JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 4, No 1, March 2025
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

ROCK Ensemble Cluster Method for People's Welfare Analysis A Mixed Data Approach

Arsilla Uswatunnisa[1]*, Devni Prima Sari [1]
[1] Departemen Matematika, Universitas Negeri Padang, Sumatera Barat, Indonesia

* Corresponding Author: arsillauswatunnisa98@gmail.com

Article Info

Article History:
Received: 18 03 2025
Revised: 21 04 2025
Accepted: 23 04 2025
Available Online: 23 04 2025

Key Words:
Ensemble Cluster
People's Welfare
ROCK

Abstract

This study clusters districts/cities in West Sumatra based on public welfare indicators using the Ensemble Cluster Method with the ROCK algorithm. This approach handles mixed data, where numeric data is clustered with Hierarchical Agglomerative Clustering, while categorical data uses ROCK. The clustering results are combined through Cluster Ensemble to improve accuracy. Secondary data from BPS 2023 includes eight indicators of people's welfare. Clustering was validated using Compactness (CP). Results showed five optimal clusters, with a CP value of 0.44. Cluster 1 has the greatest welfare challenges, while Cluster 5 shows the highest welfare. These findings can be used as a basis for formulating more targeted regional development policies.

1. INTRODUCTION

People's welfare is a measuring point for assessing the quality of development in a region. The level of people's welfare in each region can be analyzed through various indicators used to assess the success or failure of development in a region [1]. The human development index (HDI) is a measure of the quality of human life in a region using indicators that have been agreed upon globally through the United Nations Development Programme (UNDP). This index has a range of values between 0 and 100, where the higher the number, the better the quality of life and level of human development in the region [2]. In 2023, BPS noted that the HDI of West Sumatra province was 75.16, or an increase of 0.64%. Although there has been an increase compared to 2022, the HDI growth of West Sumatra in 2023 was lower than the previous year [3]. Therefore, regional grouping in West Sumatra was conducted to analyze the similarity of welfare levels between regions. This aims to assist the government in determining development priorities, considering that each region has a different level of welfare. Grouping districts/cities in West Sumatra can use multivariate analysis, namely cluster analysis. Cluster analysis is a statistical analysis used in grouping objects. The basic thing that can be done is to form a grouping of districts/cities into a group with similar characteristics.

The welfare indicator data is mixed, so the method specifically designed to effectively integrate numerical and categorical data is ensemble clustering. Numerical data is clustered using the hierarchical agglomerative method, while categorical data is clustered using ROCK. The ROCK method is also reapplied when combining clustering results [4].

Previous research has compared the ROCK ensemble method with the SWFM method on mapping districts/cities in East Java based on indicators of disadvantaged areas, where the results of the ROCK ensemble method show a higher level of accuracy compared to the SWFM method [5]. In previous research "Clustering Algorithms for Categorical Data: A Monte Carlo Study" researchers compared five methods, namely the ROCK method, k-modes, Fuzzy k-modes, k-populations, and Average Linkage in processing categorical data, where the results showed that ROCK was the method that was most resistant to overlap between clusters and had the most stable performance, so it is recommended for complex or overlapping categorical data. [6]

2. THEORETICAL BASIC

2.1 Cluster Analysis

Cluster analysis is a technique in multivariate analysis used to group a set of objects (dataset) into several groups based on similar characteristics. The purpose of clustering is to group similar objects in one group and separate different objects into other groups [7]. In order to obtain clusters that are as homogeneous as possible, similarity and dissimilarity measures are used in cluster analysis [8].

1. A similarity measure ($\boldsymbol{sim}$) is used to determine pairs of points that have similar characteristics in data. The $\boldsymbol{sim}$ between points a and b is expressed as $sim(a,b)$. The higher $sim(a,b)$ the higher the sim between the two objects. Each pair of points a and b fulfills the following conditions:
   a. $0 \leq sim(a,b) \leq 1$
   b. $sim(a,b) = 1$
   c. $sim(a,b) = sim(b,a)$
2. Measures of dissimilarity are generally applied to numerical data. The degree of dissimilarity between two points x and y is expressed by $d(x,y)$, the higher the value of $d(x,y)$, the higher the difference between the two points. so that the possibility of both joining a cluster is smaller. Each pair of points x and y satisfies the following conditions:
   a. $d(x,y) \geq 0$
   b. $d(x,y) = 0$
   c. $d(x,y) = d(y,x)$

2.2 Data Clustering Methods

Cluster analysis groups data based on its type, namely numeric and categorical. An explanation of the grouping of each type of data is as follows.

1. Clustering Numerical Data
   Numerical data clustering is based on a measure of dissimilarity, a common measure being the Euclidean distance. For example, if there are two points with $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ then the distance can be calculated using [9] :

$$d_{ij} = \sqrt{\left(\sum_{k=1}^{p} x_{ik} - x_{jk}\right)^2} \quad (1)$$

This method can be used for clustering numerical data in hierarchical and non-hierarchical methods. The hierarchical method is a method in cluster analysis that forms certain levels like a tree structure because the clustering process is carried out in stages. The hierarchical cluster method consists of two methods, namely the Agglomerative method and the Devise method. In the agglomerative method, every time a cluster is merged, the distance between clusters is updated again. Distance improvement methods that can be used to calculate the distance between two clusters [10]:

a. Single Linkage
Single linkage is grouping based on the closest distance between two points.

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\} \quad (2)$$

Where:
$d_{UV} =$ The distance from node u to node v

b. Complete Linkage
Complete linkage is grouping based on the furthest distance between two points

$$d_{(UV)W} = maks\{d_{UW}, d_{VW}\} \quad (3)$$

c. Average Linkage
Average linkage is a grouping based on the average value of the distance between objects.

$$d_{(UV)W} = \frac{1}{N_{UV}N_W}\sum_q \sum_r d_{qr} \tag{4}$$

d. Ward

Ward is the grouping of objects by minimizing the variance within the cluster.

$$d_{(UV)W} = \frac{(n_u+n_w)d_{uw}+(n_v+n_w)d_{vw}-n_w d_{uv}}{n_u+n_v+n_w} \tag{5}$$

2. Clustering Categorical Data

Categorical data can be clustered using various methods, one of which is the ROCK (Robust Clustering Using Links) method, which utilizes “links” to form clusters [11]. The stages in the ROCK method are as follows:

a. Calculating similarity ($\boldsymbol{sim}$)

Calculating the sim between the pair of the a-th point (p_a) and the b-th object (p_b) is calculated using the following formula:

$$sim(p_a, p_b) = \frac{|p_a \cap p_b|}{|p_a \cup p_b|} \tag{6}$$

Where:

$|p_a \cap p_b| =$ the number of categories in common between the a-th object and the b-th object

$|p_a \cup p_b| =$ the total number of categories contained in the a-th object and the b-th object

b. Calculating neighbors

A pair of points p_a and p_b are considered neighbors if $sim(p_a, p_b) \geq \theta$. Threshold (θ) is determined by the researcher to control the level of closeness between objects with a value range of $0 < \theta < 1$. The chosen value of θ indicates the minimum level of similarity that objects must have to be considered as neighbors.

c. Calculating the link between observations

The link value between p_a and p_b is the number of shared neighbors that p_a and p_b have. The magnitude of this link value depends on the specified threshold θ. If there are observations p_a, p_b and p_c, where p_a is a neighbor of p_c,and p_b is a neighbor of p_c, then p_a has a link with p_c, even though p_a is not a direct neighbor of p_c. Matrix $\boldsymbol{A}$ can be used to count the number of links of all possible pairs of nodes. The matrix is $n \times n$, where n is the number of nodes. The value of p_a and p_b is 1 if they are similar (neighbors), and 0 otherwise. The number of links between pairs p_a and p_b is calculated by multiplying matrix $\boldsymbol{A}$ by itself. The larger the number of links between p_a and p_b, the more likely they are to be in the same cluster.

d. Find the goodness measure of the cluster pair

Goodness measure is the merger between clusters C_a and C_b defined as follows:

$$g(C_a, C_b) = \frac{link\,(C_a, C_b)}{(n_a+n_b)^{(1+2f(\theta))} - n_a^{(1+2f(\theta))} - n_b^{(1+2f(\theta))}} \tag{7}$$

The number of links of all possible pairs of points in clusters C_a and C_b is denoted by , $(C_a, C_b) = \sum_{p_a \in C_a} \sum_{p_b \in C_b} link(p_a, p_b)$.With n_a and n_b being the number of members in clusters C_a and C_b. The function used to determine cluster merging is $f(\theta) = \frac{1}{1+2\theta}$. Objects will be grouped into the same cluster based on the largest goodness measure value.

2.3 Cluster Ensemble

The cluster ensemble method is a technique that combines the results of various clustering algorithms to obtain a more optimal division of data. Cluster ensemble is also applied to combined data by using the CEBMDC (Cluster Ensemble Based Mixed Data Clustering) algorithm [12]. The CEBMDC algotima works with the following steps:

1. Separating the combined data into two parts (numeric and categorical)

2. Performing clustering of each type of data.
3. Combining the clustering results through ensemble.
4. Using ensemble clustering with categorical data clustering method to get the final cluster.

2.4 Validation of Grouping Results

Once the clustering process is complete, the quality of the formed cluster structure can be evaluated using various validity measures. One of the measures used is Compactness (CP), which measures the average distance between each pair of data points in the same cluster. Here is the CP formula for numerical data as follows [13] :

$$CP = \frac{1}{N}\sum_{k=1}^{K} n_k \left(\frac{\sum_{x_i x_j \epsilon C_k} d(x_i,x_j)}{\frac{n_k(n_k-1)}{2}} \right) \tag{8}$$

Where, K is a Number of clusters formed, n_k is Number of points in the Kth cluster, $d(x_i, x_j)$ is Distance between the i-th point and the j-th point, and N is Total points.

The lower the Compactness (CP) value, the better the quality of the clusters formed. For categorical data, clustering validation can also be measured using compactness, which is based on the average sim between pairs of points in the same group. The following is the formula used:

$$CP *= \frac{1}{N}\sum_{k=1}^{K} n_k \left(\frac{\sum_{x_a x_b \epsilon C_k} sim(x_a,x_b)}{\frac{n_k}{2}} \right) \tag{9}$$

Where $sim(x_a, x_b)$ is the similarity between the a-th point and the b-th point. Meanwhile, the greater the CP* value, the better the resulting cluster.

3. METHOD

This study uses data on people's welfare indicators in 2023 published by the Central Bureau of Statistics (BPS). The research object includes 19 districts/cities in West Sumatra, with eight variables consisting of six numerical variables and two categorical variables. The variables used in this study reflect welfare indicators in the fields of economy, education, and health. Economic indicators include the percentage of poor people and the open unemployment rate (numerical data), as well as population density and the Gini index (categorical data). The education indicators include literacy rate and school enrollment rate (numerical data). Meanwhile, health indicators consist of life expectancy and morbidity rates (numerical data).

Table 1. Research Variables

No	Notasi	Variable
1	x_1	Percentage of Poor Population
2	x_2	Open Unemployment Rate
3	x_3	School participation rate
4	x_4	Literacy Rate
5	x_5	Life Expectancy Rate
6	x_6	Morbidity Rate
7	x_7	Population Density
8	x_8	Gini index

Data processing in this study was carried out using R software and the CEBMDC algorithm. The stages are as follows:

1. Separating numerical and categorical data
2. Clustering numeric data with hierarchical agglomerative clustering method.
3. Clustering categorical data with the ROCK method.
4. Selecting the best clustering results based on the smallest CP value for numeric data and the largest CP for categorical data.
5. Combining the best clustering results through the ensemble process.
6. Re-clustering using the ROCK method to get the final result.

4. RESULTS AND DISCUSSION

The results of data analysis in a study entitled “ROCK Ensemble Cluster Method for People's Welfare Analysis Mixed Data Approach” produced regional clustering with the CEBMCD algorithm approach. In this process, numerical data is clustered with the hierarchical agglomerative method, and categorical data is analyzed with the ROCK method.

4.1 Clustering Numerical Data

In clustering numerical data, processing using hierarchical agglomerative methods includes single linkage, complete linkage, average linkage, and ward approaches. To determine the best method, the dendrogram was cut to calculate the compactness (CP) value of each method. The method with the smallest CP value is chosen as the most optimal. The results of the CP value of data clustering can be seen in Table 2.

Table 2. Agglomerative Hierarchical Compactness Value

Methods	Total Cluster	CP
Single linkage	3	5,0933
Complete linkage	2	5,7616
Average linkage	3	5,3755
Ward	3	5,0354

Based on Table 2, the ward method has the smallest CP value, so it is used as the best method that produces 3 clusters.

Table 3. Ward Method Cluster Members

Cluster	Cluster Member
1	Mentawai Island, Pesisir Selatan, Kabupaten Solok, Sijunjung, Tanah Datar, Padang Pariaman, Agam
2	Lima Puluh Kota, Pasaman, Solok Selatan, Dhamasraya, Pasaman Barat, Padang
3	Kota Solok, Sawahlunto, Padang Panjang, Bukitinggi, Payakumbuh, Pariaman

4.2 Clustering Categorical Data

Categorical data clustering is processed using the ROCK method. With the selection of the best results based on the largest compactness (CP*) value. The greater the CP* value, the higher the similarity of objects in one cluster. CP* values for various threshold values (θ) are presented in Table 4.

Table 4. ROCK Method Clustering Compactness Value

Threshold (θ)	CP
0,3	0,3401
0,4	0,4191
0,5	0,4191
0,6	0,4736
0,7	0,4736

Based on the analysis results, $\theta = 0{,}6$ results in 5 clusters. As the list of cluster members is shown.

Table 5. Cluster Results with the ROCK Method

Cluster	Cluster Member
1	Mentawai Island, Pesisir Selatan, Solok regency, Sijunjung, Tanah Datar, Padang Pariaman, Agam, Lima Puluh Kota, Pasaman, Pasaman Barat
2	Solok Selatan and Dhamasraya
3	Padang and Payakumbuh
4	Solok city and Bukittinggi
5	Padang Panjang and Pariaman

4.3 Clustering Mix Data

For mixed data, clustering is done through ensemble cluster analysis where the results of the clustering of numerical and categorical data are transformed into new categorical data as seen in Table 6. The next analysis process follows the steps in the ROCK method as in the previous categorical data clustering. The CP* values of the clustering results are shown in Table 7.

Table 6. Combine the Clustering Result

Numeric	Categoric
1	1
2	1
2	1
⋮	⋮
3	5

Table 7. Compactness Value of Ensemble Cluster Result

Threshold (θ)	CP
0,3	0,2807
0,4	0,3157
0,5	0,3157
0,6	0,4473
0,7	0,4473

Based on Table 6, θ=0.6 again produces 5 clusters. The composition of the members of each cluster can be seen in Table 7. Meanwhile, the characteristics are analyzed based on the average value of each numerical variable in Table 8. From these average results, the characteristics of each cluster can be identified more clearly.

Table 8. Ensemble Cluster Result

Cluster	Cluster Member
1	Mentawai Island, Pesisir Selatan, solok regency, Sijunjung, Tanah Datar, Padang Pariaman, Agam, Lima Puluh Kota, Pasaman, Solok Selatan, Dhamasraya, Pasaman Barat, and Sawahlunto
2	Padang
3	Solok city and Bukittinggi
4	Padang Panjang and Pariaman
5	Payakumbuh

Table 9. Average Value of Numerical Variables for Each Cluster

Variables	Cluster average					Sumatera
	1	2	3	4	5	Barat
Percentage of Poor Population (X_1)	21,19	41,97	4.07	3,45	7,88	17,91
Open Unemployment Rate (X_2)	4,74	10,86	4,36	5,59	4,84	5,11
School participation rate (X_3)	80,88	92,43	84,96	92,82	94,30	83,88
Literacy Rate (X_4)	99,64	99,81	99,81	99,80	99,83	99,69
Life Expectancy Rate (X_5)	69,66	74,16	74,76	72,09	74,43	70,94
Morbidity Rate (X_6)	16,49	3,81	13,71	14,82	15,31	15,28

1. Clusters 1 and 2 are areas with relatively similar welfare levels where poverty and unemployment are still major issues. These two clusters require prioritization in the economic sector.
2. Cluster 3 is an area where people's welfare is quite high but still needs improvement in the education sector. The economy in this cluster is very good because the percentage of poor people is very low and the open unemployment rate in this area is low. In the education sector, the school enrollment rate is quite low and the literacy rate is moderate. In the health sector, the life expectancy rate is high and the morbidity rate is moderate.

3. Cluster 4 is an area with a medium level of people's welfare, as it requires improvement in the health sector due to a low life expectancy and a high morbidity rate. Although the areas in this cluster have a very low poverty rate, the school enrollment and literacy rates are quite good.
4. Cluster 5 is an area with superior people's welfare compared to the other clusters, which reflects areas with good economic, health, and education conditions. In the economic sector, the percentage of poor people and the open unemployment rate in this region are low. In the education sector, the percentage of school enrollment and literacy rates in this region is superior to other clusters. In the health sector, the percentage of life expectancy in this region is high and the morbidity rate is moderate.

Table 10. *Categorical Indicator Characteristics of the ROCK Ensemble Cluster*

Variable	Category	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Population Density	Not dense	8,33%	0%	0%	0%	0%
	less dense	75%	0%	0%	0%	0%
	Quite dense	16,67%	0%	0%	0%	0%
	Dense	0%	100%	100%	100%	100%
Gini Index	low	66,67%	0%	100%	100%	0%
	Medium	33,33%	0%	0%	0%	0%
	High	0%	100%	0%	0%	100%

Table 9 shows that cluster 1 is dominated by districts/municipalities with a less dense population density with a low Gini index. Clusters 2 and 5 are districts that have a very dense population density with a high Gini index. Clusters 3 and 4 show a very dense population density but a low Gini index.

5. CONCLUSION

Based on data processing, it can be concluded that the ROCK ensemble method produces the best ensemble cluster grouping at $\theta = 0{,}6$ with 5 clusters. Cluster 1 consists of 13 provinces including Mentawai Islands, Pesisir Selatan, Solok Regency, Sijunjung, Tanah Datar, Padang Pariaman, Agam, Lima Puluh Kota, Pasaman, Solok Selatan, Dhamasraya, Pasaman Barat, and Sawahlunto. Cluster 2 consists of members from Padang City, cluster 3 consists of members from Solok City and Bukittinggi. Cluster 4 consists of members from Padang Panjang and Pariaman and cluster 5 consists of Payakumbuh City.

Cluster 5 is the cluster with the most superior level of people's welfare compared to the other clusters, especially in the aspects of economy, health, and education. The regions in this cluster have a very high population density. Meanwhile, Cluster 1 and Cluster 2 reflect regions with low levels of welfare, where poverty and unemployment are still the main problems. Cluster 1 consists mostly of districts/municipalities with low to moderately dense population density, and is dominated by a low to moderate Gini index. In contrast, Cluster 2 has a very high population density, accompanied by a high Gini index, which indicates greater economic inequality. Cluster 3 needs special attention in education, as school enrollment rates are still quite low. Regions in this cluster have a very high population density, but with a low Gini index, indicating that economic inequality is not too significant. Cluster 4 faces challenges in the health sector, as evidenced by the relatively low life expectancy and high morbidity rates. The population density in this cluster is also very high but with a low Gini index, indicating a smaller level of economic inequality than the other clusters.

REFERENCES

[1] M. S. Ummah, “Pengaruh Kesejahteraan Masyarakat Melalui Analisa Indeks Pembangunan Manusia,” *Sustain.*, vol. 11, no. 1, pp. 1–14, 2019.

[2] S. Sugiyarto, J. H. Mulyo, and R. N. Seleky, “Kemiskinan Dan Ketimpangan Pendapatan Rumah Tangga Di Kabupaten Bojonegoro,” *Agro Ekon.*, vol. 26, no. 2, p. 115, 2016.

[3] BPS, “Indikator Kesejahteraan Rakyat.”

[4] T. Hidayat, R. Ruliana, Z. Rais, and M. Botto-Tobar, “Cluster Analysis Using Ensemble ROCK Method in District/City Grouping in South Sulawesi Province based on People’s Welfare Indicators,” *ARRUS J. Math. Appl. Sci.*, vol. 3, no. 1, pp. 20–30, 2023.

[5] D. H. Setiadi, “Pemetaan Kabupaten/Kota di Jawa Timur berdasarkan Indikator Daerah Tertinggal dengan Metode Data Campuran Ensemble Rock dan SWFM,” pp. 1–143, 2018.

[6] Q. Nafisah and N. E. Chandra, “Analisis Cluster Average Linkage Berdasarkan Faktor-Faktor Kemiskinan di Provinsi Jawa Timur,” *Zeta - Math J.*, vol. 3, no. 2, pp. 31–36, 2017.

[7] S. A. Mingoti and R. A. Matos, “Clustering Algorithms for Categorical Data: A Monte Carlo Study,” *Int. J. Stat. Appl.*, vol. 2, no. 4, pp. 24–32, 2012, doi: 10.5923/j.statistics.20120204.01.

[8] R. S. Wahono, *Data Mining Data mining*, vol. 2, no. January 2013. 2023.

[9] M. Musfiani, “Analisis Cluster Dengan Menggunakan Metode Partisi Pada Pengguna Alat Kontrasepsi Di Kalimantan Barat,” *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 8, no. 4, pp. 893–902, 2019.

[10] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis.* 2007.

[11] S. Guha, R. Rastogi, and K. Shim, “ROCK: a robust clustering algorithm for categorical attributes,” *Proc. - Int. Conf. Data Eng.*, pp. 512–521, 1999.

[12] Z. He, X. Xu, and S. Deng, “A cluster ensemble method for clustering categorical data,” *Inf. Fusion*, vol. 6, no. 2, pp. 143–151, 2005.

[13] C. Suhaeni, A. Kurnia, and R. Ristiyanti, “Perbandingan Hasil Pengelompokan menggunakan Analisis Cluster Berhirarki, K-Means Cluster, dan Cluster Ensemble (Studi Kasus Data Indikator Pelayanan Kesehatan Ibu Hamil),” *J. Media Infotama*, vol. 14, no. 1, 2018.