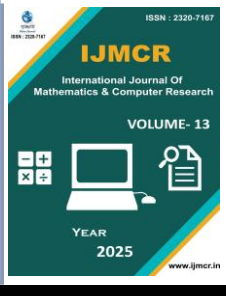
International Journal of Mathematics and Computer Research
ISSN: 2320-7167
Volume 13 Issue 12 December 2025, Page no. – 5953-5958
Index Copernicus ICV: 57.55, Impact Factor: 8.615
DOI: 10.47191/ijmcr/v13i12.03



Comparison of Silhouette, Elbow, and Gap Statistics Optimization Methods in Determining the Number of Best Clusters in K-Means Clustering Analysis

Metalia Widya Diantika[1], Agus Rusgiyono[2*], Bagus Arya Saputra[3]
[1,2,3] Department of Statistics, Universitas Diponegoro, Semarang, Indonesia

ARTICLE INFO

Published Online:
08 December 2025

Corresponding Author:
Agus Rusgiyono

ABSTRACT

Stunting is a condition of malnutrition status that is chronic in growth and development from the beginning of life. Malnutrition puts children at greater risk of death. One of the efforts to overcome stunting is to determine in advance the provinces that need to be prioritized in handling the factors that cause stunting by grouping 34 provinces in Indonesia. This study uses k-means clustering to partition data according to their respective characteristics into the form of two or more clusters, determining the optimal number of clusters through elbow optimization methods, gap statistics, and silhouette. The method used to test the best cluster results is the Davies Bouldin Index (DBI) method. The results of the elbow method clustering test produce K = 3 with a DBI value of 0.6392677; the gap statistics method produces K = 1 without DBI testing because only 1 cluster is formed, while the silhouette method produces K = 2 with a DBI value of 0.2116945. This shows that the results of clustering k-means with the silhouette method produce better cluster quality because it has a lower DBI value than other methods.

KEYWORDS: Stunting; Clusterin; K-Means; Silhouette; Elbow; Gap Statistics

I. INTRODUCTION

Stunting is a condition of chronic malnutrition during the period of growth and development from the beginning of life [1]. According to [2], almost half of all deaths in children under five are caused by malnutrition. This is because malnutrition puts children at greater risk of dying from common infections, increases the frequency and severity of infections, and delays recovery from the infection. One of the efforts to overcome stunting is to determine in advance which provinces need to be prioritized for handling the factors that cause stunting by grouping 34 provinces in Indonesia in 2021. Based on previous research conducted by [3] with the title LBW and SGA Impact Longitudinal Growth and Nutritional Status of Filipino Infants, it was found that there is a relevant relationship between low birth weight (BBLR) babies weighing less than 2.5 kg and stunting. According to [4], argue that BBLR, poor nutritional intake, inadequate sanitation, and infections during growth trigger stunted growth and development and give birth to stunted children. Stunting factors include breastfeeding, children's food patterns, infections, food intake and supply, sanitation, and environmental health.

Clustering is a method or algorithm used to find clusters naturally according to variables in a predefined dataset [5]. K-Means clustering is a method of grouping by partitioning data according to its respective characteristics into the form of two or more groups. K-Means clustering is one of the simplest types of "unsupervised machine learning algorithms".

The number of optimal clusters in the data grouping can be determined through elbow optimization, gap statistics, and silhouette methods. The elbow method determines the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point. Gap Statistics determines the number of optimal clusters more consistently than other measurements, while the Silhouette method makes it is possible to see the quality and strength of clusters in determining the number of best clusters, as well as how good or bad an object is placed in a cluster.

In the previous study conducted by [6] with the title Application of the K-Means Clustering Algorithm for Prediction of Student Academic Achievement Scores, it discussed the grouping of student graduation predicates using Euclidean distances with the k-means clustering method. The results of the research were obtained by 4 groups of student graduation predicates, namely very good, good, adequate, and lacking.

Based on this, this study uses the k-means clustering method to group 34 provinces in Indonesia in 2021 based on the factors that cause stunting in the region. The number of best clusters in this study was determined through a comparison of elbow

optimization, gap statistics, and silhouette methods. The purpose of this study is to obtain the best clustering results to determine the provinces that need to be prioritized for handling the factors that cause stunting, as an effort to overcome the stunting rate in the region.

II. THEORETICAL FRAMEWORK

Cluster analysis is a multivariate technique that has the main purpose of grouping objects based on their characteristics [7]. Cluster analysis has assumptions that need to be met, namely:

A. Representative

Testing (sample representing the population) can be conducted through a Kaiser Meyer Olkin (KMO). According to [8], the formula for KMO is:

$$\text{KMO} = \frac{\sum_{j=1}^{p}\sum_{l=1,j\neq l}^{p} r_{jl}^2}{\sum_{j=1}^{p}\sum_{l=1,j\neq l}^{p} r_{jl}^2 + \sum_{j=1}^{p}\sum_{l=1,j\neq l}^{p} a_{jl,m}^2} \quad (1)$$

where p is the number of variables, n is the number of objects, and x_{ij} is the value of object i on variable j. The term x_{il} represents the value of object i on variable l. The coefficient r_{jl} denotes the Pearson correlation coefficient between variable j and variable l, while $a_{jl,m}$ is the partial correlation coefficient between variable j and variable l, controlling for m variables.

B. Non-Multicollinearity

Multicollinearity refers to the presence of a linear relationship or a high degree of correlation among independent variables. According to [9], one way to detect the presence of multicollinearity is by calculating the Variance Inflation Factor (VIF) using the following formula:

$$\text{VIF} = \frac{1}{\left(1-R^2\right)} \quad (2)$$

R^2 is the coefficient of determination of a variable bound to its free variable. The limit of the VIF value is 10. If the VIF is lower than 10, then there is no multicollinearity [9].

Clustering analysis groups data based on similarity or the magnitude of differences to determine how similar the objects are. One of the dissimilarity measures used in this study is the Euclidean distance. The Euclidean distance between two points is defined as the square root of the sum of the squared differences between the corresponding variables of the objects. The formula for calculating the Euclidean distance is as follows [10]:

$$d\left(x_i, c_k\right) = \sqrt{\sum_{j=1}^{p}\left(x_{i,j} - c_{k,j}\right)^2} \quad (3)$$

where

k : 1, 2, …, q

$d(x_i, C_k)$: the Euclidean distance between object i on variable j and the center of cluster k on variable j

$x_{i,j}$: the value of object i on variable j

$c_{k,j}$: the value of the cluster center (centroid) k on variable j

p : the number of observed variables

q : the total number of clusters

The k-means method aims to group existing data into several groups, where the data in one group have different characteristics from those in other groups [6]. The basic steps of the k-means algorithm are as follows:

1. Determine the number of K clusters to be formed.
2. Randomly select the initial cluster centers (centroids).
3. Calculate the distance of each object to each centroid using the Euclidean distance formula:

$$d(x_i, c_k) = \sqrt{\sum_{j=1}^{p}(x_{i,j} - c_{k,j})^2}$$

4. Assign each object to the nearest centroid. An object becomes a member of the k-th cluster if its distance to the k-th centroid is the smallest compared to its distances to the other centroids.
5. Update the centroid by computing the average value of all objects belonging to each cluster, using the following equation:

$$\text{C}_{k,j} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{i,j} \quad (4)$$

where

k : 1, 2, 3, …, q

j : 1, 2, 3, …, p

$C_{k,j}$: the centroid of cluster k on variable j

$x_{i,j}$: the value of object i on variable j

n_k : the number of objects in cluster k

6. Repeat steps 3–5 until the cluster memberships no longer change.

There are several methods for determining the optimal number of clusters. Below are some techniques commonly used to optimize the selection of the number of clusters.

1) Elbow

The identification of a data cluster aims to minimize the difference between cluster points. The elbow method is a technique used in determining the optimal number of clusters by taking into account the percentage of the comparison of the number of clusters with the shape of the elbow at a point [11]. The determination of the best cluster value by the elbow method was obtained by calculating the lowest sum of squared error (SSE) value in the data, with the following formula:

$$\text{SSE} = \sum_{k=1}^{k}\sum_{x_i \in S_k}\left(x_i - c_k\right)^2 \quad (5)$$

SSE represents the sum of squared errors, k is the number of clusters, x_i is the value of object i that belongs to cluster k, S_k is the set of objects in cluster k, and c_k is the center (centroid) of cluster k.

2) *Gap Statistics*

Gap Statistics is one of the effective methods for determining the optimal number of clusters in a dataset. Its purpose is to identify the most stable number of clusters compared with other measurement approaches. The Gap Statistics method compares the value of the log(W_k) curve between clusters formed from the observed data and clusters generated from reference data with a uniform distribution. A uniform distribution is a distribution in which each random variable has an equal probability. The difference between paired objects within a cluster is calculated using the following formula:

$$Dr = \sum_{i,j \in C_r} d_{i,j} \quad (6)$$

where

$d_{i,j}$: the squared Euclidean distance between object i and object j

D_r : the total Euclidean difference for cluster r

r : the number of clusters.

1. Then, the within-cluster sum of squares (W_k) is calculated for n objects using the following formula:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \quad (7)$$

The gap value is obtained by calculating the difference between 2. the standardized approach of uniformly distributed reference data and the observation data.:

$$\text{gap}_n(k) = E^*\{\log(W_k)\} - \log(W_k) \quad (8)$$

where $E^*\{\log(W_k)\}$is the expected value of the logarithm of W_{kb}^*under Bresampling iterations of the reference data, defined as: 3

$$E^*\{\log(W_k)\} = [\tfrac{1}{B}] \sum_{b=1}^{B} \log(W_{kb}^*) \quad (9)$$

where

$b = 1,2,3,\ldots,B$

B: the total number of resampling iterations from the uniform distribution

$k = 1,2,3,\ldots,K$

K: the maximum number of clusters

The criterion for determining the optimal number of clusters is based on the highest gap statistic value or the first value at which the gap statistic shows only a minimal increase, as k increases [12].

3) Silhouette

The Silhouette value ranges from -1 to 1. When the value is between 0 and -1, it indicates that the cluster is widely spread or that the distance between points within the cluster is large. If the Silhouette value is close to 1, it means that the distance between points within the cluster is small, while the distance between points in different clusters is large. Thus, an ideal Silhouette value approaches 1 [5].

The optimal number of clusters is the one with the highest Silhouette value. The calculation of the optimal k value using the Silhouette Index is as follows:

1. Calculates the average difference of the i object with all objects in a cluster

$$a(i) = \frac{1}{n_k - 1} \sum_{h \in Cl_k, h \neq i} d(i,h) \quad (10)$$

where

$a(i)$: the average difference of object i with all objects in one Cluster

h : another object in one Cluster k

d(i,h) : the difference between objects i and h

n_k : the number of objects on the Cluster k

Cl_k : a collection of k cluster objects

Calculates the average difference of the i-i object with all objects that are in other clusters

$$d(i,v) = \frac{1}{n_v} \sum_{h \in Cl_v, h \neq i} d(i,h) \quad (11)$$

where d(i,v) is the average difference of object i with all objects in the other clusters v, nv is the number of objects in the v th cluster, and Clv is the set of objects of the V cluster, where k≠v

Specifies its minimum value of b(i), which indicates the difference in the average of object i for the nearest cluster to its neighbor

$b(i) = \min d(i,v)$ (12)

where b(i) is the minimum value of the average difference of object i with all objects in other clusters v

Calculating silhouette values

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (13)$$

S(i) is the value of s, b(i) is the minimum value of the average difference of object i with all objects between clusters, and a(i) is the average of the difference of object i with all objects in a cluster. Next, it can calculate the silhouette coefficient defined as the average s(i), as follows:

$$SC = \frac{1}{n} \sum_{i=1}^{n} s(i) \quad (14)$$

where n is the number of observations, and SC is the value of the Silhouette Coefficient.

The Davies Bouldin Index (DBI) is a method of measuring cluster evaluation in a grouping. The DBI calculation was tested in terms of cohesion and separation values. The value of cohesion is defined as the value of the proximity of a cluster's data to the centroid of its cluster. Meanwhile, the separation value is defined as the difference between the centroids of other

clusters. The validation calculation of the Davies Bouldin Index includes [13]:

1. Calculate the value of $S_k$to determine the cohesion within each cluster using the following equation:

$$S_k = \frac{1}{n_k} \sum_{i=1}^{n_k} d(x_i, C_k) \quad (15)$$

where

S_k: the average distance between each object $i$and the centroid of the cluster k

$k = 1,2, \dots, q$

q: the total number of clusters

n_k: the number of objects in a cluster k

C_k: the centroid of the cluster k

$d(x_i, C_k)$: the distance between objects i and the centroid of the cluster k

2. Calculate the value of $M_{kv}$to determine the separation or difference between clusters by measuring the distance between the centroids of the cluster $k$and the centroid of another cluster v. The formula is:

$$M_{kv} = d(C_k, C_v), k \neq v \quad (16)$$

where

M_{kv}: the distance between the centroid of the cluster k and the centroid of the cluster v

C_k: centroid of cluster k

C_v: centroid of cluster v

$d(C_k, C_v)$: the distance between the two centroids

3. Calculate the ratio to compare the cohesion of the cluster k and cluster v:

$$R_{kv} = \frac{S_k + S_v}{M_{kv}} \quad (17)$$

4. Calculate the maximum cluster similarity value

$$R_k = \max_{k \neq v} (R_{kv}) \quad (18)$$

This value represents the maximum similarity between cluster k and any other cluster v.

5. Calculate the DBI value

$$\text{DBI} = \frac{1}{K} \sum_{k=1}^{K} R_k \quad (19)$$

where DBI denotes the Davies–Bouldin Index, and $K$is the total number of clusters. The above equation shows K as the number of clusters. If the DBI value obtained is lower, it shows the more optimal the number of clusters obtained. The profiling of optimal clustering results explains the characteristics of each, so the inclination of each cluster can be observed. The characteristics of clusters can be seen by calculating the average of the members of each variable in the study.

III. RESEARCH METHOD

The data in this final project research is secondary data on stunting factors in toddlers in Indonesia, based on 34 provinces in 2021, data obtained from the 2021 Indonesian Health Profile Book. Stunting factor data in toddlers includes five variables, namely: inadequate sanitation, malnourished toddlers, adequate doctors, low birth weight, and short nutritional status. The data analysis in this study uses R software, with the following stages:

1. Input data on the causes of stunting for toddlers
2. Testing representative assumptions using the SME test
3. Testing of the assumption of multicollinearity using the VIF value. If there is multicollinearity, PCA can be performed, and the results of the PCA value calculation are used as a substitute for the previous data value.
4. Clustering uses the k-means method through the calculation of Euclidean distances and grouping objects in each cluster.
5. Determine the number of clusters through elbow, gap statistics, and silhouette methods
6. Calculate the coefficient value of each cluster using the validation of the Davies-Bouldin index
   a. Calculating the average difference of the object with the centroid cluster followed
   b. Calculating the difference between centroids in one cluster and centroids in other clusters
   c. Calculating the ratio as a comparison of the k cluster with the fifth cluster
   d. Determine the ratio between clusters that have the maximum value
7. Interpretation of the results of data grouping based on the characteristics of each cluster formed. The lowest Davies-Bouldin index value will be selected as the optimal number of clusters.

IV. RESULTS AND DISCUSSION

The sample assumption test representing the population using KMO was carried out to determine the sufficiency requirements of a sample. The results of the KMO test on the data on the causes of stunting under five are shown in the Table 1. below:

Table 1. Results of the Value of SMEs

Variable	SMEs
Unsanitary	0.59
Malnutrition in toddlers	0.74
Physician sufficiency	0.66
Low birth weight	0.60
Short nutritional status	0.68

Research data that have been representative of the population can be analyzed at a later stage with multicollinearity testing, in order to detect linear relationships between variables. The results of data processing in this study obtained a VIF value lower than 10. Therefore, it can be concluded that each variable does not have a multicollinearity relationship.

Table 2. VIF Value

Variable	VIF
Unsanitary	2.035156
Malnutrition in toddlers	1.475299
Physician sufficiency	2.193533
Low birth weight	1.564707
Short nutritional status	2.096464

This study uses the k-means method and Euclidean distance measure to group data based on predetermined cluster sizes, namely K=1, K=2, to K=10. The manual calculation of the k-means is:

1. Determine the K value of the cluster formed
2. Assign the initial cluster center randomly.
3. Calculate the difference between each object and each cluster center using Euclidean distance.
4. Grouping each object into the nearest cluster center. Data can be part of the kth cluster if the difference of data to the kth centroid is lower than the difference to the other centroid
5. Assign a new cluster center using the average calculation of the objects on each cluster
6. Calculate the difference of each object to each new centroid using the Euclidean difference for the 2nd iteration
7. Grouping each object on the nearest centroid.
8. Since there are still objects that change clusters with Euclidean differences in the iteration process, repeat the calculation steps 5 to 7. The calculation stops when there is no change in the members of each cluster.

\

The next stage is to determine the best cluster value using the elbow, gap statistics, and silhouette methods. After data processing, the best K value was obtained using the elbow method at K=3, the gap statistics method at K=1, and the silhouette method at K=2. The results of the K value determination chart using these three methods can be observed below:

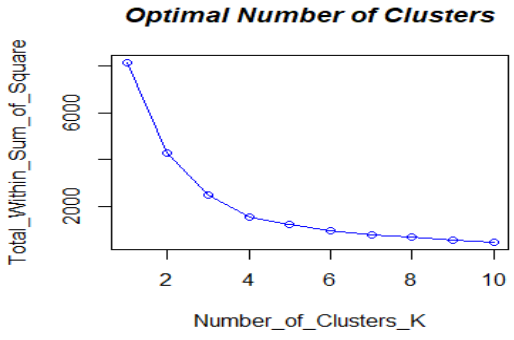


Figure 1. Elbow Method Output

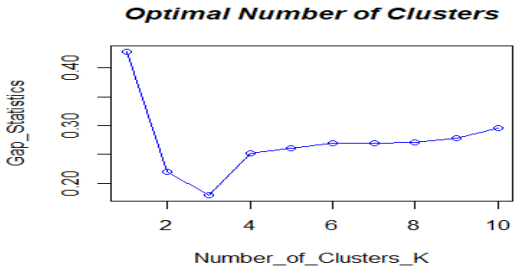


Figure 2. *Output* of *the Gap Statistics Method*

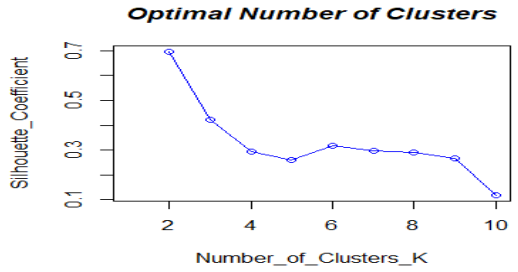


Figure 3. *Silhouette* Method *Output*

After obtaining the number of best clusters from each optimization method, the validation process is continued on the cluster with DBI calculation to determine the number of best clusters. The results of the clustering test with the elbow method obtained DBI = 0.6392677, with gap statistics, the DBI test was not carried out because there was only one cluster, while with the silhouette method, the value DBI = 0.2116945 was obtained.

Based on the results of the above tests, it can be determined that clustering with the silhouette method has a better cluster quality because the silhouette DBI value is lower than other methods. The optimal number of clusters with K=2, will be carried out in the profiling stage to see the characteristics of each cluster. The characteristics of each cluster can be represented by looking at the average of the members of each variable. Here is the average of each variable in the cluster that is formed:

Table 3. Average Variables of Stunting Factors for Toddlers

Variable	Cluster	
	1	2
Unsanitary	17.82%	59.19%
Malnutrition in toddlers	0.90%	1.70%
Physician sufficiency	9.46%	49.50%
Low birth weight	3.61%	4.80%
Short nutritional status	8.01%	10.40%

Table 3. shows that the average cluster of factors causing stunting for toddlers is in cluster 2. Cluster 2 means that the provinces in cluster 2 have higher factors that cause stunting under five than cluster 1. Cluster 1 is seen to have an average cluster smaller than cluster 2. This means that the members of cluster 1 are provinces with low stunting factors for toddlers . Based on this, it can be interpreted that Papua is a province that

needs attention from the Indonesian government, because it has factors that cause high stunted toddlers, especially in improper sanitation factors, toddlers with malnutrition, adequate doctors, BBLR, and the status of toddlers with short nutrition.

V. CONCLUSION

The results of the above calculation and analysis can be concluded as follows:

1. The grouping of k-means methods from 34 provinces is:

a. At K=1, cluster members covering all provinces, consisting of 34 provinces

b. In K=2, members were obtained in the 1st cluster, which was 33 provinces and the 2nd cluster, which was 1 province.

c. In K=3, the 1st cluster members were obtained from 1 province, the 2nd cluster was 9 provinces and the 3rd cluster consisted of 24 provinces

2. The results of clustering of factors that cause stunting under five in Indonesia using the k-means method obtained the optimal number of clusters, namely at K=2. This can be observed from the validation results of DBI = 0.2116945, which is the lowest DBI result. The grouping gave the result that the 1st cluster consisted of 33 provinces, while the 2nd cluster consisted of 1 province.

3. The results of cluster profiling showed that in the number of 2 clusters, it was obtained that the average cluster with the highest stunting factors for toddlers was in cluster 2, which means that the members of cluster 2 are provinces with high factors that cause stunting for toddlers. Cluster 1 has a lower average cluster than cluster 2, which means that cluster 1 members are provinces with low stunting factors for toddlers. Therefore, it is expected that the government in Indonesia will pay more attention to the provinces in cluster 2, namely Papua, which is a province with a high average of stunting factors for toddlers compared to cluster 1, so that in these provinces, it can suppress the factors that cause stunting under five.

REFERENCES

1. Margawati, A., & Astuti, A. M. Pengetahuan ibu, pola makan dan status gizi pada anak stunting usia 1–5 tahun di Kelurahan Bangetayu, Kecamatan Genuk, Semarang. *Jurnal Gizi Indonesia*, 6(2), 82–89. **2018.**
2. UNICEF. Child Malnutrition. https://data-uniceforg.topic/nutrition/malnutrition/ **2020.**
3. Blake, R. A., et al. LBW and SGA Impact Longitudinal Growth and Nutritional Status of Filipino Infants. *PLoS ONE*, 11(7), 1–13. **2016.**
4. Arifin, D. Z., Irdasari, S. Y., & Sukandar, H. Analisis Sebaran dan Faktor Risiko Stunting pada Balita di Kabupaten Purwakarta. *Jurnal Pustaka Unpad*, 1–9. **2012.**
5. Malik, A., & Tuckfield, B. *Applied Unsupervised Learning with R*. Packt Publishing. **2019.**
6. Helilintar, R., & Farida, I. N. Penerapan Algoritma K-Means Clustering untuk Prediksi Prestasi Nilai Akademik Mahasiswa. *Jurnal Sains dan Informatika*, 4(2), 80–87. **2018.**
7. Awalluddin, A., & Taufik, I. Analisis Cluster Data Longitudinal pada Pengelompokan Daerah Berdasarkan Indikator IPM di Jawa Barat. *Prosiding Seminar Nasional Metode Kuantitatif*, 187–194. **2017.**
8. Widarjono, A. *Analisis Statistik Multivariat Terapan*. UPP STIM YKPN. **2010.**
9. Gujarati, D. *Dasar-dasar Ekonometrika Jilid 2*. Erlangga. **2009.**
10. Johnson, W. A., & Wichern, D. W. *Applied Multivariate Statistical Analysis* (6th ed.). Pearson Prentice Hall. **2007.**
11. Madhulatha, T. S. An Overview on Clustering Methods. Journal of Engineering, 2(4), 719–725. **2012.**
12. Silvi, R. Analisis Cluster… Jurnal MANTIK, 4(1), 22–31. **2018.**
13. Kartikasari, M. D. Self-Organizing Map Menggunakan Davies–Bouldin Index… Jambura Journal of Mathematics, 3(2), 187–196. **2021.**
14. Kemenkes RI. Profil Kesehatan Indonesia. Kemenkes RI. **2021.**