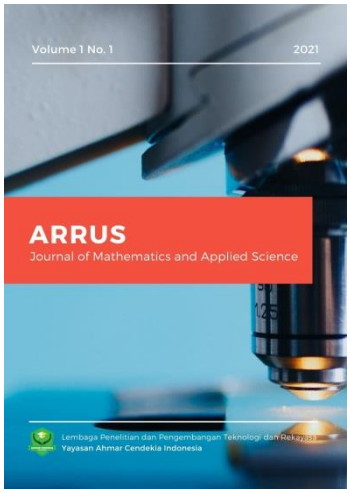
https://doi.org/10.35877/mathscience1761

OPEN ACCESS

ISSN : 2776-7922 (Print) / 2807-3037 (Online)



RESEARCH ARTICLE

Cluster Analysis Using Ensemble ROCK Method in District/City Grouping in South Sulawesi Province based on People's Welfare Indicators]

Taufiq Hidayat, Ruliana[1], Zulkifli Rais[1,*], & Miguel Botto-Tobar[2,3]

[1]Statistics Study Program, Faculty of Mathematics and Natural Sciences, Makassar State University, Indonesia
[2]Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[3]Research Group in Artificial Intelligence and Information Technology, University of Guayaquil, 090510, Guayaquil, Ecuador

*Corresponding author: Zulkifli Rais, Statistics Study Program, Faculty of Mathematics and Natural Sciences, Makassar State University, Indonesia

E-mail: zulkifli.rais89@unm.ac.id

Abstract: Cluster analysis is a data mining technique used to group data based on the similarity of attributes of object data. One of the problems that are often encountered in cluster analysis is data with a mixed categorical and numerical scale. The clustering stage for mixed data using the ensemble ROCK (Robust Clustering using links) method is carried out by combining clustering outputs from categorical and numeric scale data. The method used for categorical data is the ROCK method and the method used for numerical data is the Hierarchical Agglomerative method. The best clustering method is determined based on the criteria for the ratio between the standard deviations within the group (SW) and the smallest standard deviation between groups (SB). Based on 24 observation objects in the regencies and cities of the Province of South Sulawesi, the ROCK ensemble method with a value of 0.1 produces three clusters with a ratio value of 2,27 x10-16 based on the combination of the output results of the ROCK method and the Hierarchical Agglomerative method

Keywords Data Mining, Cluster Analysis, ROCK, Agglomerative Hierarchy, Cluster Ensemble ROCK

1. Introduction

Cluster analysis is a multivariate method that aims to group a sample of subjects on the basis of a set of changers measured into several different groups so that the same subjects are placed in the same group (Cornish, 2007). According to Simamora (2005), cluster analysis is a statistical analysis technique aimed at placing a set of objects into two or more groups based on the similarities of objects on the basis of various characteristics.

The problem that is often encountered in cluster analysis is that the data used are of mixed type (numerical and categorical). A method often used in grouping mixed data is to transform categorical data into numerical or vice versa. Grouping with cluster ensemble can handle mixed data by grouping numerical and categorical data separately using the grouping method for each of these data, then the grouping results are combined and viewed as categorical data which is then grouped by grouping methods for categorical data (Dewangan et al., 2010).

For numeric data, grouping objects can use both hierarchy and non-hierarchical methods. The hierarchy method is used when the researcher does not specify the desired number of

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

clusters, while the non-hierarchy method is used when the number of clusters is determined by the researcher. For grouping categorical data, hierarchical and non-hierarchical methods can also be used, in addition to other methods that can be used, namely the ROCK (Robust Clustering Using Links) method (Guha et al. , 1999) .

The welfare of the people in each region is different. Therefore, it can be done to group the regions in South Sulawesi Province to see the welfare conditions of rakyat in one area with other regions so that it can assist the provincial government ofSouth Sulawesi in compiling and determining development priorities. The grouping methodthat can be used to handle mixed data is the ensemble cluster method by grouping numerical and categorical data separately using a grouping method for each of these data, then the grouping results are combined and viewed as categorical data which is then grouped with a grouping method for categorical data (Belinda et al., 2019).

2. Literature Review

2.1. Numerical Data and Categorical Data

Data is a fact that is processed into information. Data used in statistical analysis (statistical data) based on the type of variables are grouped into two, namely numerical data and categorical data. Numerical data is data with quantitative variables that produce numerical information. Numerical data can be grouped into two, namely discrete data (enumeration results), for example the number of children in the family and continuous data (measurement results), for example height and weight. Categorical data is data with qualitative variables resulting from classifying or helping data (attribute data). Agresti (2007) states that categorical data have a measurement scale consisting of a set of categories, for example political philosophy that can be measured as a liberal category, a moderate category, or a conservative category.

2.2. Cluster Analysis

Cluster analysis is a method in multivariate analysis to group n observations into C groups (C n ≤) based on their characteristics, then the results of the analysis can be considered as a reference for grouping new data on pre-formed clusters. Cluster analysis is one of the primitive methods, so there is no need for assumptions to be used to group data, because grouping is based on similarity and inexperience (Johnson & Winchern, 2007). The data structure for cluster analysis with n observations and m variables is shown in Table 1 sas follows.

Table 1. Cluster Analysis Data Structure

Observations to-	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	⋯	$\mathbf{X_m}$
1	X_{11}	X_{21}	X_{31}		X_{m1}
2	X_{12}	X_{22}	X_{32}		X_{m2}
3	X_{13}	X_{23}	X_{33}		Xm_3
⋮	⋮	⋮	⋮		⋮
N	X_{1n}	X_{2n}	X_{3n}	⋯	X_{mn}

2.3. Agglomerative Hierarchy Method

1) Single Linkage

Single linkage is a grouping based on the closest distance or many similarities. If two objects are separated by a close distance then the objects are combined into one group and so on. If d_{UV} is a measure of indifference between the Uth group and the Vth group then, the measure of distance used between the group (UV) and W is as in the following equation (Johnson & Winchern, 2007).

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

where,

d_{UW} :distance of group *U* and group W

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

d_{VW} :distance of group V and group W
$d_{(UV)W}$:minimum distance between UV and W groups.

2) Complete Linkage

Complete linkage is a method where clusters are formed by grouping objects that have the farthest distance or little similarity. If two objects are separated by a long distance then they are combined into one group and so on. If d_{UV} is a measure of indifference between the U-th group and the Vth group then, the measure of distance used between the th group(UV) and the W-th is as in the following equation (Johnson & Winchern, 2007).

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

where,

d_{UW} :distance of group U and group W
d_{VW} :distance of group V and group W
$d_{(UV)W}$:distance maximum between the UV and W groups.

3) Average Linkage

Average linkage is a method by which a cluster is formed based on the average value of the distance of all individuals in one group with the average distance of all individuals in another group. If d_{UV} is a measure of the inexperience between the U-th group and the Vth group then, the measure of distance used between the th group(UV) and the W-th can be imitated as in the following equation (Johnson & Winchern, 2007).

Average linkage is a method by which a cluster is formed based on the average value of the distance of all individuals in one group with the average distance of all individuals in another group. If d_{UV} is a measure of the inexperience between the U-th group and the Vth group then, the measure of distance used between the th group(UV) and the W-th can be imitated as in the following equation (Johnson & Winchern, 2007).

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

where,

d_{UW} :distance of group U and group W
d_{VW} :distance of group V and group W
N_{uv} :number of observations in the UV group
N_W :number of observations in group W
d_{qr} :the distance between the q-th observation in the UV group and the r-th observation in the W group.

2.4. ROCK Method

Traditional methods that use the concept of distance between points for grouping categorically type variables are considered inappropriately used in categorical data. Therefore, a method of grouping agglomerative hierarchies was developed which is used for categorical data, namely the ROCK (Robust Clustering using links) method (Guha. , et al, 2000).

In the ROCK method , a new concept was formed, namely link is used to measure the similarity / proximity between a pair of data points. Observations that have a high degree of relationship are combined in one group, while low ones are separated from the grouped data (Guha, et al, 2000).

Grouping on the ROCK algorithm will stop when in the following circumstances.

a) The number of expected groups has already been met, or
b) There are no links between groups.

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Grouping by the ROCK method consists of several steps. Here are the steps for grouping with the ROCK method (Guha, et al, 2000)

a) Specifies the initialization for each data point as the cluster initially.
b) Calculating Similarity

The measure of similarity between the i-th and jth observation pairs is calculated by the formula in the following equation (Guha, et al, 2000).

$$\mathrm{sim}(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}, i \neq j$$

c) Menentukan Tetangga (Neighbors)

The observation of X_i and X_j is expressed as neighbors if the value of sim $(X_i, X)_j \geq \theta$. The threshold value of θ used usually ranges from 0 to 1. The threshold value of θ can be determined by adjusting the existing data.

d) Counting Links

Links (X_i, X_j) between objects are obtained from the number of common neigbor between X_i and X_j. If the link values (X_i, X_j) are large then most likely X_i and X_j are in the same clusters

e) Goodness Measure

In the ROCK algorithm, the merger of groups is basedon a measure of goodness between groups. Goodness measure is an equation that calculates the number of links divided by the possibility of links being formed based on the size of their groups. The goodness measure can be calculated by the formula in the equation as follows (Dutta et al. , 2005).

$$g\left(C_i, C_j\right) = \frac{\mathrm{link}(C_i, C_j)}{(n_i + n_j)^{1+2f\theta} - n_i{}^{1+2f\theta} - n_j{}^{1+2f\theta}}$$

with, $\mathrm{link}(C_i, C_j)$ = is the number of links of all possible pairs of objects present in $\sum_{Xi \in ci, Xj \in cj} \mathrm{link}\,(X_i, X_j)$$C_i$ and C_j. n_i and n_j is the number of members in the i-th group and the j-th group.

2.5. Cluster Ensemble

Grouping of categorical and numeric mixed data can be done by dividing the data into purely categorical and purely numerical data. Numeric data and categorical data are grouped according to data types separately. The results of the grouping are then combined using the cluster ensemble method. Cluster ensemble is a method of combining several results from different grouping algorithms so that a combined solution is obtained as the final solution. The result of the individual grouping algorithm is categorical and therefore the case of the ensemble cluster can be viewed as a case of grouping of categorical data.

Ensemble grouping consists of two stages of the algorithm. The first stage is to group with several algorithms and save the results of the grouping. The second stage is to use the consensus function to determine the final cluster of the groups already obtained in the first stage (Hee et al. , 2002)

2.6. Performance of Clustering Results

Grouping performance measurements are used to determine the validity of a grouping. A good group is to have a high homogeneity in the group and a high homogeneity between groups (Hair. , et al, 2001). Performance measurements in numerical data can be known

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

from the ratio of S_W and S_B. The value of the standard deviation within the group or within (S_W) can be formulated in the following equation (Bunkers & James, 1996).

$$S_W = \frac{1}{C} \sum_{C=1}^{C} S_C$$

where

Sc :save n standard c-th group

C : the number of groups formed

The value of the standard deviation between groups or between (S_B) can be formulated in the following equation (Bunkers and James, 1996)

$$S_B = \left[\frac{1}{C-1} \sum_{C=1}^{C} \left(\bar{X}_c - \bar{X} \right)^2 \right]^{1/2}$$

where

$\bar{X}_c$: :c-th group mean

$\bar{X}$: :overall average of the group k

2.7. People's Welfare

The welfare of the people has always been an interesting topic to discuss. As in every country the main goal in development is the improvement of the welfare of the people. As is the case in Indonesia, the welfare of the people is one of the state goals stated in the Preamble to the 1945 Constitution paragraph IV. The welfare of the people is basically a condition whose form is dynamic or in other words its quantitative value will never stop because it will continue to change along with the development of human life needs (Alwi & Hasrul, 2018).

The total population of South Sulawesi Province in 2015 was 8,520,304 people with a population density of about 186.18 people per square kilometer. The population continues to increase year-on-year by 1.8% or 10.1 million. When the population continues to grow, it means that the government must also continue to increase the number of decent living facilities for its people. In addition to population problems, the unemployment rate in South Sulawesi Province is also still high. According to 2015 data, the open unemployment rate reached 218,311 people or about 5.95%. Compared to the previous year with the open unemployment rate amounting to 212,857 people or about 5.08%. The implication is that the open unemployment rate will increase even more, if there is no change in strategy in job creation. Therefore, one solution to overcome this is to identify characteristics based on the level of welfare of the people of each region so that the government can take or decide on good/ targeted policies and strategies in development (Alwi & Hasrul, 2018).

3. Research Method and Materials

3.1. Types of Research

This research uses descriptive statistical methods in the analysis of the characteristics of the data used, as well as with the approach or exploration of data on people's welfare indicators. In this ROCK ensemble method, numerical data is grouped with the agglomerative hierarchy method while in the categorical data, the ROCK method is used.

3.2. Data Collection Techniques

The data used in this study is secondary data, namely data obtained indirectly or obtained from existing sources. The data is sourced from the publication of the Central Statistics Agency of South Sulawesi Province in 2020.

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

3.3. Operational Definition of Variables

The description of the variables to be used in this study is given in Table 2 as follows.

Table 2 Research Variables

Variable	Information	Data Type
X1	Percentage of Poor Population	Numerical
X2	Open Unemployment Rate	Numerical
X3	Literacy Rate	Numerical
X4	Life Expectancy	Numerical
X5	Population Density	Category
X6	Gini Index	Category

3.4. Data Analysis Techniques

The data analysis techniques in this study are:

a) Make descriptive statistics on the variables of the percentage of the poor, open unemployment rate, literacy rate, life expectancy, population density and gini index.

b) Grouping variables of people's welfare indicators in Provinsi Sulawesi Selatan using the ROCK ensemble method.

c) In grouping with the ROCK ensemble method is as follows.

1) Divide research variables into categorical data and numerical data. Numerical data consists of the percentage of the poor population, open unemployment rate, literacy rate, life expectancy figures, while the categorical variable consists of population density and gini index
2) Clustering numerical data with the hierarchy method and the distance size used is euclidean distance. The hierarchy methods used are single linkage, complete linkage and average linkage.
3) Calculates the best grouping performance based on the S_W/S_B ratio value of the grouping results by the single linkage, complete linkage and average linkage methods.
4) Clustering categorical data with rock method with value
5) Determines the optimum number of groups based on the minimum S'W/S'B ratio value based on the rock method grouping results at each θ value.
6) Clustering the combined data using the Ensemble ROCK method threshold value (θ) of 0.1; 0.2; 0.3; 0.4; 0.5; 0,6; 0,7; 0.8 and 0.9. In grouping with the ROCK method, an R-project statistics program (Package R) with the name package cba will be used.
7) Determines the number of optimum groups based on the minimum S'W/S'B ratio value based on the results of grouping the ROCK Ensemble method on each θ value.
8) Make descriptive statistics of data based on the grouping results that have been obtained.

d) Interpret the results of the analysis and draw conclusions.

4. Results and Discussion

4.1. Descriptive Analysis

Based on Table 3 of descriptive statistics for categorical changes in Population Density obtained from the 2020 catalog of people's welfare indicators sourced from the Central Statistics Agency Provinsi Sulawesi Selatan explained that there are 41.66% moderate density levels of districts and cities in Sulsel, and low and high density levels of residents of districts and cities in South Sulawesi 29.17% each.

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 3 Descriptive Statistics of Categorical Changes in Population Density

Population Density	Frequency	Onrsentase(%)
Low	7	29,17
Keep	10	41,66
Tall	7	29,17
Sum	24	100

Table 4 Descriptive Statistics of Categorical Changes of the Gini Index

Gini Index	Frequency	Onrsentase (%)
Keep	14	58,33
Tall	10	41,67
Very High	0	0
Sum	24	100

Based on Table 4 of descriptive statistics for categorical changes in the gini index catalog of people's welfare indicators in 2020 sourced from the Central Statistics Agency Provinsi Sulawesi Selatan explained that there was a 58.33% moderate level of the gin index and a high level of the gini index of 41.67%.

The descriptive statistics for numerical data are as follows:

Table 5 Descriptive Statistics of Numerical Changes

Changers	N	Minimum	Maksimum	Average
Percentage of Poor Population	24	4,54	14,58	9,42
Open Unemployment Rate	24	2,31	15,92	4,96
Literacy Rate	24	85,24	98,59	91,89
Life Expectancy	24	66,39	73,39	68,49

Based on Table 5, the descriptive analysis for the numerical changer explains that the 2020 people's welfare index data is as many as 24 districts and cities. It was explained that the average percentage of poor people is 9.42% with the highest percentage of poor people, which is 14.58% and the lowest percentage of poor people is 4.54%. Thepercentage of poor people followsthe open unemployment rate with an average open unemployment rate of 4.96 with the highest open unemployment rate of 15.92 and the lowest open unemployment rate of 2.31 in one year. The literacy rate has an average of 91.89 with the highest literacy rate value of 98.59 and the lowest literacy rate of 85.24 in one year and the lowest life expectancy of 66.39 years.

4.2. Clustering

Clustering consists of 3 stages based on data separation, namely categorical data, numerical data, and mixed data.

4.2.1. Clustering data kategorik

In this study,θ a value of 0.1 was used; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8 and 0.9. The value is determined by the researcher adjusted for the distance of the observation object and the expected clustering results. The result of clustering the ROCK method. The best clustering results are determined from the smallest S_W and S_B ratio values. Based on Table 6, it explains that the ratio and smallest values are = 0.3 with a value of 0.189 which is the best θcluster result in the ROCK method for categorical data.

The best cluster results for the ROCK method with a value of 0.3 which produces 3 θclusters, namely cluster 1, cluster 2, and cluster 3 with members of each cluster are shown in Table 7.

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 6 Ratio Values of Rock Method Cluster Results

Value θ	Ratio S_W dan S_B
0.1	0.302
0.2	0.255
0.3	**0.189**
0.4	0.252
0.5	0.252
0.6	0.252
0.7	0.252
0.8	0.252
0,9	0.252

Table 7 Cluster Results of rock method with Value θ = 0.1

Cluster	Cluster members
Cluster 1	Bulukumba, Sinjai, Maros, Pangkep, Bone, Soppeng, Wajo, Toraja Utara
Cluster 2	Bantaeng, Jeneponto, Takalar, Gowa, Barru, Sidrap, Pinrang, Enrekang, Luwu, Luwu Timur, Mak assar, Pare-pare, Palopo
Cluster 3	Selayar, Tana Toraja, Luwu Utara

4.2.2. Clustering Numeric data

The first stage carried out in the agglomerative hierarchy method is to declare (initialize) each object of observation as a group with a single member. The next stage is to form a matrix of distances between observation objects. The distance used in this study is the euclidean distance calculated using the equation . The distance obtained from the 24 observation objects is expressed in a matrix **d** measuring 24 x 24.

The number of clusters formed based on dendograms for all three methods is two clusters to five clusters. After the clustering results are obtained, the next stage is to calculate the validity index of the size of the optimum number of clusters using the dunn index. The estimation of the number of optimum groups is carried out by looking at the largest value of the cluster validity index.

Table 8 Dunn Index Value Results

Number of Clusters	Single Linkage	Complete Linkage	Average Linkage
2 Cluster	0.397	0.247	0.276
3 Cluster	**0.496**	0.270	0.352
4 Cluster	0.363	0.308	0.356
5 Cluster	0.328	**0.415**	**0.365**

Based on Table 8, the results of the dunn index validity value show that the optimum clusters formed for the three methods are 3 clusters for the single linkage method, 5 clusters for the complete linkage method and 5 clusters for the average linkage method. After obtaining the optimum number of clusters, the best clustering method of the three methods is then selected based on the ratio values of S_W and S_B the smallest of each method. The formed ratio values are presented in Table 9.

Table 9 Ratio Values Cluster Results agglomerative hierarchy .

	Nilaand S_W	Nilai S_B	Ratio
Single Linkage	1.524	1.044	1.459
Complete Linkage	0.227	0.764	0.297
Average Linkage	0.129	0.858	**0.151**

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

The best method is obtained, namely Average Linkage which has the smallest S_W and S_B ratio value of 0.151. This shows that clustering the numerical data of the Average Linkage method with 5 clusters is appropriate clustering for the agglomerative hierarchy method.

The following is a table of cluster members for the average linkage method.

Table 10 Cluster Members Average linkage method

Cluster	Cluster Members
Cluster 1	Selayar, Bulukumba, Bantaeng, Takalar, Gowa, Sinjai, Maros, Pangkep, Barru, Bone, Soppeng, Wajo, Pinrang, Enrekang, Luwu, North Luwu, East Luwu
Cluster 2	Jeneponto
Cluster 3	Tana Toraja, North Toraja
Cluster 4	Makassar
Cluster 5	Pare-Pare, Palopo

4.2.3. Mixed Data Clustering

The first stage in analyzing rock ensemble clusters for mixed data is to cluster each type of data using their respective methods. The clustering results for the categorical data obtained using the ROCK method are expressed as output 1, as well as the clustering results for numerical data obtained using the agglomerative hierarchy method are expressed as output 2. Next, the two clustering output results are stated as categorical changes (ensemble stage) which are then clustered using the ROCK method.

In this analysis, several values are used, such as in clustering categorical data, namely a value θ of 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8 and 0.9. The best clustering results are determined from the ratio and smallest values. The ratio value and can be seen in table 11, as follows.

Table 11 Ratio Values cluster results of rock ensemble method

Value θ	Ratio Value
0.1	**$2{,}27 \times 10^{-16}$**
0.2	$1{,}13 \times 10^{-01}$
0.3	$1{,}79 \times 10^{-01}$
0.4	$1{,}47 \times 10^{-01}$
0.5	$1{,}47 \times 10^{-01}$
0.6	$1{,}47 \times 10^{-01}$
0,7	$1{,}47 \times 10^{-01}$
0.8	$1{,}47 \times 10^{-01}$
0.9	$1{,}47 \times 10^{-01}$

Table 11 shows that the smallest ratio value is clustering with a value of = 0.1 with a ratio of $S\theta_W$ and S_B worth 2.27×10^{-16}. The value indicates that the standard deviation in the cluster is $2{,}27 \times 10^{-16}$ of the standard deviation between clusters. In other words, the variance of data in clusters provides a smaller deviation value than the variance between clusters.

The best cluster result for the value = 0.1 is the first θrunning result that produces 3 clusters, namely cluster 1, cluster 2, and cluster 3 with members of each cluster shown in Table 12 as follows:.

Table 12 Cluster Results of rock Ensemble Method with Value = 0.1

Cluster	**Cluster** Members
Cluster 1	Tana Toraja, Toraja Utara
Cluster 2	Selayar, Bulukumba, Sinjai, Maros, Pangkep, Bone, Soppeng, Wajo, Luwu Utara
Cluster 3	Bantaeng, Jeneponto, Takalar, Gowa, Barru, Sidrap, Pinrang, Enrekang, Luwu, Luwu Timur, Makassar, Pare-pare, Palopo

4.3. Interpretation of Clustering results

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

The results of cluster 1 are a group of 2 districts, namely Tana Toraja, Toraja Utara. This group has an average valueof 12.05% of the poor population, an open unemployment rate of 2.88, a literacy rate of 90.91 and a life expectancy of 73.34 years. The prominent population density in this group is the moderate population density of 3.47%. Then the low-level population density and the high-level population density of 2.43%. The dominant gini index is the medium-level gini index, which is 4.86%, and the high-level gini index is 3.13%.

The results of cluster 2 are a group of 9 districts, namely Selayar, Bulukumba, Sinjai, Maros, Pangkep, Bone, Soppeng, Wajo, and Luwu Utara. This group has an average valueof 10.10% of the poor population, an open unemployment rate of 3.88, a literacy rate of 91.45 and a life expectancy of 67.99 years. Based on the results of the dominant population density in this group, it is a moderate population density of 15.62%. Then the low-level population density and the high-level population density of 10.49% each. Then and the dominant gini index is the medium-level gini index which is 21.87% and the high-level gini index is 1 5.62%.

The results of cluster 3 are a group consisting of 13 districts and cities, namely Bantaeng, Jeneponto, Takalar, Gowa, Barru, Sidrap, Pinrang, Enrekang, Luwu, Luwu Timur, Makassar, Pare- Pare, Palopo. This group has an average value of the percentage of the poor population of 8.54.%, an open unemployment rate of 6.07, a literacy rate of 92.35 and a life expectancy of 69.92 years. Based on the results, the prominent level of population density in this group is the moderate population density of 22.57%. Then the population density is low and high-level is 15.8%. and the most dominating gini index is the medium-level index , which is 31.6%.

5. Conclusion

For Conclusions, the main conclusions of the study may be presented in a short Conclusions section, which may stand alone.

1. The results of clustering categorical data using the ROCK method with values of 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8 and 0.9. Based on the ratio values of SθW and SB smallest shows that value = 0.3 is the best value in θcluster analysis for categorical data.
2. The results of clustering numerical data using the agglomerative hierarchy method show that the best method for numerical data is the average linkage method with 5 optimum clusters.
3. The results of clustering categorical and numerical mixed data using the ROCK ensemble method with 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8 and 0.9 show that the value = 0.1 is the best value in θcluster analysis for categorical and numerical mixed data.

References

Ariska, N. (2017). Analisis Cluster dengan Metode Ensemble ROCK untuk Data Berskala Campuran Kategorik dan Numerik (Kasus : Mahasiswa Aktif Program Studi Statistika). Makassar: Universitas Negeri Makassar.

Alwi, W., & Hasrul, M. (2018). Analisis Klaster Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, *6*(1), 35.

Badan Pusat Statistik. (2020). *Indikator Kesejahteraan Rakyat*. BPS: Makassar

Belinda, N. S., HG, I. R., & Yozza, H. (2019). Penerapan Analisis Cluster Ensemble Dengan Metode Rock Untuk Mengelompokkan Provinsi Di Indonesia Berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal Matematika UNAND*, *8*(2), 108.

Bunkers, M. J., Miller, J. R., & DeGaetano, A. T. (1996). Definition of climate regions in the Northern Plains using an objective cluster modification technique. *Journal of Climate*, *9*(1), 130-146

Cornish, R. (2007). *Statistics*: 3.1 *Cluster Analysis. Leicestershire, UK: Loughborough University Mathematics Learning Support Centre*.

https://doi.org/10.35877/mathscience1761

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Dewangan, R. R., Sharma, L. K., & Akasapu, A. K. (2010). *Fuzzy Clustering Technique for Numerical and Categorical Dataset. International Journal on Computer Science and Engineering (IJCSE), NCICT Special Issue*, 75-80.

Dewi, A. (2012). Metode Cluster Ensemble Untuk Pengelompokan Desa Perdesaan di Provinsi Riau.

Guha, S., Rastogi, R., & Shim, K. (1999). ROCK: A Robust Clustering Algorthim for Categorical. In *International Conference on Data Engineering* (hal. 512–521).

Hair, Joseph F. , Black, Jr, William C. Babin, Barry J. & Anderson, R. E. (2010). Pearson - Multivariate Data Analysis, 7/E - Joseph F. Hair, Jr, William C. Black, Barry J. Babin & Rolph E. Anderson. *Pearson New International Edition*, 816.

Han, J., & Kamber, M. (2001). Data mining: Data Mining Concepts and Techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*.

Hee, Z., Xu, X. i., & Deng, S. (2002). Clustering Mixed Numeric and Categorical data: A Cluster Ensemble Approach. China: Harbin Institute of Technology.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (Vol. 6). London, UK:: Pearson.

Kantardzic, M. (2011*). Data Mining: Concepts, Models, Methods and Algorithms*, A John Wiley & Sons. *Inc. Hoboken, New Jersey*.

Putri, K. A. K. (2021). Pengelompokan Daerah di Jawa Tengah Berdasarkan Indikator Kesejahteraan Menggunakan Metode *Ensemble Cluster Rock* (*Doctoral dissertation*, Universitas Muhammadiyah Semarang).

Rahayu, D. P. (2013). Analisis Karakteristik Kelompok dengan Menggunakan Pendekatan Cluster Ensemble. *Jurnal Matematika Sains dan Teknologi*, *14*(1), 01–10

Rencher, A. C. (2002). A Review of "Methods of Multivariate Analysis, Second Edition." In *IIE Transactions* (Vol. 37, Nomor 11).

Sari, I. A., & Saputro, D. R. S. (2021). Algoritme Quick RObust Clustering using linKs (QROCK) untuk Clustering Data Kategorik. *PRISMA, Prosiding Seminar Nasional Matematika*, *4*, 640–644.

Simamora, B. (2005). *Analisis multivariate pemasaran*. Gramedia Pustaka Utama.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). Anomaly Detection. *Introduction to Data Mining; Goldstein, M., Harutunian, K., Smith, K., Eds*, 651-680.

Tyagi, A., & Sharma, S. (2012). Implementation Of ROCK Clustering Algorithm For The Optimization Of Query Searching Time. *International Journal on Computer Science and Engineering (IJCSE)*, *4*(05), 809–815.

Wulandari, L. (2019). Evaluasi Daerah Tertinggal di Jawa Timur Berdasarkan Indikator Kementrian Negara Pembangunan Daerah Tertinggal (KPDT) menggunakan *Ensemble Robust Clustering Using Link (ROCK)* (*Doctoral dissertation*, UIN Sunan Ampel Surabaya).

Yulianto, S., & Hidayatullah, K. H. (2016). Analisis Klaster Untuk Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat. *Statistika*, *2*(1), 56–63.