Clustering analysis of senior high school in West Java based on educational facilities

M A Putri and S Abdullah
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Corresponding author’s email: sarini@sci.ui.ac.id

Abstract. The problem of education in senior high school remains. One of the roots of the problem is the lack of educational facilities. The government has responded to this case by giving the state budget (APBN) for education. Nevertheless, the large number of senior high schools in Indonesia is one of the barriers to efficient distribution of APBN funds. Thus, clustering analysis of senior high school facilities in Indonesia especially in West Java is expected to be a solution. The number of observations is 13,486 with nine categorical variables recorded on a website of the Ministry of Education and Culture in August 2019. The method used is Robust Clustering Using Link (ROCK). To get a clearer profile of the clusters, we used ROCK method modified with Nested Clustering. The results of this study indicate that 14 clusters were formed and have their profiles. Cluster 3 is a relatively good cluster while cluster 1a is a relatively poor cluster. Clusters representing poor facilities were found to be dominant in West Java Province.

Keywords: Categorical clustering, Jaccard coefficient, robust clustering, similarity

1. Introduction
The existence of adequate educational facilities is needed to support an organized and sustainable education process. The government has given attention to support education in Indonesia by allocating the state budget (APBN). APBN in 2019 allocated 20 % or Rp.492.55 trillion for education in Indonesia [1]. The Minister of Finance considers that the Ministry of Education and the Local Government has not allocated education funds which has a significant impact both in terms of planning and execution. However, the large number of schools in Indonesia has had an impact on providing APBN funds to each school that cannot be done quickly and accurately. The Ministry of Education and Culture (2019) states that the total number of schools in Indonesia is 270,171 schools [2].

Senior high school is one of the most important levels considered by the implementation of the 12-year compulsory education. Thus, based on existing data and the problems, researchers are interested clustering senior high schools in Indonesia especially in West Java Province based on educational facilities. The purpose of this study is to know the condition of senior high schools in Indonesia, then recommend to the government in allocating and distributing APBN funds by prioritizing senior high schools based on the level of need in terms of educational facilities. It is expected that improvements of educational facilities can be done quickly and appropriately. Therefore, in this study, we did clustering analysis of senior high school in Indonesia especially in West Java Province based on educational facilities.
2. Materials and method
The method used to cluster the data is Robust Clustering Using Link (ROCK). ROCK is a method of clustering data using the concept of distance measured by a coefficient called a Jaccard coefficient to measure the similarity or closeness between a pair of data points [3]. According to Reddy and Kavitha (2012), ROCK method has several advantages, i.e. to have a better level of accuracy when compared to other hierarchical methods, and is highly scalable, ensuring that it can handle large quantities of data [4].

2.1. Variables
In recording to the educational facilities, the government uses indicators in the form of numerical and categorical data. Whereas in clustering, [5] explained that it cannot treat the same method in clustering categorical data and numeric data. This is because the clustering of categorical data is more complex compared to numerical data. However, the government has provided an ideal standard of several educational facilities regulated in government regulations, ministerial regulations, and laws. These regulations are foundations to do categorization of numerical data. Thus, the data in this study are categorical data with the following categories of information (table 1).

| No | Variable                      | Label of category                        |
|----|-------------------------------|------------------------------------------|
| 1  | Electrical power              | 1 : Not available                        |
|    |                               | 2 : < 900 VA                             |
|    |                               | 3 : 900–2200 VA                          |
|    |                               | 4 : 2200–4400 VA                         |
|    |                               | 5 : 4400–14,000 VA                       |
|    |                               | 6 : ≥ 14,000 VA                          |
| 2  | Number of laboratories        | 1 : 0                                    |
|    |                               | 2 : 1 to 4                               |
|    |                               | 3 : 5                                    |
|    |                               | 4 : > 5                                  |
| 3  | Library                       | 1 : Available                            |
|    |                               | 2 : Not available                        |
| 4  | Ratio of students to teachers | 1 : < 15                                 |
|    |                               | 2 : 15 to 25                             |
|    |                               | 3 : > 25                                 |
|    |                               | 4 : No teachers                          |
| 5  | Ratio of students to classrooms | 1 : < 20                                |
|    |                               | 2 : 20 to 36                             |
|    |                               | 3 : > 36                                 |
|    |                               | 4 : No classrooms                        |
| 6  | Water adequacy                | 1. Not available                         |
|    |                               | 2. Inadequate                            |
|    |                               | 3. Adequate                              |
| 7  | School status                 | 1. Public school                         |
|    |                               | 2. Private school                        |
| 8  | Water source                  | 1. Tap water or pump water               |
|    |                               | 2. Rainwater, river, unprotected springs, unprotected wells |
|    |                               | 3. Water springs                         |
|    |                               | 4. Wells                                 |
| 9  | Electrical source             | 1 : Available                            |
|    |                               | 2 : Not available                        |
2.2. Robust clustering using link (ROCK)

Clustering is a technique of clustering data into groups with characteristics that are close to each other [3]. Johnson and Wichern define clustering as a grouping of all observations based on the level of similarity of observational characteristics [6]. Thus, any observations that are similar to other observations will be in the same cluster.

Robust clustering using links or ROCK is a hierarchical clustering algorithm that uses the concept of links to measure the similarity and closeness between observations [3]. Observations with a high link (relationship level), will be grouped into one cluster. While observations that have a low link will be separated from the grouping. The number of links between the two observations states the number of neighbors owned. After the initial calculation of the number of links between observations, the algorithm starts by assuming each cluster to be a single object and then combines the clusters based on the goodness measure value. The merger is continued until one of the following two criteria is met, i.e. the number of clusters expected reached and there are no links between clusters.

The stages of the ROCK algorithm according to [3] are as the following.
1. Initiating by assuming of each observation considered as a cluster that has one member.
2. Calculate the similarity between groups using equation 1 (the Jaccard coefficient)

\[ \text{sim}(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}, \quad i \neq j \]  

where \( X_i \) is the categorical set of the \( i \)-th observation, while \( X_j \) is the categorical set of the \( j \)-th observation. The categorical set is a set whose members are in the form of categorical data. \( |X_i \cap X_j| \) state the number of identical members between \( X_i \) and \( X_j \), and \( |X_i \cup X_j| \) state the number of joint members of \( X_i \) and \( X_j \).
3. Determine the number of clusters (\( n \)) and determine the threshold value (\( \theta \)) to form a neighbor matrix. The value \( \theta \) is the parameter chosen by the researcher to determine the close relationship between observations, whether a couple of observations are neighbors or not. Threshold values are between 0 and 1.
4. Form a neighbor matrix \( A \) based on the value \( \theta \). The element of neighbor matrix \( A \) in the \( i \)-th row, \( j \)-th column is 1 if \( \text{sim}(X_i, X_j) \geq \theta \) and 0 if \( \text{sim}(X_i, X_j) < \theta \).
5. Count the number of links, there are two ways to count explained as follows.
   a. The link between the two clusters with each cluster contains 1 observation:
      If \( C_i \) is the \( i \)-cluster which have one observation \( X_i \), then the link is calculated by equation 2.
      \[ \text{Link}(C_i, C_j) = \text{Link}(X_i, X_j) = \left| T_{X_i} \cap T_{X_j} \right|, \quad i \neq j \]  
       where \( T_{X_i} \) is the neighboring set of the \( i \)-th observation, \( T_{X_j} \) is the neighboring set of the \( j \)-th observation, and \( \left| T_{X_i} \cap T_{X_j} \right| \) is the number of same elements of neighboring set between \( X_i \) and \( X_j \).
   b. The link between two clusters with each cluster contains more than 1 observation:
      If \( C_i \) is the \( i \)-th cluster which is a set of \( X_i \) members that are placed into one group. Then, the link is calculated by equation 3.
      \[ \text{Link}(C_i, C_j) = \sum_{X_i \in C_i, X_j \in C_j} \text{Link}(X_i, X_j), \quad i \neq j \]  
       where \( \text{Link}(C_i, C_j) \) states the number of links between each possible pair of observations that are in \( C_i \) and \( C_j \).
6. Calculate the goodness measure value to join two groups using the following equation,

\[ g(C_i, C_j) = \frac{\text{Link}(C_i, C_j)}{(n_i + n_j)^{1/2} - n_i^{1/2} n_j^{1/2}} \]  \hspace{1cm} (4)

where \( n_i \) and \( n_j \) each represents the number of members in the i-th and j-group and \( f(\theta) = \frac{1-\theta}{1+\theta} \).

7. Combine group with the largest goodness measure, then recalculate the links between groups and update the new goodness measure values.

8. Perform steps 6 through 8 until the number of expected groups is reached or until there are no more links between groups.

9. Repeat steps 1 to 8 with different \( \theta \) values.

10. Calculate the ratio between \( S_W \) and \( S_B \) for each \( \theta \) value by equation 5 and equation 6, respectively.

11. Compare the results from step 10 for each \( \theta \) value to determine the optimal \( \theta \) value with the criterion of a small ratio of \( S_W \) and \( S_B \).

2.3. Validation ROCK clustering

Determining the optimum number of groups (\( n \)) and the optimum threshold value (\( \theta \)) are important stages in clustering. Cluster validation will measure the performance results of clustering. A good cluster is one with a high homogeneity between members in the group and high heterogeneity between groups [4]. \( S_W \) is the mean of the standard deviation in groups, while \( S_B \) is the mean of standard deviations between groups calculated by equation 5 and equation 6, respectively [7]. To get a good clustering result, choose a value of \( \theta \) which has the smallest \( S_W \) and \( S_B \) ratio, which means that both the homogeneity within groups and heterogeneity between groups reach maximum values.

\[ S_W = \frac{1}{j} \sum_{j=1}^{j} \left[ \frac{n_i - \frac{1}{2} \sum_{c=1}^{g} \frac{1}{n_{cj}} \sum_{k=1}^{K} n_{kj} 2^{1/2}}{n - g} \right] \]  \hspace{1cm} (5)

\[ S_B = \frac{1}{j} \sum_{j=1}^{j} \left[ \frac{\left( \frac{1}{n_{cj}} \sum_{k=1}^{K} n_{kj} 2^{1/2} \right)^2 - \left( \frac{1}{2n} \sum_{k=1}^{K} n_{kj} 2^{1/2} \right)^2}{g - 1} \right] \]  \hspace{1cm} (6)

where \( n \) is the number of observations, \( j \) is the number of variables, \( g \) is the number of clusters, \( k \) is the number of categories, \( n_{kj} \) is the number of observations with the \( k \)-th category and the \( j \)-th variable, \( n_{cj} \) is the number of observations with the \( c \)-th group and the \( j \)-th variable, while \( n_{kj} \) is the number of observations in category \( k \), group \( c \), and variable \( j \) [8].

3. Results and discussion

3.1. Statistical descriptive

The data in this study is secondary data consisting of educational facilities of senior high schools in Indonesia including public and private schools. The observations used are 13,486 schools registered in a website the Ministry of Education and Culture in August 2019. The data can be accessed through the website dapo.dikdasmen.kemdikbud.go.id [2]. Charts were produced using package ggplot2 in R [9].
Figure 1 shows the condition of senior high school education facilities in Indonesia. Educational facilities with relatively good conditions include electricity sources and adequate water because 95% of high schools have adequate electricity and water. Educational facilities with relatively good conditions but need attention include the library, ratio of student per classroom, electrical power, and water sources because there are 17% of senior high schools that do not have a library, 4% of senior high schools do not have classes, and 13% of senior high schools do not have a good source of water which is obtained from rain water, river water, unprotected springs, and unprotected wells. Meanwhile, educational facilities with relatively poor conditions and need special attention include the number of laboratories and the ratio students to teachers. The number of public and private senior high schools in Indonesia is almost the same.

3.2. Determination value of threshold ($\theta$)

First, initiate use $\frac{S_w}{S_B}$ with several $\theta$ values. The purpose of initiation to find range of optimum $\theta$, which mean have small value of ratio $\frac{S_w}{S_B}$ and resulting meaningful interpretation. In this study, the meaning of meaningful interpretation is each cluster has unique characteristic. Afterward, the optimum $\theta$ obtained are $\theta = 0.01; 0.05; 0.1; 0.2$. Because of several clusters have similar characteristics, it was decided to reduce the size of $n$. The next $n$ is $n = 9$ with $\theta = 0.01; 0.05; 0.1; 0.2$. After that, evaluate ratio $\frac{S_w}{S_B}$ for each $\theta$, and the optimum $\theta$ is $\theta = 0.2$ which has the smallest value of ratio $\frac{S_w}{S_B}$ (see table 2), resulting in a meaningful interpretation.

After obtaining nine clusters, it turns out that Cluster 1 has not meaningful interpretation, in this study, the meaning of has not meaningful interpretation is Cluster 1 has not unique characteristic. To get a clearer profile of the cluster, we modified the ROCK method, called Nested Clustering, which requires clustering all members in Cluster 1. In this step, we use $\theta = 0.2$ and $n = 6$ then obtained cluster 1a, 1b, 1c, 1d, 1e, and 1f as partitions of Cluster 1. See figure 2 for the distribution for each cluster’s number of senior high schools.

Thereafter, describe the character of each cluster based on mode threshold (see table 3), i.e. average of number of senior high schools for each category expressed as a percentage. A category will be a mode if the its percentage is greater than or equal to the mode threshold. That category is therefore a character of the cluster.

![Figure 1. Descriptive statistics for categorical variables](image-url)
Table 2. Value of ratio $\frac{S_w}{S_B}$ for each $\theta$.

| $\theta$ | $S_w$ | $S_B$ | $S_w/S_B$ |
|---------|-------|-------|-----------|
| 0.01    | 0.422 | 23.301| 0.0181    |
| 0.05    | 0.422 | 23.319| 0.0181    |
| 0.1     | 0.414 | 22.491| 0.0184    |
| 0.2     | 0.423 | 23.730| 0.0178    |

Table 3. Mode threshold for several number of categories.

| Number of category | Mode threshold |
|--------------------|----------------|
| 2                  | 50%            |
| 3                  | 33%            |
| 4                  | 25%            |
| 6                  | 17%            |

Figure 2. (a) Distribution chart of number of senior high school for each cluster *before* modified with Nested Clustering, and  (b) Distribution chart of number of senior high school for each cluster *after* modified with Nested Clustering.
For example, the number of laboratories has 4 categories, so that the mode threshold is 25%. In table 4, only category 2 has a percentage greater than 25%, which means Cluster 4 has a characteristic in the form of clusters with a number of laboratories between 1 and 4. For complete characteristics of all clusters, see table 5 and table 6. In general, Cluster 3 is the relatively good cluster because it has the most of good facilities and Cluster 1a is the relatively poor cluster because has the most of poor facilities.

3.3. Condition of educational facilities in Indonesia

Figure 3 represents the general condition of educational facilities in Java, Sumatra, Kalimantan, Sulawesi, and Papua. Based on point density, it is seen that the distribution of senior high schools in Java is very dense in almost every region. On the island of Sumatra, it also looks dense in the south and north, while for other parts it is still relatively rare. On the island of Kalimantan, it is seen that senior high schools are spread evenly in several parts. However, there are a few senior high schools in the north, this can be explained by North Kalimantan being youngest province, so the number of senior high schools in the area is relatively low. On the island of Sulawesi, it is seen that the points are more closely packed at the tip of the island and are spread out on the edges of the island. This indicates that high schools in Sulawesi tend to be on the coast. Meanwhile, on the island of Papua, it is seen that the distribution of points is very sparse, meaning that there are still few of senior high schools in Papua. Thus, it can be concluded that the distribution of senior high schools in Indonesia is uneven.

### Table 4. Characteristic of cluster 4.

| Facility        | Category | Cluster 4 (%) | Meaning |
|-----------------|----------|---------------|---------|
| Number of laboratory | 1 | 0 (0%) | 0 |
|                 | 2 | 360 (70%) | 1–4 |
|                 | 3 | 69 (13%)  | 5 |
|                 | 4 | 83 (16%)  | >5 |

### Table 5. Characteristic of cluster 1a, 1b, 1c, 1d, 1e and 1f.

| Facility          | C1a | C1b | C1c | C1d | C1e | C1f |
|-------------------|-----|-----|-----|-----|-----|-----|
| Water adequacy    | Adequate | Adequate | Adequate | Adequate | Adequate | Adequate |
| Number of laboratories | 1 to 4; 0 | 1 to 4; 0 | 1 to 4 | 1 to 4 | 1 to 4 | 1 to 4; 0 |
| Library           | Available | Available | Available | Available | Available | Available |
| Electrical power  | Cat 1, 3, 5, and 6 | Cat 3, 5, and 6 | Cat 3, 4, 5, and 6 | Cat 3, 4, 5 | Cat 3 | Cat 3, 4, and 5 |
| School status     | Public School | Private School | Public School | Private School | Public School | Private School |
| Water source      | Cat 1 and 4 | Cat 4 | Cat 1 | Cat 1 and 2 | Cat 1 and 4 |
| Electrical source | Available | Available | Available | Available | Available | Available |
| Ratio of students to teachers | < 15 | < 15 | < 15; 15 to 25 | < 15 | < 15 | < 15; 15 to 25 |
| Ratio of students to classrooms | > 36 | < 20; 20 to 36 | 20 to 36 | < 20; 20 to 36 | < 20; 20 to 36 | < 20; > 36 |
Table 6. Characteristic of cluster 2, 3, 4, 5, 6, 7, 8 and 9.

| Facility          | C2    | C3     | C4     | C5     | C6     | C7     | C8     | C9     |
|-------------------|-------|--------|--------|--------|--------|--------|--------|--------|
| Water adequacy    | Adequate | Adequate | Adequate | Adequate | Adequate | Adequate | Adequate | Adequate |
| Number of laboratories | 1 to 4 | 1 to 4; 5; >5 | 1 to 4 | 1 to 4 | 1 to 4 | 1 to 4 | 1 to 4 | 1 to 4 |
| Library           | Available | Available | Available | Available | Available | Available | Available | Available |
| Electrical power  | Cat 4, 5, and 6 | Cat 6 | Cat 4, 5 and 6 | Cat 3, 4, and 6 | Cat 4 and 5 | Cat 4 and 5 | Cat 3, 4, 5, and 6 | Cat 3 |
| School status     | Public School | Public School | Public School | Public or Private School | Public School | Public School | Private School | Public School |
| Water source      | Cat 1 and 2 | Cat 1 | Cat 1 | Cat 1 and 4 | Cat 4 | Cat 1 | Cat 1 and 4 | Cat 1 and 4 |
| Electrical source | Available | Available | Available | Available | Available | Available | Available | Available |
| Ratio of students to teachers | < 15 | < 15 | 15 to 25 | < 15, 15 to 25 | < 15 | < 15 | 15 to 25, > 25 | < 15 |
| Ratio of students to classrooms | < 20, 20 to 36 | 20 to 36 | 20 to 36, > 36 | < 20 | 20 to 36 | 20 to 36 | 20 to 36 | 20 to 36 |

Based on the color distribution, yellow is dominant in Java. This indicates the existence of senior high schools classified as cluster 1b dominant in Java, which means they have adequate water, the number of laboratories is below the minimum standard, they generally have a library, the water source is relatively good (because it is predominantly sourced from Plumbing / PAM, pumps, and wells), the ratio of students to teachers is above standards, and ratio of students to classrooms is average or better than standards. In general, most senior high schools in Java must pay attention to several facilities such as increasing the number of laboratories, making the number of teachers and classrooms equal to standards. Meanwhile, on the island of Sumatra, yellow is common, but only in the north and south. On the island of Kalimantan, light green is mildly dominant in the east which indicates the dominance of cluster 3, which is a cluster that has relatively good educational facilities.

3.4. Condition of educational facilities in West Java Province

Figure 4 represents in general the condition of educational facilities in West Java Province. Based on point density, it is seen that the distribution of high schools is quite dense seen in Bekasi City, Depok City, Bogor City, Bandung City, Bandung Regency, Tasikmalaya City, and Garut Regency. Whereas for other regencies/cities it is rare, meaning the small number of senior high schools in these areas.

Based on the color distribution, in general in the West Java Province there is a domination of yellow, which is an indicator for cluster 1b that characterizes of senior high schools with fewer laboratories than standards (or none at all). The indicated yellow dominance is in Bekasi Regency, Bekasi City, Bogor Regency, Bandung City, and Garut Regency. Meanwhile, the dominance of the colors that indicate cluster 4 is in Karawang regency, which means the number of laboratories and the ratio of students per class are substandard.
Figure 3. Distribution map of cluster in Indonesia

Figure 4. Distribution map of cluster in West Java

For details, see figure 5a to find out the distribution of clusters in West Java Province. It is known that the largest percentage of clusters in West Java Province is Cluster 1b. Furthermore, in figure 5b Cluster 1b is dominant in Garut Regency, meaning that there are many of senior high schools in Garut Regency in need of more laboratories. A suggestion for Garut Regency are as much as possible to increase the number of laboratories so that students are better served.

Note the percentage of senior high schools in Cluster 3 which represent relatively good cluster is only 1% of the total senior high schools in West Java. For Cluster 1a which represent relatively
not good cluster it does not dominate but the percentage is greater than in Cluster 3, meaning that senior
high schools in West Java are more likely to be substandard. Thus, it can be concluded that the condition
of senior high schools in West Java requires special attention from the local government to improve the
condition of senior high schools. The improvements needed are adding the number of laboratories, an
electricity supply for senior high schools that do not have electricity, standardizing the number of
teachers, and the addition of classrooms.

4. Conclusion
Clusters representing poor facilities are indicated dominant in West Java Province. High schools in
Garut Regency require more of laboratories. For further research, researchers can use other clustering
methods to compare with the ROCK clustering method and consider using spatial analysis. For the local government, the results of this study are expected to provide an illustration for the local government so that they are able to immediately make policies that are adjusted to the needs of each senior high school in each region in order to improve the quality of education, especially in the aspect of educational facilities.

Acknowledgments
This work was financially supported by Universitas Indonesia under research grant PITTA B with grant contract number NKB-0665/UN2.R3.1/HKP.05.00/2019.

References
[1] Kementrian Keuangan 2019 Informasi APBN 2019 available at https://www.kemenkeu.go.id/media/11213/buku-informasi-apbn-2019.pdf
[2] Kementrian Pendidikan dan Kebudayaan 2019 Data Pokok Pendidikan Dasar dan Menengah available at http://dapo.dikdasmen.kemdikbud.go.id/school/C5D4DADD1BDE7D948BE4
[3] Guha S, Rastogi R and Shim K 1999 Inform. Syst. 25 345-66
[4] Reddy M V J and Kavitha B 2012 Int. J. Database Theory Appl. 5 121-33
[5] Cleff T 2013 Univariate data analysis Exploratory Data Analysis in Business and Economics (Germany: Springer)
[6] Johnson R A and Wichern D W 2002 Applied Multivariate Statistical Analysis (Upper Saddle River: Prentice hall)
[7] Latifah T D 2018 Segmentasi Pelanggan Susu Formula Menggunakan Cluster Ensemble Berbasis Algoritme ROCK BSc. Final Projects (Bogor: Departemen Statistika, FMIPA, Institut Pertanian Bogor)
[8] Alvionita, Sutikno and Suharsono A 2017 IOP Conf. Seri.: Earth Environ. Sci. 58 012029
[9] Wickham H 2009 Ggplot2: Elegant Graphics for Data Analysis (Germany: Springer)