Comparative Analysis Clustering Algorithm for Government's Budget Performance Data

Isnen Hadi Al Ghozali *[1], Ibnu Afan[2], Triardani Lestari[3]
[1,2]Magister Ilmu Komputer, Universitas Budi Luhur, Jakarta, Indonesia
[3]Directorate General of Budget, Ministry of Finance
e-mail: *[1] **2111601163@student.budiluhur.ac.id** , [2]2111601510@student.budiluhur.ac.id , [3]triardani@kemenkeu.go.id

Abstract

The government's budget performance is a benchmark for the government's success in optimizing people's money to achieve national goals. Even though performance measurement has reached the Work Unit level, the data formed still do not have a specific grouping, in the sense of unstructured data. The purpose of this research is to find the best clustering algorithm for classifying budget performance data. The data used is budget performance data for 19,460 Indonesian Government Work Units. The data is sourced from the SMART application and the OM SPAN application. This research uses a comparative study approach for the K-Means algorithm, DBSCAN, and agglomerative hierarchical clustering (AHC). Evaluation of the clustering results formed using the Davies-Bouldin Index (DBI) method. The AHC algorithm with k = 6 achieved the lowest DBI value of 0.3583472. The DBI value for the DBSCAN algorithm with MinPts = 10 is 0.5398259. However, the AHC algorithm is not good in terms of ease of implementation. Therefore, the K-means algorithm with parameters k = 10 is the best alternative. The K-Means algorithm gets a DBI value of 1.052678. The K-Means algorithm produces 10 clusters. Based on knowledge extraction, it is determined that cluster 2 and cluster 5 are ideal clusters in terms of budget performance. While the clusters that require attention are cluster 1, cluster 3, cluster 4, and cluster 8.

Keywords— k-means, DBSCAN, AHC, clustering, budget performance

1. INTRODUCTION

Government budget performance refers to the evaluation of the utilization of ministry or agency budgets as recorded in budget documents. The government's budget achievements serve as a yardstick for evaluating the government's ability to efficiently utilize public funds to accomplish national objectives. Spending performance measurements serve the purpose of retrospective analysis as well as future forecasting. Retrospectively, the government archives and preserves historical data regarding past activity. Evaluating historical budget performance can serve as the foundation for future policy implementation.

Budget performance is a measure of the effectiveness of the government's fiscal policies [13]. To obtain a credible measurement, it is necessary to pay attention to the characteristics of government organizations. In 2021, the government of Indonesia had 19,460 Work Units carrying out government tasks. The average national spending performance score for the Work Unit level reached 87.40 and was categorized as "Good". This result is slightly lower than the average expenditure performance score at the Ministry/Agency level, which reached 92.34 and is categorized as "Excellent".

Even though performance measurement has reached the Work Unit level, the data formed still does not have a specific grouping, in the sense of unstructured data. This will certainly make it difficult for regulators to adopt fiscal policies that are following their characteristics so that

performance achievement remains at an optimum level. It is at this point that data science is needed. There is a clustering algorithm in data science that groups data based on cluster structure into data sets with the greatest similarities in the same cluster and the greatest differences in different clusters [26]. Theoretically, clustering algorithms are divided into centroid-based clustering, density-based clustering, distribution-based clustering, and hierarchical clustering.

For the centroid-based clustering category, the K-means algorithm is a popular algorithm. The K-means algorithm groups N data points into k clusters by minimizing the sum of the squares of the distance between each point and the centroid (mean of the nearest cluster) [24]. Determining the value of k becomes crucial in this algorithm. Several previous studies have suggested improving the K-means algorithm with attribute reduction, a better initialization technique [24], the canopy algorithm [25], k-means [26], ball k-means [27], and firefly algorithms [28].

Density-based clustering organizes data based on the density of points in the data space, rather than just areas of the same density. However, this algorithm has trouble with data that has different densities and high dimensions. The DBSCAN algorithm is an alternative that is often used [15]. The advantage of this algorithm is its ability to detect outliers [17]. Previous fields of study that used this algorithm include inductive technology [11], urban rail passenger aggregation distribution [12], and crowdsourcing logistics pricing [14]. DBSCAN enhancement was carried out with neighbor similarity, a fast nearest neighbor query [15], and network space [16].

Hierarchical clustering is a mathematical model or exploratory tool to demonstrate categorizing large volumes of different groups or tree form data sets based on similarities without prior knowledge [3]. Hierarchical clustering is divided into two groups, agglomerative (AHC) and divisive hierarchical clustering (DHC) [6]. AHC is an algorithm that has been developed in many previous studies in the areas of hotspot clustering [1], student activity [2], and judicial practice [4]. Several recent studies emphasize the development of more efficient algorithms [5] [6] [7].

This research focuses on using unsupervised learning to get the best grouping of budget performance measurement data that doesn't yet have cluster data. Previous studies have generally focused on using only one clustering algorithm in handling data. For example, research [18] uses K-Means to classify GRDP Growth Rate data, research [12] uses DBSCAN for urban rail passenger aggregation distribution, and research [2] uses AHC to categorize learning activities in online learning. The data studied was generally observational data in the fields of education, health, and law. This study investigates budget performance data, which has previously been reported only within a limited scope.

2. RESEARCH METHOD

The data used for the clustering analysis is the Government of Indonesia's budget data at the Work Unit level in 2021. The data used is secondary data owned by the Ministry of Finance in the SMART application. The raw set of data used is 19,460 observations with 10 attributes, namely Work Unit code (kd_ori_wu), Work Unit location (loc_wu), personnel expenditure budget (b51_wu), goods expenditure budget (b52_wu), capital expenditure budget (b53_wu), budget absorption (n_real), consistency of fund withdrawal plan (n_consist), achievement of output volume (n_cro), and efficiency value (n_ne).

Table 1. The Attributes Of The Working Unit Data

Attribute Data type Description	**Attribute Data type Description**	**Formula**
kd_ori_wu	Working Unit name	-

Attribute Data type Description	Attribute Data type Description	Formula
loc_wu	Working Unit location	-
b51_wu	Working Unit's personnel expenditure	Σpersonnel expenditure
b52_wu	Working Unit's goods expenditure	Σgoods expenditure
b53_wu	Working Unit's capital expenditure	Σcapital expenditure
budget_wu	Working Unit's budget	Σpersonnel expenditure + Σgoods expenditure + Σcapital expenditure
block_wu	Budget block	Σpersonnel expenditure block + Σgoods expenditure block + Σcapital expenditure block
n_real	budget realization	$P = \frac{RA}{PA} \times 100\%$ P : budget realization score RA : budget realization AA : budget allocation
n_consist	Disbursement plan consistency	$K = \frac{\sum_{i=1}^{n}\left(\frac{RPDK_n - \lvert RPDK_n - RAK_n \rvert}{RPDK_n}\right) x100\%}{n}$ K : consistency of budget absorption planning RAK_n : cumulative budget realization up to month n $RPDK_n$: cumulative fund withdrawal plan up to the nth month n : number of months
n_cro	Achievement of output realization	$CRO = \prod_{i=1}^{m}\left(\left(\frac{RVK_i}{TVK_i} x \left(\prod_{j=1}^{n}\frac{Realization_j}{Target_j}\right)^{\frac{1}{n}}\right)^{\frac{1}{m}}\right)$ CRO : achievement of activity output RVK : realization of activity output volume TVK : target activity output volume m : number of activity outputs n : number of activity output indicators
n_ne	Efficient value	$E = \frac{\sum_{i=1}^{n}((PAKi \times CKi) - RAKi)}{\sum_{i=1}^{n}(PAKi \times CKi)}$ $NE = 50\% + \left(\frac{E}{20} \times 50\right)$ E : efficiency PAK_i : output budget ceiling i RAK_i : realization of output budget i Ck_i : output achievements i NE : efficiency value

To obtain the government's budget performance clustering, this research will test three clustering algorithms, namely the K-Means algorithm, the DBSCAN algorithm, and the AHC algorithm. The K-Means algorithm is a popular clustering algorithm that is used to divide data sets into several clusters according to the proximity of data points. The K-Means algorithm is run based on the following steps:

1. The initial input data includes variable D as a collection of input data, $D=\{x_1,x_2,...,x_n\}$, and the i^{th} data,, $x_i, \in x_i \in R^d$ (d-dimensional space).
2. Initial parameters, namely K as the desired number of clusters, C is defined as a collection of cluster centers, $C=\{c_1,c_2,...,c_K\}$, and m is the number of iterations or convergence criteria.
3. Define the cluster center, c_k as the k^{th} cluster center, $\in c_k \in R^d$.
4. Define a cluster using the formula:

$$S_k=\{x_i \mid \text{argmin}_{cj} \mid\mid x_i - c_j \mid\mid, 1 \leq j \leq K\} \quad (1)$$

With:

S_k : k[th] cluster
x_i, : i[th] data
c_j : The j[th] set of cluster centers
K : number of clusters

5. Perform the algorithm iteration for each t value:
   a. Data sharing with clusters:

$$S_k^{(t)} = \{x_i \mid \text{argmin}_{cj} \mid\mid x_i - c_j^{(t)} \mid\mid, 1 \leq j \leq K\} \text{ untuk } 1 \leq I \leq n \quad (2)$$

   b. Cluster Center Update:

$$c_k^{(t+1)} = \frac{1}{\left|S_k^{(t)}\right|} \sum_{x_i \in S_k^{(t)}} x_i \text{ untuk } 1 \leq k \leq K \quad (3)$$

with:
$S_k^{(t)}$: k[th] cluster for t value
x_i, : i[th] data
$c_j^{(t)}$: set of j[th] cluster centers for t values
K : number of clusters

6. The output results are the final cluster $\{S_1, S_2, ..., S_K\}$.

The DBSCAN algorithm is a clustering algorithm that groups data according to density. The DBSCAN algorithm can be explained in the following steps:

1. The initial input data includes variable D as a collection of input data, $D=\{x1, x2, ..., xn\}$, and the i[th] data, $x_i \in R^d$ (d-dimensional space).
2. Initial parameters, namely ε, are the maximum distance between two adjacent data points, and MinPts is the minimum number of points in the ε-circle of a point so that the point is considered a core point.
3. Define points, namely P as a data point, Q as a neighboring data point of P, and Nε(P) as a neighboring ε-circle of P.
4. Define neighborhood, $P \xrightarrow{neighbor} Q$ is point Q, which is a neighbor of point P.
5. Define Core Points: P is a core point if |Nε(P)| ≥ MinPts.
6. Define Direct Neighbourhood, P $\xrightarrow{\text{directly-reachable}}$ Q, Point Q is a direct neighbor of point P if Q is a neighbor of P and P is the core of the point.
7. Define Boundary Points (Border): P is a boundary point if P is not a core point but has a direct neighbor who is a core point.
8. Define noise: P is noise if P is not a core point and there are no other points that are direct neighbors of P.
9. Algorithm iteration: select points P from D randomly. If P has not been reached, calculate Nε(P).
   If |Nε(P)| < MinPts, mark P as noise.
   If |Nε(P)| ≥ MinPts, form a new cluster and add all reachable points of P to the cluster.
   Repeat this step for all points newly added to the cluster.

The Agglomerative Hierarchical Clustering (AHC) algorithm is an algorithm that combines data points gradually to form a cluster hierarchy. The steps in building the AHC algorithm are as follows:

1. The initial input data includes variable D as a collection of input data, $D=\{x1, x2, ..., xn\}$, and the i[th] data, $x_i \in R^d$ (d-dimensional space).
2. Determine the distance matrix $D_{dist}(i,j)$ to express the distance between x_i and x_j.
3. Define initial clusters, with each point x_i initially considered a separate cluster.

4. Algorithm iteration: select the two closest clusters, defined by *Ca* and *Cb* , which are the two clusters that have the closest distance based on D_{dist}. Merge the two clusters into a new cluster, $C_{new}=C_a \cup C_b$. Update D_{dist} to account for C_{new} as a single entity. Eliminate the old clusters, C_a and C_b, from the cluster list. Add C_{new} to the cluster list. Repeat these steps until there is only one cluster remaining.
5. The results obtained are in the form of a cluster hierarchy that forms an agglomeration tree.

To evaluate the results of clustering, one of the recommended methods is the Davies-Bouldin Index (DBI) method. The DBI method uses cohesion and separation values to generate an index. The cohesion value is the closeness of the data to the centroid of its cluster that is followed. Separation is the distance between centroids in the cluster. The smaller the DBI value (as long as it is greater than zero), the better the cluster formation. The formula for calculating DBI is as follows:

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\left(R_{i,j}\right) \quad (4)$$

with $R_{i,j}$ can be obtained through the following equation:

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (5)$$

with SSW_i, SSW_j dan $SSB_{i,j}$ obtained through the following equation:

$$\text{SSW}_i = \frac{1}{m_i}\sum_{n=1}^{m_i} d\left(x_i, c_j\right) \quad (6)$$

$$\text{SSW}_j = \frac{1}{m_j}\sum_{n=1}^{m_j} d\left(x_i, c_j\right) \quad (7)$$

$$SSB_{i,j} = d(c_i,c_j) \quad (8)$$

With:

$R_{i,j}$: the ratio between cluster i and cluster j
SSW_i, SSW_j : sum of squares within cluster i and j
$SSB_{i,j}$: sum of the square between cluster i and j
$d\left(x_i, c_j\right)$: distance of the i^{th} data point to the j^{th} centroid
$d(c_i,c_j)$: distance from the i^{th} centroid to the j^{th} centroid
k : number of clusters
mi,mj : number of data in the ith and jth clusters

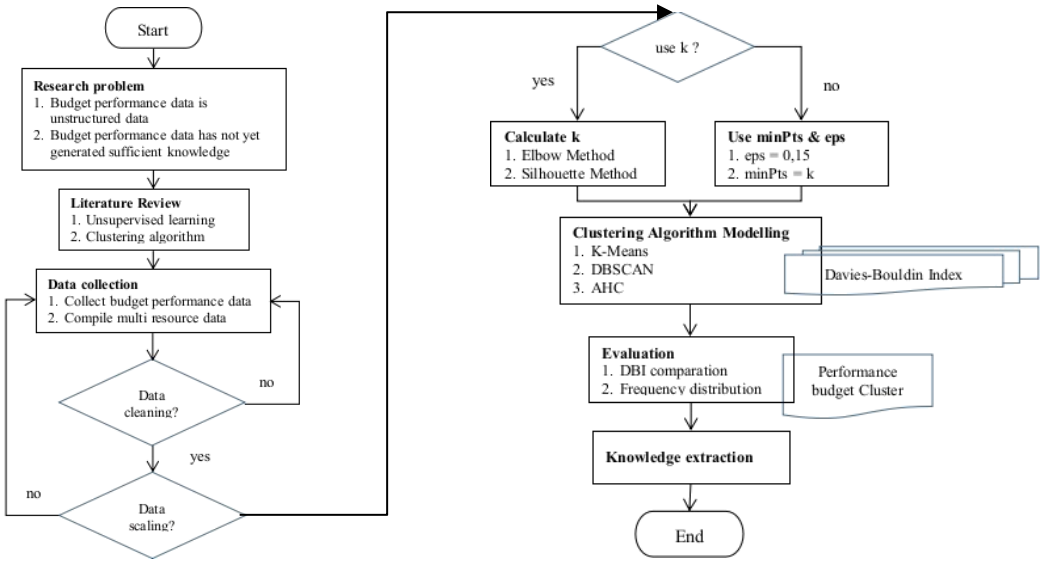


Figure 1. Research Framework

3. RESULTS AND DISCUSSION

3.1. Identification Problem and Literature Review

Based on observations on the official website of the Directorate General of Budget, Ministry of Finance, and Minister of Finance Regulation Number 22/PMK.02/2021, no grouping of budget performance data was found for the Work Unit level. Although there will be 19,460 work units carrying out government tasks in 2021. This makes budget performance data unstructured. The implication is that budget performance data cannot be used to generate knowledge that will later be useful for decision-making.

To handle unstructured data, unsupervised learning methods can be used. Clustering algorithms are the best alternative to handle this data. Based on previous research, there are three popularly used algorithms: K-Means [24] [25] [26] [27] [28], DBSCAN [11] [12] [14] [15] [16], and AHC [1] [2] [4].

3.2. Data Collection

Based on the SMART application data for 2021, 19,460 Work Units were obtained. The Work Units consist of 1,474 Work Units located in Jakarta (code 1), 17,784 Work Units spread across 34 provinces (code 2 - 35), and 211 Work Units overseas (code 50 - 59). Each Work Unit has three main attributes, namely the location of the Work Unit, budget attributes, and budget performance attributes. The OM SPAN application report is the source of data records for budget attributes. From the SMART application reports, data recordings regarding budget performance attributes are obtained. Data is initially stored in tabular form and subsequently converted to.csv format. The budget attributes are translated into personnel expenditure attributes (b51_wu), goods expenditure attributes (b52_wu), capital expenditure attributes (b53_wu), total budget (budget_wu), and budget blocks (block_wu). Meanwhile, the budget performance attributes consist of budget realization (n_real), disbursement plan consistency (n_consist), achievement of output realization (n_cro), and efficient value (n_ne).

3.3. Data pre-Processing

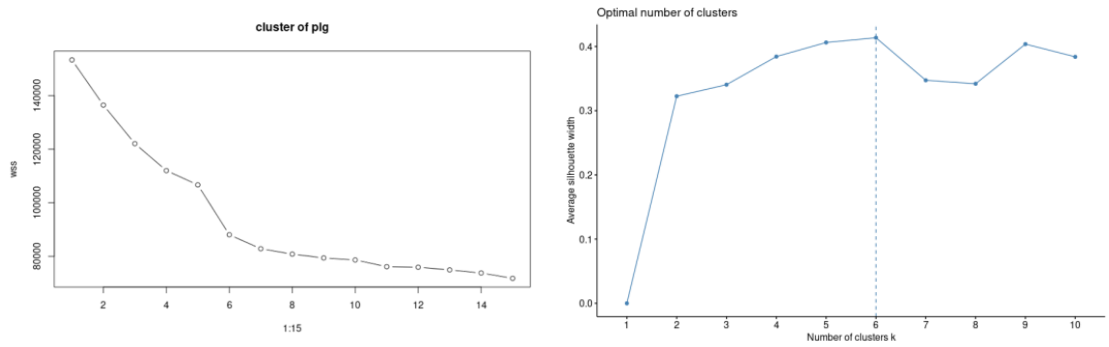
Before using the data in the clustering algorithm, data pre-processing is first carried out. The first stage will be data cleaning by selecting complete data records so that fields containing N/A are not processed further. At this stage, 294 incomplete data fields were found, leaving 19,166 data fields. In the second stage, the most relevant attributes will be selected to form the basis of the clustering algorithm. For the budget attribute, the total budget attribute (budget_wu) and budget block (block_wu) were selected. For the budget performance attribute, the attributes of achievement of output realization (n_cro), and efficient value (n_ne) were selected.

Table 2. Summary of the Working Unit data

	kd_ori_wu	loc_wu	budget_wu	block_wu	n_cro	n_ne
Length	19166					
Class	character					
Mode	character					
Min		1	1.00E+05	0	0.2	0
Median		12	5.80E+09	0	100	63.51
Mean		14.12	6.44E+10	2.26E+09	97.24	68.32
Max		59	8.37E+13	6.43E+12	100	100

3.4. Determination of Parameters k, eps, and MinPts

For the centroid-based clustering algorithm, determining the value of k is crucial. In the K-Means algorithm, determining the value of k can affect the performance of the clusters formed [28]. The k parameter is also used in the AHC algorithm to determine cluster boundaries. To determine the optimum k parameter, you can use the Elbow Method and the Silhouette Method. Figure 2 shows the results of calculating the k parameters from the dataset.

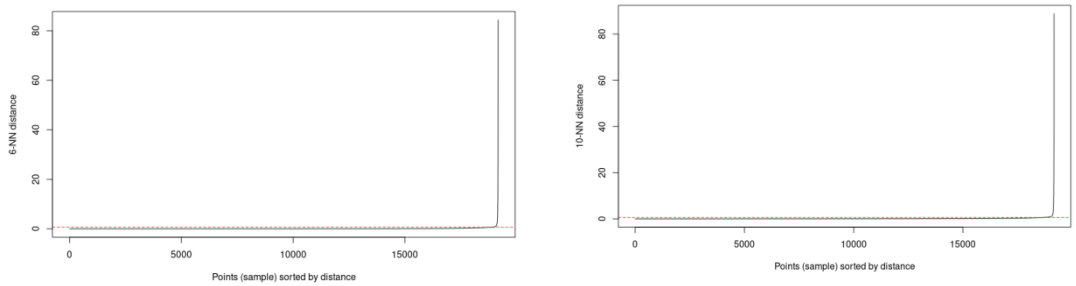


(a) (b)

Figure 2 (a) Elbow Method (b) Silhouette Method

Based on Figure 2, the Elbow method shows the optimum k parameter when k = 10. Meanwhile, the Silhouette Method shows the optimum k parameter when k = 6. Therefore, the two k values will be used in the K-means and AHC algorithms to get the best clusters.

The DBSCAN algorithm does not use k parameters, but eps and MinPts. To determine the optimum eps and MinPts values, the Knee Method can be used. DBSCAN empirically employs MinPts = 4 [12]. However, the minimum MinPts value is d + 1, In this case, the MinPts value = 6. In the previous calculation, the k = 6 and k = 10 values were obtained, these two values will be used to determine the optimum eps and MinPts values. Based on Figure 3, the results are: 1) for k = 6, eps = 0.65, and MinPts = 6; and 2) for k = 10, eps = 0.65, and MinPts = 10.



(a) (b)

Figure 3. Knee Method (a) k = 6 (b) k = 10

3.5. Build Clustering Algorithm

After determining the k, eps, and MinPts parameters, the next step is to run the clustering algorithm to get the best cluster. The results obtained after running the K-Means algorithm are as follows:

Table 3. K-Means Cluster Size

Cluster	Cluster Size	
	k = 6	k = 10
1	4541	3266
2	532	2679
3	4292	504
4	3135	6
5	15	3765
6	6651	2411
7		982
8		763
9		4782
10		8
Total	19.166	19.166

In the K-Means algorithm, for k = 6, the smallest cluster size is 15 and the largest cluster size is 6,651. When k = 10, the smallest cluster size is 6 and the largest cluster size is 4,782. These results indicate that at k = 10, the data distribution tends to be better because the distribution of data in each cluster is more even, even though there are two clusters whose size values differ greatly from those of the other clusters.

The second algorithm that will be used is the DBSCAN algorithm. The results obtained after running the DBSCAN algorithm are as follows:

Table 4. DBSCAN Cluster Size

Cluster	Cluster Size	
	eps = 0.65 MinPts = 6	eps = 0.65 MinPts = 10
0	151	209
1	18828	18735
2	180	14
3	7	179
4		12
5		17
Total	19.166	19.166

At MinPts = 6, the clusters formed are 3 clusters with 151 data points of noise. When MinPts = 10, the number of clusters formed increases to 5 with 309 data points of noise. At the two MinPts values, cluster 1 still has the largest cluster size.

The third algorithm that will be used is the AHC algorithm. The results obtained after running the AHC algorithm are as follows:

Table 5. AHC Cluster Size

Cluster	Cluster Size	
	k = 6	k = 10
1	19148	19142
2	9	1
3	3	5
4	1	5
5	4	3
6	1	4
7		1
8		3
9		1

Cluster	Cluster Size	
	k = 6	k = 10
10		1
Total	19.166	19.166

The number of clusters formed using the AHC algorithm is the same as the number of clusters formed using the K-Means algorithm. For cluster size, the AHC algorithm is the same as the DBSCAN algorithm; the largest is in cluster 1, with a value that is much different from the other clusters.

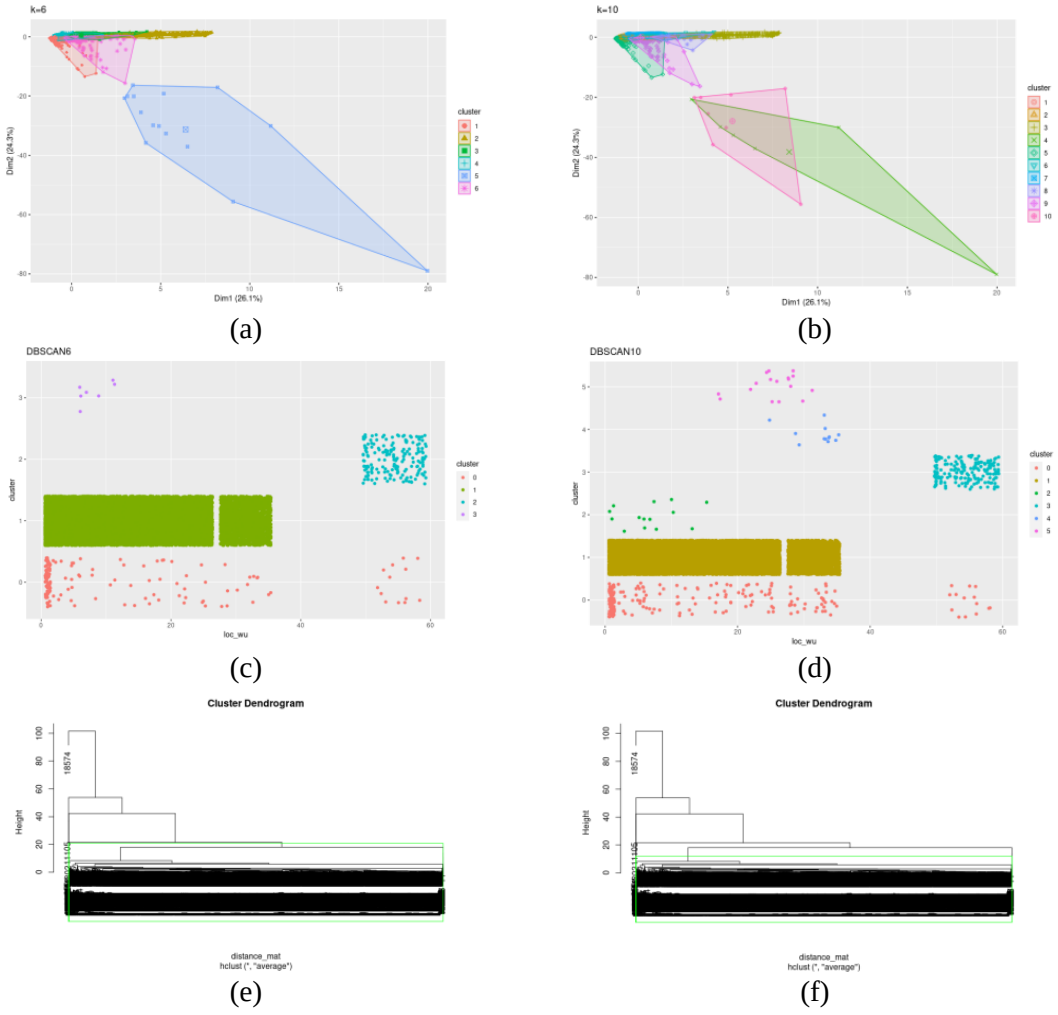


Figure 4. Cluster Visualization: K-Means, k = 6 (b) K-Means, k = 10 (c) DBSCAN, MinPts = 6 (d) DBSCAN, MinPts = 10 (e) AHC, k = 6 (f) AHC, k = 10

3.6. Evaluation Model

In the previous stage, six clustering models had been obtained. To evaluate the results of clustering, one of the recommended methods is the Davies-Bouldin Index (DBI) method. The DBI values of the three algorithms used are as follows:

Table 6. DBI

DBI	K-Means	DBSCAN	AHC
k = 6	1.105577	0.5479952	0.3583472
k = 10	1.052678	0.5398259	0.3755633

Based on Table 9, the parameter k = 10 will generally produce a smaller DBI value, except for AHC. When using the smallest DBI value approach, the best algorithms are AHC, DBSCAN, and then K-Means. According to the statistical approach, the AHC algorithm with parameters k = 6 produces the best cluster for budget performance datasets. These results are because the AHC algorithm forms cluster 1, which has the same characteristics as the data as a whole. This implies that the frequency distribution in cluster 1 is close to the total data (19148 of 19166 data). When compared to the study [8], which obtained three large clusters], the AHC algorithm results are still not optimal.

DBI results for the K-Means algorithm show the highest value among the other two algorithms. Based on the statistical approach, the resulting clusters are not as good as the other two algorithms. However, the K-Means algorithm has the advantage of a more even frequency distribution in each cluster. For k = 10, only cluster 4 and cluster 10 have a very small frequency distribution. The DBSCAN algorithm is moderate, with the advantage of being able to detect noise (outliers) from the dataset. The DBSCAN algorithm's features can be used to build the next stage of machine learning [22].

These results must be tested further concerning the parameters of ease of implementation for regulators. Clusters resulting from the AHC algorithm have the potential to cause non-cooperative behaviors because, in the context of large-scale group decision-making, policies are only taken based on characteristics that are too general [21]. A study [9], which concluded a simple baseline for low-budget active learning for complex data such as image data, confirms that the K-Means algorithm is an alternative for classification purposes. The ten clusters formed by the K-Means algorithm do have their complexity in terms of defining the characteristics of each cluster. But this is a big plus because it lets regulators make policies that fit the specifics of the clusters in question. Furthermore, the K-Means algorithm has been widely implemented to cluster data related to state finances, including Personal Income Levels in Romania [10], capital allocation for Small and Medium-Sized Enterprises (MSMEs) [19], operating cash flow [20], and determinants of SMEs' performance [23]. Based on these things, we choose the K-Means algorithm as an alternative to obtain knowledge from budget performance datasets.

Table 7. Cluster Attributes Means

cluster	loc_wu	budget_wu (billion Rp)	block_wu (billion Rp)	n_cro	n_ne	Characteristic
1	11 - 23	15.68	0.30	99.21	56.82	Budget under National means, budget block under National means, output realization above National means, efficient value under National means
2	13 - 28	32.92	0.56	99.49	94.03	Budget under National means, budget block under National means, output realization above National means, efficient value above National means
3	1 - 56	24.47	0.43	34.06	34.70	Budget under National means, budget block under National means, output realization under National means, efficient value under National means
4	1 - 1	9,403.67	3,110.00	81.93	59.85	Budget above National means, budget block above National means, output realization under National means, efficient value under National means

cluster	loc_wu	budget_wu (billion Rp)	block_wu (billion Rp)	n_cro	n_ne	Characteristic
5	1 - 13	78.05	2.38	99.58	93.73	Budget above National means, budget block a little above National means, output realization above National means, efficient value above National means
6	21 - 59	13.36	0.26	98.22	54.64	Budget under National means, budget block under National means, output realization above National means, efficient value under National means
7	28 - 59	30.87	0.64	99.09	88.73	Budget under National means, budget block under National means, output realization above National means, efficient value above National means
8	1 - 35	27.75	0.80	93.60	16.65	Budget under National means, budget block under National means, output realization under National means, efficient value under National means
9	1 - 10	65.50	2.29	99.18	56.32	Budget above National means, budget block almost same National means, output realization above National means, efficient value under National means
10	1 - 1	41,800.00	25.20	89.52	43.96	Budget above National means, budget block above National means, output realization under National means, efficient value under National means
National	1 - 59	64.40	2.26	97.24	68.32	

Based on the perspective of budget performance, cluster 2 and cluster 5 are ideal clusters because they have achieved high output realization and high-efficiency values, even though they have a small budget. The difference between the two clusters is only in their location. For Work Unit locations, only location 1 stands alone in two different clusters. In general, there is no significant polarization between locations within the country and abroad.

On the budget attribute, cluster 4 and cluster 10 are the two clusters with the largest average budgets. Both of these clusters possess the identical location code attribute, which is location 1. Regulators need to consider the right mix of fiscal policy considering that the two clusters still record low efficiency scores. For the budget blocking attribute, cluster 4 is a cluster that needs policy review because there are indications that large budget blocks have an impact on low-efficiency values. Most of the clusters recorded small budget block values, this shows the regulator's commitment to optimizing the budget to achieve the best performance.

For the dimensions of output achievement, output achievement in cluster 3 is still relatively low. This has implications for the low value of efficiency. Cluster 3 is one of the clusters that needs attention from the regulator. It is hoped that the regulator can further identify the obstacles faced by cluster 3 in realizing the output. Further identification can help regulators formulate prudent fiscal policies.

For the efficiency value attribute, nationally, it is still not encouraging. Cluster 8 in particular had the lowest efficiency value, with quite extreme values. In addition, cluster 1 also

has a low-efficiency value, even though the actual output is high. These results need further identification. Regulators need to check the validity and completeness of the achievement data inputted by the Work Unit. This is because there is a possibility that the low score is due to administrative negligence in inputting performance achievement data. The low-efficiency score in cluster 8 also needs to be viewed with skepticism so that regulators are not mistaken in formulating policies.

4. CONCLUSIONS

Comparing the three clustering techniques, the AHC algorithm with k = 6 has the lowest DBI, 0.3583472. However, the AHC method is difficult to implement. AHC method results accrue in cluster 1 (19148 out of 19166 data points) due to their frequency distribution. Same with DBSCAN results; frequency distribution accumulates in cluster 1. Policy design may be difficult because decision-makers have trouble distinguishing data features. This implies a biased policy. Thus, the optimal option is K-means with parameters k = 10. The K-Means algorithm's DBI is 1.052678. The K-Means method creates 10 clusters.

Based on knowledge extraction, it is determined that cluster 2 and cluster 5 are ideal clusters in terms of budget performance. While the clusters that require attention are cluster 1, cluster 3, cluster 4, and cluster 8. We suggest further identification related to the completeness of the Work Unit performance achievement data to find out the possibility of administrative errors during the input process of budget performance achievements.

For further research, we suggest comparing the results with data from subsequent years to measure the consistency of the clustering algorithm. Research can be continued by using clustering results to predict performance values and classification algorithms. Associative algorithms can also be used to determine the best policy mix in the budgeting sector.

REFERENCES

[1] K. P. Simanjuntak and U. Khaira, "Hotspot Clustering in Jambi Province Using Agglomerative Hierarchical Clustering Algorithm," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 1, no. 1, pp. 7–16, 2021.

[2] A. Triayudi and I. Fitri, "A new agglomerative hierarchical clustering to model student activity in online learning," *TELKOMNIKA*, vol. 17, no. 3, pp. 1226-1235, 2019.

[3] S. Sreedhar Kumar, M. Madheswaran, B. A. Vinutha, H. Manjunatha Singh, and K. V. Charan, "A brief survey of unsupervised agglomerative hierarchical clustering schemes," *International Journal of Engineering and Technology (UAE)*, vol. 8, no. 1, pp. 29-37, 2019.

[4] S. Wu, J. Lin, Z. Zhang, and Y. Yang, "Hesitant Fuzzy Linguistic Agglomerative Hierarchical Clustering Algorithm and Its Application in Judicial Practice," *Mathematics*, vol. 9, no. 4, p. 370, Feb. 2021, doi: 10.3390/math9040370.

[5] I. G. N. M. Jaya and H. Folmer, "Identifying Spatiotemporal Clusters by Means of Agglomerative Hierarchical Clustering and Bayesian Regression Analysis with Spatiotemporally Varying Coefficients: Methodology and Application to Dengue Disease in Bandung, Indonesia," *Geographical Analysis*, vol. 53, no. 4, pp. 767–817, 2021, doi: 10.1111/gean.12264.

[6] A. Naeem, M. Rehman, M. Anjum, and M. Asif, "Development of an efficient hierarchical clustering analysis using an agglomerative clustering algorithm," *Current Science*, vol. 117, no. 6, pp. 1045–1053, 2019.

[7] M. R. Ridwan and H. Retnawati, "Application of Cluster Analysis Using Agglomerative Method," *Numerical: Jurnal Matematika dan Pendidikan Matematika*, vol. 5, no. 1, pp. 33–48, 2021.

[8] I. A. Mezinova, J. B. Amirkhanyan, O. V. Bodiagin, and M. M. Balanova, "The Relationship between the Country‘s Global Competitiveness and its National MNEs," *Visegrad Journal on Bioeconomy and Sustainable Development*, vol. 8, no. 2, pp. 87–92, Nov. 2019, doi: 10.2478/vjbsd-2019-0017.

[9] K. Pourahmadi, P. Nooralinejad, and H. Pirsiavash, "A Simple Baseline for Low-Budget Active Learning," 2021, doi: 10.48550/ARXIV.2110.12033.

[10] V.-C. Bulai, A. Horobeț, and L. Belascu, "Improving Local Governments' Financial Sustainability by Using Open Government Data: An Application of High-Granularity Estimates of Personal Income Levels in Romania," *Sustainability*, vol. 11, no. 20, p. 5632, Jan. 2019, doi: 10.3390/su11205632.

[11] S. Babichev, S. Vyshemyrska, and V. Lytvynenko, "Implementation Of Dbscan Clustering Algorithm Within The Framework Of The Objective Clustering Inductive Technology Based On R And Knime Tools," *Radio Electronics, Computer Science, Control*, vol. 0, no. 1, Apr. 2019, doi: 10.15588/1607-3274-2019-1-8.

[12] X. Li, P. Zhang, and G. Zhu, "DBSCAN Clustering Algorithms for Non-Uniform Density Data and Its Application in Urban Rail Passenger Aggregation Distribution," *Energies*, vol. 12, no. 19, p. 3722, Jan. 2019, doi: 10.3390/en12193722.

[13] V. Martynenko, Y. Kovalenko, I. Chunytska, O. Paliukh, M. Skoryk, and I. Plets, "Fiscal Policy Effectiveness Assessment Based on Cluster Analysis of Regions," *International Journal of Computer Science and Network Security*, vol. 22, no. 7, pp. 75–84, Jul. 2022, doi: 10.22937/IJCSNS.2022.22.7.10.

[14] Z. Li, Y. Li, W. Lu, and J. Huang, "Crowdsourcing Logistics Pricing Optimization Model Based on DBSCAN Clustering Algorithm," *IEEE Access*, pp. 1–1, 2020, doi: 10.1109/ACCESS.2020.2995063.

[15] S.-S. Li, "An Improved DBSCAN Algorithm Based on the Neighbor Similarity and Fast Nearest Neighbor Query," *IEEE Access*, vol. 8, pp. 47468–47476, 2020, doi: 10.1109/ACCESS.2020.2972034.

[16] T. Wang, C. Ren, Y. Luo, and J. Tian, "NS-DBSCAN: A Density-Based Clustering Algorithm in Network Space," *ISPRS International Journal of Geo-Information*, vol. 8, no. 5, p. 218, May 2019, doi: 10.3390/ijgi8050218.

[17] S. Wibisono, M. T. Anwar, A. Supriyanto, and I. H. A. Amin, "Multivariate weather anomaly detection using DBSCAN clustering algorithm," *Journal of Physics: Conference Series*, vol. 1869, no. 1, p. 012077, Apr. 2021, doi: 10.1088/1742-6596/1869/1/012077.

[18] N. A. Febriyati, A. D. GS, and A. Wanto, "GRDP Growth Rate Clustering in Surabaya City uses the K-Means Algorithm," *International Journal of Information System & Technology*, vol. 3, no. 2, pp. 276–283, 2020.

[19] A. Hidayah, "Implementing Data Clustering to Identify Capital Allocation for Small and Medium Sized Enterprises (SMEs)," *ASEAN Marketing Journal*, vol. X, no. 1, pp. 66–74, 2018.

[20] R. Wulaningrum, V. E. Satya, M. Kadafi, D. Y. A. S. Fala, and A. Azizah, "Operating Cash Flow Analysis of Indonesian Provincial Government," Mar. 2022, pp. 571–576, doi: 10.2991/assehr.k.220301.094.

[21] X. Chao, G. Kou, Y. Peng, and E. H. Viedma, "Large-scale group decision-making with non-cooperative behaviors and heterogeneous preferences: An application in financial inclusion," *European Journal of Operational Research*, vol. 288, no. 1, pp. 271–293, 2021, doi: 10.1016/j.ejor.2020.05.047.

[22] C. De Lucia, P. Pazienza, and M. Bartlett, "Does Good ESG Lead to Better Financial Performances by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe," *Sustainability*, vol. 12, no. 13, p. 5317, Jul. 2020, doi: 10.3390/su12135317.

[23] C. Cicea, I. Popa, C. Marinescu, and S. C. Ștefan, "Determinants of SMEs' performance: evidence from European countries," *Economic Research-Ekonomska Istraživanja*, vol. 32, no. 1, pp. 1602–1620, Jan. 2019, doi: 10.1080/1331677X.2019.1636699.

[24] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?," *Pattern Recognition*, vol. 93, pp. 95–112, 2019, doi: 10.1016/j.patcog.2019.04.014.

[25] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, Jun. 2019, doi: 10.3390/j2020016.

[26] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[27] S. Xia et al., "A Fast Adaptive k-means with No Bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020, doi: 10.1109/TPAMI.2020.3008694.

[28] H. Xie et al., "Improving K-means clustering with enhanced Firefly Algorithms," *Applied Soft Computing*, vol. 84, p. 105763, 2019, doi: 10.1016/j.asoc.2019.105763.