed. The optimal number of clusters selection is carried out by looking at the graph of the average Silhouette Width (SW). The Silhouette method uses an average value approach to estimate the optimal K from the formed clusters. The higher the average value, the better the cluster.

After calculating the average SW value for clusters ranging from 2 to 10 with the K-Medoids algorithm, it can be seen in **Figure 6** that 4 clusters yield the highest value of 0.7404. Hence the optimal number of clusters is obtained by $K = 4$. However, $K = 3$ can also be selected as the number of other optimal clusters because the average SW value of $K = 3$ that is not much different from K = 4, which is 0.7281. Moreover, $K = 3$ and K = 4 have the highest average value compared to the others.

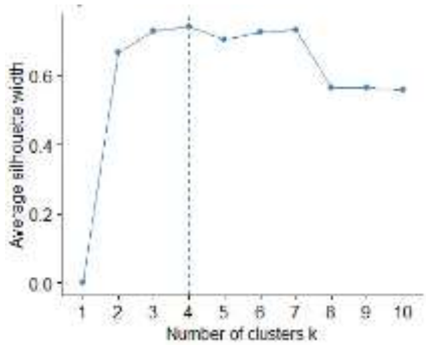


Figure 6. Optimal Number of Clusters using K-Medoids

Based on the purpose of this study to cluster State Universities into three categories of scientific publication productivity, then $K = 3$ is selected as the optimal number of clusters. Hereinafter, the clustering of the State Universities in Indonesia is carried out into three categories by using K-Medoids Algorithm.

The application of the K-Medoids algorithm begins with the determination of three medoids for the 3 targeted clusters. Medoid values as obtained are shown in **Table 5** with different medoid values from the centroid determination results.

Table 5. Medoid for *K*-th Cluster

Cluster	**1**	**2**	**3**
Medoid	56	90	13

From **Table 5**, it can be seen that the medoid or object that represents the centroid in cluster 1 is the 56th object, while the medoid in cluster 2 and cluster 3 are the 90th and the 13th object, respectively. Thus, the object which has the closest distance to the medoid of a cluster will enter the cluster. The number of members in each cluster can be seen in **Figure 7**.

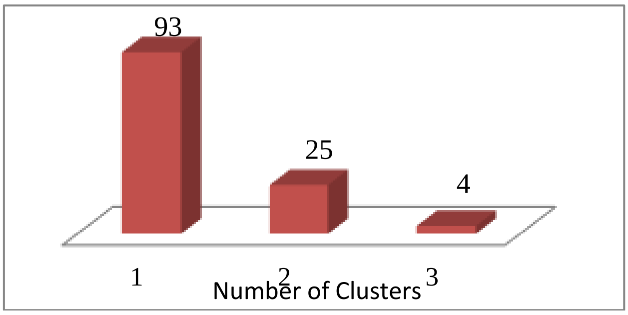


Figure 7. Number of Members in Each Cluster using K-Medoids

Figure 7 shows that clustering with the K-Medoids method produces a number of state university (hereinafter referred to as “PTN”) objects that are somewhat different for each cluster. Cluster 1 has 93 PTN members. Cluster 2 getting 25 PTN members. Whereas Cluster 3 gets 4 PTN members. The clustering results also can be seen visually in **Figure 8** below:

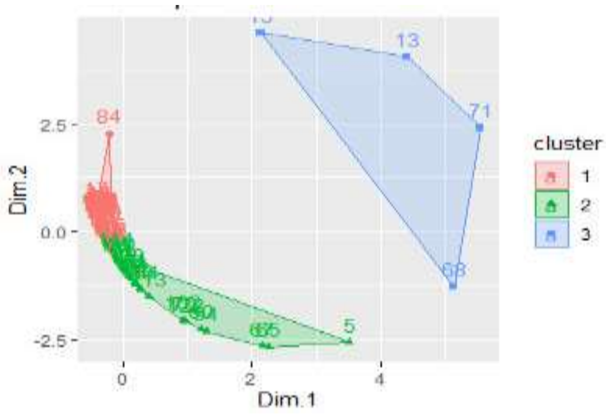


Figure 8. Cluster Plot K-Medoids

Based on **Figure 8** it can be seen that the state university membership in each cluster, where in cluster 3 there are 4 PTN members, namely Bandung Insitute of Technology (ITB), Sepuluh Nopember Institute of Technology (ITS), Gadjah Mada University (UGM), and Indonesia University (UI). Cluster 2 has 25 PTN members consisting of Bogor Agricultural University, Airlangga University, Andalas University, Bengkulu University, Brawijaya University, Diponegoro University, Hasanuddin University, Jember University, Jenderal Soedirman University, Lambung Mangkurat University, Lampung University, Malang State University, Padang State University, Semarang State University, Surabaya State University, Yogyakarta State University, Padjadjaran University, Ganesha University of Education, Indonesian Education University, Riau University, Sam Ratulangi University, Sriwijaya University, North Sumatra University, Syiah Kuala University, and Udayana University. Meanwhile, 93 other state universities are members of Cluster 1.

Furthermore, the interpretation of the clustering results using K-Medoids is carried out by looking at the characteristics of each cluster formed. In this study, the characteristics of each cluster are carried out based on the numerical variables involved in the clustering process and the categorical variables observed in the clustering objects.

a. Characteristics of Each Cluster Based on The Numeric Variables involved

The interpretation of the K-Medoids clustering results can be seen from the characteristics of each cluster based on the mean of each cluster. According to **[20]**, interpretation is carried out by calculating the average cluster results so that it can provide a logical description of the cluster results. The average calculation is obtained by using the original data. Therefore, based on the numerical variables used in the clustering of State Universities, the characteristics of each cluster can be seen in **Figure 9**.

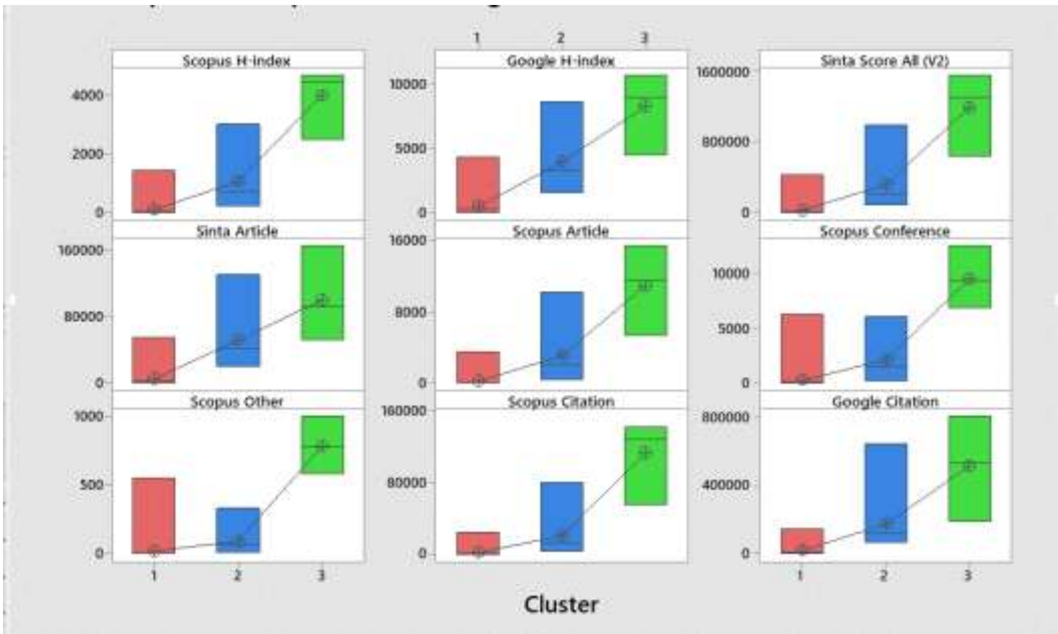


Figure 9. Characteristics of K-Medoids Clusters Based on Numerical Variables

The Boxplot chart in **Figure 9** shows that based on the numerical variables involved, cluster 3 has the highest mean value. In order to, it can be implied that the PTN group with high productivity of scientific publications is in cluster 3, there are ITB, ITS, UGM, and UI. The second highest mean value is in cluster 2, so that cluster 2 is a group of PTN that have medium productivity of scientific publications which consists of 25 members. While the lowest mean value for each variable is in cluster 1 which implies that the PTN group with low productivity is in cluster 1 which consists of 93 members.

b. Characteristics of Each Cluster Based on The Categorical Variables

In this study, there are two categorical variables observed in the clustering of PTNs. The variables are PTN Status and PTN Location. The PTN status variable has 3 categories, namely PTN-BLU, PTN-BH, and PTN-Satker. Meanwhile, the PTN location variable is divided into PTN in Java and outside Java. Based on the PTN status, the characteristics of each cluster can be determined by using cross-tabulation as in **Table 6** below:

Table 6. Cross-Tabulation K-Medoids Clusters Based on PTN Status

Cluster	PTN Status						Total
	PTN-BLU		PTN-BH		PTN-Satker		
1	10	41.67%	1	8.33%	82	95.35%	93
	10.75%		1.08%		88.17%		100%
2	14	58.33%	7	58.33%	4	4.65%	25
	56%		28%		16%		100%
3	0	0%	4	33.33%	0	0%	4
	0%		100%		0%		100%
Total	24	100%	12	100%	86	100%	122

Note:
(pink) percentage value based on PTN status
(green) percentage value based on cluster

Based on **Table 6**, it can be seen that cluster 1 is dominated by PTN-Satker. Meanwhile in cluster 2, more than half of the members are PTN-BLU. Whereas in cluster 3, all members are PTN-BH. In this case, it can be interpreted that only PTN-BH has high productivity of scientific publications, followed by the majority of PTN-BLU which have medium productivity, and almost all PTN-Satker has low productivity.

Apart from the PTN status, the characteristics of each cluster can be reviewed based on the location of PTN, namely PTN in Java and PTN outside Java. By using cross-tabulation, the cluster characteristics can be seen in **Table 7** below:

Table 7. Cross-Tabulation K-Medoids Clusters Based on PTN Location

Cluster	PTN Location				Total
	Java		Outside Java		
1	32	66.67%	61	82.43%	102
	34.41%		65.59%		100%
2	12	25.00%	13	17.57%	25
	48%		52%		100%
3	4	8.33%	0	0%	4
	100%		0%		100%
Total	48	100%	74	100%	122

Note:
(pink) percentage value based on PTN location
(green) percentage value based on cluster

From Table 7, it can be seen that cluster 1 is dominated by PTN located outside Java. In cluster 2, the number of PTN members located in Java and outside Java is almost equal, while all PTN members in cluster 3 are located in Java. Based on this, it can be concluded that only PTN in Java have high productivity of scientific publications, while most PTN outside Java have low productivity.

3.5 Comparison of Clustering Results

The comparison of the clustering results from the two methods based on the global SW value and the optimal number of clusters can be seen in Table 8 berikut:

Table 8. Cluster Validation Results

Method	SW	*K*
K-means	0.8018	3
K-Medoids	0.7281	3

From Table 8, it can be seen that the SW value for the validation of the higher education clustering in Indonesia based on the scientific publication productivity with K-Means method is greater than K-Medoids Method. By using the same number of clusters in this case, it can be said that the K-means Method gives better clustering results, although both methods structurally produce excellent strong structures with the same number of clusters.

4. CONCLUSIONS

This study has provided two cluster methods, namely K-Means and K-Medoids, that applied to clustering of universities based on the productivity of scientific publications. The conclusions of this study are:

1. Based on cluster validation using the Silhouette value, these two methods provide the same optimum number of clusters, but the Silhouette value is different. The Silhouette value in the K-means is higher than K-Medoids cluster. Therefor the K-Means cluster gives better clustering results.
2. Based on the results of clustering of PTN in Indonesia using the K-Means method, there are 4PTN-BH located in Java with high productivity of scientific publications, namely ITB, ITS, UGM, and UI. PTN with medium scientific publication productivity consists of 16 PTNs which were dominated by PTN-BH and PTN-BLU positioned in Java. They are IPB, Airlangga University, UNIBRAW, Diponegoro University, UNHAS, Jember University, Lampung University, Malang State University, Semarang State University, Yogyakarta State University, Padjadjaran University, Ganesha University of Education, UPI, Sam Ratulangi University, North Sumatra University, and Udayana University. PTN with low scientific publication productivity consisted of 102 PTNs which were dominated by PTN-Satker with most locations outside Java.
3. Based on the clustering results using K-Medoids, there are four PTN-BH in Java that have high productivity of scientific publications, namely ITB, ITS, UGM, and UI. The PTN group with medium productivity consists of 25 members, most of them are PTN-BLU in Java and outside Java. Meanwhile, the PTN group with low productivity has 93 members, which are dominated by PTN-Satker outside Java.

REFERENCES

[1] *Undang-undang RI nomor 22 Tahun 1961*, vol. 62, no. clase VIII. 1961.
[2] Sevima, “2 Pengertian & Perbedaan PTN-BH, PTN-BLU, PTN-Satker - SEVIMA,” 2019. https://sevima.com/pengertian-perbedaan-ptn-bh-ptn-blu-ptn-satker/
[3] Lukman, Yuniasih, I. Maryati, M. A. Silalahi, and A. Sihombing, *Kekuatan 50 institusi ilmiah Indonesia: Profil publikasi ilmiah terindeks Scopus*. Indonesia, 2016. [Online]. Available: https://medium.com/@arifwicaksanaa/pengertian-use-case-a7e576e1b6bf
[4] Lukman, “Strategi Peningkatan Publikasi Ilmiah Bereputasi,” *Surabaya PDII LIPI Jakarta.*, no. September, pp. 1–47, 2016.
[5] B. A. Pribadi and R. Delfy, “Implementasi Strategi Peta Konsep (Concept Mapping) Dalam Program Tutorial Teknik Penulisan Artikel Ilmiah Bagi Guru,” *J. Pendidik. Terbuka Dan Jarak Jauh*, vol. 16, no. 2, pp. 76–88, 2015, [Online]. Available: http://jurnal.ut.ac.id/index.php/jptjj/article/view/316
[6] R. A. Johnson and D. W. Wichern, “Applied Multivariate Staistical Analysis,” *Applied Multivariate Statistical Analysis*. pp. 671–757, 2007.

[7] D. Triyansyah and D. Fitrianah, “Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing,” *J. Telekomun. dan Komput.*, vol. 8, no. 3, p. 163, 2018, doi: 10.22441/incomtech.v8i3.4174.

[8] J. O. Ong, “Implementasi Algotritma K-means clustering untuk menentukan strategi marketing president university,” *J. Ilm. Tek. Ind.*, vol. vol.12, no, no. juni, pp. 10–20, 2013.

[9] F. Nasari and C. J. M. Sianturi, “Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat,” *CogITo Smart J.*, vol. 2, no. 2, pp. 108–119, 2016, doi: 10.31154/cogito.v2i2.19.108-119.

[10] S. Budiman, D. Safitri, and D. Ispriyanti, “Perbandingan Metode K-Means Dan Metode Dbscan Pada Pengelompokan Rumah Kost Mahasiswa Di Kelurahan Tembalang Semarang,” *J. Gaussian*, vol. 5, no. 4, pp. 757–762, 2016.

[11] L. Bai, J. Liang, and F. Cao, “A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters,” *Inf. Fusion*, vol. 61, pp. 36–47, 2020, doi: 10.1016/j.inffus.2020.03.009.

[12] R. Liu, “Data Analysis of Educational Evaluation Using K-Means Clustering Method,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3762431.

[13] Ghufron, B. Surarso, and R. Gernowo, “Implementation of K- Medoids Clustering for High Education Accreditation Data,” *J. Ilm. KURSOR*, vol. 10, no. 3, pp. 119–128, 2020.

[14] S. Sindi, W. R. O. Ningse, I. A. Sihombing, F. I. R.H.Zer, and D. Hartama, “Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia,” *J. Teknol. Inf.*, vol. 4, no. 1, pp. 166–173, 2020, doi: 10.36294/jurti.v4i1.1296.

[15] I. Makruf, Lubna, Khasanah, R. Sulaeman, and D. A. Harahap, “Educational data mining on higher education level education costs using clustering techniques in Indonesia,” *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. 6, pp. 2084–2089, 2020, doi: 10.5373/JARDCS/V12I6/S20201169.

[16] R. Yusuf and D. Novian, “Aplikasi Image Retrieval Pada Varian Ikan Cupang Menggunakan K-Means dan K-Medoids Agoritm,” *Jurnal Teknik*, vol. 12, no. 2. 2014.

[17] M. Milkhatun, A. A. F. Rizal, N. W. W. Asthiningsih, and A. J. Latipah, “Performance Assessment of University Lecturers: A Data Mining Approach,” *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 6, no. 2, pp. 73–81, 2020, doi: 10.23917/khif.v6i2.9069.

[18] S. Gultom, S. Sriadhi, M. Martiano, and J. Simarmata, “Comparison analysis of K-Means and K-Medoid with Ecluidience Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012092.

[19] E. Herman, K. E. Zsido, and V. Fenyves, “Cluster Analysis with K-Mean versus K-Medoid in Financial Performance Evaluation,” *Appl. Sci.*, vol. 12, no. 16, 2022, doi: 10.3390/app12167985.

[20] P. Arora, Deepali, and S. Varshney, “Analysis of K-Means and K-Medoids Algorithm for Big Data,” *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016, doi: 10.1016/j.procs.2016.02.095.

[21] Z. Karaca, “the Cluster Analysis in the Manufacturing Industry With K-Means Method: an Application for Turkey,” *Eurasian J. Econ. Financ.*, vol. 6, no. 3, pp. 1–12, 2018, doi: 10.15604/ejef.2018.06.03.001.

[22] R. Mickey and A. C. Rencher, *Methods of Multivariate Analysis*, vol. 93, no. 443. 1998. doi: 10.2307/2669873.

[23] Athifaturrofifah, R. Goejantoro, and D. Yuniarti, “Perbandingan Pengelompokan K-Means dan K-Medoids Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Studi Kasus : Data Titik Panas Di Indonesia Pada 28 April 2018),” *J. EKSPONENSIAL*, vol. 10, no. 2, pp. 143–152, 2019.

[24] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Canada: A John Wiley & Sons, Inc., Publication, 2005.