The comparison of k-modes clustering and ROCK clustering to the poverty indicator in Samadua Subdistrict, South Aceh

H Sofyan¹, M Iqbal¹,²*, M Marzuki¹, M Muhammad²
¹Statistics Department, Syiah Kuala University, Indonesia
²Regional Planning and Development Agency of Aceh Province (Bappeda), Indonesia

*E-mail: m.iqbal@mhs.unsyiah.ac.id

Abstract. The cluster methods including k-modes clustering and ROCK clustering are rarely used for analyzing categorical data. Much of the research data, however, are categorical, such as the poverty indicator dataset. Many studies have examined each of these two methods, but there is still limited study for comparing them. This study compared between k-modes clustering and ROCK clustering on a dataset of poverty indicators at a category scale. There were 1449 households with the lowest 40% income in Samadua Sub-district, South Aceh Regency were used as data in this study. The result of this research was that the optimal number of clusters for k-modes clustering was 2. Similarly, the optimal number of clusters for ROCK clustering was 2 with a threshold value (θ) of 0.29. The SW/SB and R-Squared values for the k-modes clustering were 0.0149 and 0.7558, respectively, compared to 0.0682 and 0.5732 for ROCK clustering. From these two values, it can be concluded that the k-modes clustering is better than ROCK clustering. The significant difference between the two clusters based on deciles of households with the lowest 40% income of the k-modes clustering is ownership of defecation facilities, ownership of television and motorbikes.

1. Introduction
Grouping data (observation) into a group / cluster based on the same characteristics which is similar, is the method of statistics which commonly used nowadays. This grouping method was conducted by the basis of similarity from each observation. This grouping will be more convenient when the relationship between each characteristic has been analyzed.

The Cluster Method is based on the type of dataset that will be used. If the dataset is numeric, therefore the cluster method that is used such as hierarchical clustering, non-hierarchical clustering (k-means, k-median), and fuzzy clustering (fuzzy c-means). Whereas if the dataset is category type, therefore the cluster method that can be used are cluster k-modes and Robust Clustering using Links (ROCK clustering).

The k-modes is one of the clustering methods that is used for the category dataset. k-modes is modification method from K-means method, which in k-means method the data type should be numeric ratio or interval. Dataset of category type is commonly based on the certain criteria in conducting the clustering. Those criteria will be later formed the categories in each variable of indicator [1].

Besides the k-modes, there is also ROCK clustering method that is commonly known for the category type of data. It is compatible for the extreme data (outlier). This method is common for
applying the link concept as the similarity measurement in the cluster forming. The link value is determined by the number of threshold ("θ") where it range between 0 until 1 [2, 3].

One of the highest poverty rate provinces in Indonesia is Aceh. The poverty in Aceh in 2018, according to The Central Bureau of Statistic (BPS) of Aceh, carry through 15.97% (839 thousand inhabitants). This number has decreased from 2017. Moreover, this number is the lowest poverty rate since 2002. However, the poverty number in Aceh remains distant above the national number of povertys, which is 9.82% [4].

South Aceh is one of the 23 Districts/Cities in Aceh Province. This district has poverty rates as much as 14.07% (32.51 inhabitants) on 2017. The number remains in higher position than the national poverty percentage, yet under the Aceh Province, namely 15.97%. South Aceh become the 5 lowest districts according to the poverty percentage rate of Aceh Province [5].

Samadua is one of the subdistricts in South Aceh which located close to the district capital, Tapaktuan. Samadua is the seventh most populated subdistrict in 2017, the population reached to 15.337 inhabitant. The growth of the population in Samadua is the third least, it was 0.54 in 2016 until 2017 [4]. The number of households that earn income that least than 40% in Samadua reach 1449 inhabitants, with the total population of 5608 inhabitant (Integrated Database, 2015).

Another poverty study was conducted by Ulfah and ‘Uyun [6]. It concerned on the group determination of impoverished household (RTM) in Special Region of Yogyakarta by using the k-means and fuzzy c-means (FCM) methods. There were 15 poverty indicator which were examined which was separated into 10 aspect, namely the income, food, clothing, residence, clean water, health, education, property, lighting, and inhabitant number. The obtained result from the study is that k-means method was better compared to the FCM method since the obtained accuracy value for k-means was 83.33% (FCM accuracy value was 50.00%). This conclusion is similar to the conducted research by Hidayat et al. [7]. The k-means method is more proper to be used for grouping in poverty in Central Java.

This research was conducted in two part, namely descriptive analysis and group determination. The descriptive analysis was conducted to analyze the figure of poverty in Samadua Subdistrict, South Aceh District. The group determination was conducted by determining the optimum group and the best group by using k-modes and ROCK clustering for the grouping of poverty household numbers. This research was expected to provide the statistical information for the poverty household numbers grouping in Samadua Subdistrict, South Aceh District.

2. Materials and method
The data that had been used for this research was collected from Integrated Database (BDT in Indonesia) of The Regional Planning and Development Agency (BAPPEDA) of Aceh Province in the year of 2015. Those are 1449 household data which had 40% lowest income in Samadua Subdistrict, South Aceh District. The data consist of poverty indicators which were assigned by BPS. The poverty indicators that were used for this research is only 12 indicators, namely the floor type on shelter (X1), the type of wall of the shelter (X2), the disposal facility usage (X3), the main lighting source (X4), the main source of drinking water (X5), the main fuel for cooking (X6), the ownership of KIS/BPJS Health Insurance/Jamkesmas (X7), the income source of the household head (X8), the education background of the household head (X9), the ownership of refrigerator (X10), the ownership of television (X11), and the ownership of motorcycle bike (X12).

The data analysis steps which were conducted started from the data coding for each variable, descriptive analysis, optimal clusters determining from k-modes clustering and ROCK clustering based on the SW/SB ratio value and R-squared, conducting the k-modes clustering and ROCK clustering comparison method based on the SW/SB ratio value and R-squared [8], and conducting the further analysis according to the optimal method obtained.
3. Discussion analysis

3.1. Descriptive analysis
This research was using 12 predictor variables from the poverty indicators, village variables, and welfare rate variables. Entire predictor variables were analyzed descriptively by observing the mode from the 40% lowest income households in Samadua Subdistrict, South Aceh District, as shown in Table 1. The 40% lowest income households generally had the unplastered floor/red brick floor with the percentage of 88.06% (1276 households). For the type of the shelter wall, it is dominated by the regular wall as much as 83.85% (1215 households).

Table 1. Modes for predictor variables

| Variables                                           | Mode                     | Numbers | Percentage (n=1449) |
|-----------------------------------------------------|--------------------------|---------|--------------------|
| Floor type of the shelter                           | Unplastered floor / red bricks | 1276    | 88.06%             |
| Wall type of the shelter                            | Regular wall             | 1215    | 83.85%             |
| Disposal facility usage                             | Private use              | 687     | 47.41%             |
| The air lighting source                             | PLN Electricity          | 1418    | 97.86%             |
| The main drinking water source                       | Protected water well     | 510     | 35.20%             |
| The main cooking fuel                               | Firewood                 | 912     | 62.94%             |
| Ownership of KIS/ BJS Health Insurance / Jamkesmas | No                       | 869     | 59.97%             |
| The main head of household income source            | Daily construction worker| 650     | 44.86%             |
| Head of household education background              | Elementary / Extraordinary disabling elementary school | 839   | 57.90%             |
| Ownership of refrigerator                           | No                       | 1075    | 74.19%             |
| Ownership of television                             | Yes                      | 950     | 65.56%             |
| Ownership of motorcycle bike                         | No                       | 808     | 55.76%             |

The 40% lowest income households generally had their own facility of disposal (47.41%), they also used PLN electricity as the main source of lighting (97.86%), used the protected water well to as the source of drinking water (35.20%), used firewood as the main cooking fuel (62.94%). The 40% lowest income households also did not have any access to KIS/ BPJS Health Insurance/ Jamkesmas (59.97%), did not own any refrigerator (74.19%), did not own any motorcycle bike (55.76%), and did not own any television (65.56%). Whereas for the household main income source variable, approximately they worked as daily constructional workers (44.86% household), and their highest education background only reached Elementary/ Extraordinary elementary school for disable person (57.90%).

3.2. Cluster analysis

3.2.1. K-modes clustering
The clusters numbers determination on k-modes clustering was conducted by observing the ratio value from standard deviation on the group (SW) and the standard deviation between group (SB). The cluster numbers in this research range between 2 until 10. Then, the comparison of ratio value from each cluster was conducted. The clusters numbers with the least ratio is the optimal cluster numbers. Ratio between SW and SB from each cluster number was shown on Table 2.
The determination of optimum cluster number is also regarded to the R-squared value. R-squared is the ratio of squares between groups numbers (SSB) with the total numbers of squares (SST). If the R-squared value resulted is close to 1, the better formed clusters. R-squared which was formed from each cluster is shown in Table 2.

The optimal cluster numbers are 2 clusters (k=2), it has the least SW/SB ratio, namely 0.1049. This is also supported by the R-squared value which is high, namely 0.7558 or 75.58%. It can be concluded that if the cluster numbers are 2 clusters (k=2), it is the optimal clusters numbers that are formed.

### 3.2.2. ROCK clustering

The determination of optimal cluster numbers on ROCK clustering is similar to the determination of optimal cluster numbers on k-modes clustering, namely by determining the SW/SB ratio with the R-squared. On ROCK clustering, first of all, we need to determine the maximum number of clusters which available to be formed. The optimum cluster numbers will be formed after the threshold or theta (θ) value is given. In this research, the maximum numbers of clusters formed are 10 clusters. Whereas, for the threshold are 0.01, 0.05, 0.10, 0.11, 0.13, 0.14, 0.15, 0.16, 0.17, 0.20, 0.21, 0.22, 0.23, 0.24, 0.25, 0.26, 0.27, 0.29, and 0.30.

The threshold values of 0.31, 0.32, 0.33, 0.34, and 0.35 were not analyzed due to the similar result as the 0.30. The threshold 0.29 has the least SW and SB ratio, namely 0.0682 as Table 3 summarizes. The R-squared value which was formed is 0.5732 or 57.32%. This R-squared value was the highest R-squared which was approached to 1. By the threshold value of 0.29, hence the optimal cluster numbers formed are 2 clusters.

### Table 2. Ratio value of SW/SB and R-squared based on the cluster numbers on k-modes clustering

| No. | Cluster numbers (k) | Ratio (SW/SB) | R-squared  |
|-----|---------------------|---------------|------------|
| 1   | 2                   | 0.0149        | 0.7558     |
| 2   | 3                   | 0.0304        | 0.5999     |
| 3   | 4                   | 0.0355        | 0.6217     |
| 4   | 5                   | 0.0498        | 0.5274     |
| 5   | 6                   | 0.0417        | 0.6656     |
| 6   | 7                   | 0.0928        | 0.3259     |
| 7   | 8                   | 0.0872        | 0.3901     |
| 8   | 9                   | 0.0960        | 0.3762     |
| 9   | 10                  | 0.1032        | 0.3699     |

### Table 3. SW/SB ratio and R-squared value based on the cluster numbers in the ROCK clustering

| No. | Threshold (θ) | Number of formed clusters (k) | SW/SB  | R-squared  |
|-----|--------------|-------------------------------|--------|------------|
| 1   | 0.01         | 2                             | 0.3518 | 0.0478     |
| 2   | 0.05         | 2                             | 0.3225 | 0.0567     |
| 3   | 0.10         | 8                             | 0.0813 | 0.4861     |
| 4   | 0.11         | 8                             | 0.1583 | 0.1997     |
| 5   | 0.12         | 8                             | 0.1086 | 0.3464     |
| 6   | 0.13         | 8                             | 0.1836 | 0.1564     |
| 7   | 0.14         | 6                             | 0.1021 | 0.3750     |
| 8   | 0.15         | 2                             | 0.0807 | 0.4900     |
| 9   | 0.16         | 2                             | 0.0896 | 0.4320     |
| 10  | 0.17         | 2                             | 0.1243 | 0.2880     |
| 11  | 0.18         | 2                             | 0.0721 | 0.5458     |
3.2.3. Determination method of optimum clusters

The optimum determination method was conducted to compare k-modes clustering with the optimal cluster numbers and ROCK clustering with optimal cluster numbers. There were 2 cluster for the cluster (k) on k-modes clustering method, and so were for the ROCK clustering method, comparison of both methods used the measurement instrument of SW/SB ratio and R-squared. The comparison for the determination of optimum method in this research is shown in Table 4. The least SW/SB ratio value which was obtained by k-modes clustering was 0.0149. The R-squared value from k-modes clustering is also the highest compared to the ROCK clustering, namely 0.7558 or 75.58%. Whereas, the value of SW/SB and R-squared from ROCK clustering consecutively were 0.0682 and 0.5732 or 57.32%.

| Measurement Instruments | Cluster Methods |
|-------------------------|-----------------|
|                         | K-modes clustering (k = 2) | ROCK clustering (k = 2) |
| SW/SB                   | 0.0149           | 0.0682          |
| R-squared               | 0.7558           | 0.5732          |

3.3. Further analysis of k-modes clustering

The grouping of the 40% lowest income households resulting 2 optimal clusters by using k-modes clustering. The number of whole grouped household were 1449 households. The obtained grouping result are in cluster 1, there were 790 households and there were 659 households in cluster 2. These are the mode from each variable in each cluster.

| No. | Variables                        | Cluster 1                      | Cluster 2                      |
|-----|----------------------------------|-------------------------------|-------------------------------|
| 1   | Floor type of the shelter        | unplastered floor/ red bricks | unplastered floor/red bricks |
| 2   | Wall type of the shelter         | Regular wall                  | Regular wall                  |
| 3   | Disposal facility usage          | No disposal facility          | Private use                   |
| 4   | The main lighting source         | PLN Electricity               | PLN Electricity               |
| 5   | The main drinking water source   | Pumped well                   | Pumped well                   |
| 6   | The main cooking fuel            | Firewood                      | Firewood                      |
| 7   | Ownership of KIS/ BJS Health Insurance/Jamkesmas | Yes | Yes |
| 8   | The main head of household income source | Daily construction workers | Daily construction workers |
9 Head of household education background
Elementary / Extraordinary disable elementary school
Elementary / Extraordinary disable elementary school

| No. | Villages                  | Cluster 1 | Cluster 2 | Cluster 1 Percentage | Cluster 2 Percentage |
|-----|--------------------------|-----------|-----------|----------------------|----------------------|
| 1   | Ujung Tanah              | 16        | 18        | 47%                  | 53%                  |
| 2   | Payonan Gadang           | 57        | 42        | 58%                  | 42%                  |
| 3   | Lubuk Layu               | 33        | 30        | 52%                  | 48%                  |
| 4   | Suaq Hulu                | 41        | 33        | 55%                  | 45%                  |
| 5   | Luar                     | 38        | 20        | 66%                  | 34%                  |
| 6   | Tampang                  | 13        | 13        | 50%                  | 50%                  |
| 7   | Jilatang                 | 36        | 36        | 50%                  | 50%                  |
| 8   | Ujung Kampung            | 9         | 8         | 53%                  | 47%                  |
Cluster 2 had 659 households (Table 7). The number of households consecutively by welfare rates from the highest to the least are decile 4, decile 3, decile 2, and decile 1. Decile 2 and decile 4 have nearly similar number of households. The difference between decile 1 and decile 4 reached 54.63%. The number of households on decile 1, decile 2, decile 3, and decile 4 consecutively were 13 households (1.97%), 78 households (11.84%), 195 households (29.59%), and 373 households (56.60%).

Cluster 1 had higher proportion number of household percentage on decile 1 than the cluster 2. So, did on the decile 2 and decile 3. The comparison of proportion number of household percentage between cluster 1 and cluster 2, for decile 1, decile 2, and decile 3 consecutively were 13.80%:1.97%, 27.72%:11.84%, dan 32.53%:29.59%. Whereas, for decile 4, cluster 2 had higher proportion number of household percentage than cluster 1. The difference of the proportion number of household percentage on decile 4 reached 30.65% with the proportion number of household percentage on cluster 1 and cluster 2 consecutively were 25.95% and 56.60%.

Table 7. Comparison of welfare rates on each cluster

| Welfare Status | Cluster 1 | Cluster 2 |
|----------------|-----------|-----------|
|                | Number    | Percentage| Number    | Percentage|
| Decile 1       | 109       | 13.80%    | 13        | 1.97%     |
| Decile 2       | 219       | 27.72%    | 78        | 11.84%    |
| Decile 3       | 257       | 32.53%    | 195       | 29.59%    |
| Decile 4       | 205       | 25.95%    | 373       | 56.60%    |
| Total          | 790       | 100.00%   | 659       | 100.00%   |
4. Conclusions
The comparison between two methods of grouping namely k-modes clustering and ROCK clustering were conducted for 12 poverty indicators in Samadua Subdistrict, South Aceh District. The result of the analysis provided several following conclusions:

- By using grouping method of k-modes clustering, 2 optimum clusters (k=2) were obtained. Whereas by using grouping method of ROCK clustering, 2 optimum clusters (k=2) were also obtained at the threshold value of 0.29 (θ=0.29).
- By the comparison between k-modes clustering with k=2 and ROCK with k=2 clustering, obtaining k-modes clustering is the best method since k-modes clustering had SW/SB ratio value lower than the ROCK clustering, namely 0.0149 (k-modes clustering) and 0.0682 (ROCK clustering). The R-squared value of k-modes clustering was higher than the ROCK clustering’s, namely 0.7558 (k-modes clustering) and 0.5732 (ROCK clustering).
- Cluster 1 and cluster 2 has the similar mode on the variables of the floor type of shelter, the wall type of shelter, the main source of lighting, the main source of drinking water, the main cooking fuel, the ownership of KIS/BPJS Health Insurance/Jamkesmas, the main income source of household heads, the education background of the household heads, and the ownership of refrigerator with the each variable category consecutively were unplastered floor/ red bricks, regular wall, PLN electricity, pumped well, firewood, owning the access for KIS/BPJS Health Insurance/Jamkesmas, daily construction workers, Elementary / Extraordinary Elementary School, and the non-ownership of refrigerator. Whereas for the variable of the use of disposal facility, households on cluster 1 generally did not own the disposal facility and cluster 2 did have the disposal facility privately. Households in cluster 1 also generally did not own any television and motorcycle bike. Otherwise, households on cluster 2 own the television and motorcycle bike. This could be concluded that it is necessary to renovate the disposal facility in Samadua Subdistrict since the 40% lowest income households which did not own disposal facility privately reached 52.59%, particularly for households on cluster 1.

References
[1] Huang J Z 2009 Encyclopedia of Data Warehousing and Mining, Second Edition. Pennsylvania: IGI Global.
[2] Guha S, Rastogi R, Shim K 1999 Proceed. the 15th Int. Conf. on Data Eng., 512–521.
[3] Tyagi A, Sharma S 2012 Int. J. Comp. Sci. Eng. (IJCSE), 4 (5) 809–815.
[4] Badan Pusat Statistik Kabupaten Aceh Selatan 2018 Kabupaten Aceh Selatan dalam Angka 2018. Tapaktuan: Badan Pusat Statistik Kabupaten Aceh Selatan. (in Bahasa Indonesia).
[5] Badan Pusat Statistik Provinsi Aceh 2017 Aceh dalam Angka 2017. Banda Aceh: BPS Aceh. (in Bahasa Indonesia).
[6] Ulfah A N, ‘Uyun S 2015 Jurnal Teknik Informatika dan Sistem Informasi. 1 (2) 139–148. (in Bahasa Indonesia).
[7] Brock G, Pihur V, Datta S, Datta S 2008 Solid State Comm., 25 (4) 371–372.
[8] Dewi A, Sutikno, Kuswanto H 2012 Metode Cluster untuk Pengelompokan Desa Pedesaan di Provinsi Riau. 4 1–16.